Implementation of Optimization Techniques for Multi-Document Summarization

Ashwini Jadhav[1], Dr. K. V. Metre[2]

[1]PG Student, Computer Dept., MET BKC, Adgaon, Nasik, Maharashtra, India

[2]Professor, Dept. of Computer Engineering, MET BKC Adgaon, NasikMaharashtra,India.

Abstract

Presentation-based learning is an effective way of learning in which students or employee of the organization are able to receive continuous feedback from teammates or from their coaches. It is the pictorial way to represent the work. Before going towards the presentation of the work presenters have to work on the slides. These slides of presentation are made from the article, some academic papers or with the help of internet. It results into wastage of more timing to create slides rather than focusing on preparation of the presentation. In this paper, we are analysing the way to automatic generation of the presentation slides from academic papers. Due to this the presenters can prepare their formal slides in a quicker way. Therefore, we are proposing PPSGen system to address this problem of existing system. PPSGen have lots of advantages over baseline methods.

1 Introduction

Presentation slides is the most effective and popular way of presentation. Especially it is effective for academic conferences and useful for students as well as employees of the organization. Presentation-based learning is pictorial or visible way of learning which helps employees to collect the feedback immediately from their instructor, team-mates or from their staff. Currently, many of software's are available to make slides for researcher's paper. For example, Microsoft Power Point and OpenOffice, these help only in formatting of the slides. It does not provide the contents for slides. In this paper, we are proposing a method to automatic generate slides that are well-structured from the specified academic papers. This approach helps presenter's to reduce the time as well as efforts required for the final presentation.

Academic papers have same structure like, abstract, introduction, problem statement, proposed system, system architecture, experimental results, conclusion etc. whereas, presentation slides have different structure which depends on presenter's. In this paper, we propose PPSGen method for automatically generating the presentation slides. It helps to generate draft slides of the paper for final presentation. Current, methods of constructing slides simply extract the sentences from the papers. For that they required more structured form of the paper. PPSGen is a very challenging task which supports vector regression (SVR) model with a number of useful features and also integer linear programming is used to elaborate objective functions, align key phrases and sentences. SVR-based sentence scoring model is used to assign an importance score for each sentence in the academic paper. PPSGen system address the problem of existing system and it has lots of advantages over baseline methods as it generates the slides with better quality than baseline methods.

ISSN: 2278 – 7798

2 Literature Survey

PPSGen:Learning-Based PresentationSlides Generation for Academic Papers[1] Academic papers have same structure like, abstract, introduction, problem statement, proposed system, system architecture, experimental results, conclusion etc. whereas, presentation slides have different structure which depends on presenter's. In this paper, we propose PPSGen method for automatically generating the presentation slides. It helps to generate draft slides of the paper for final presentation. Current, methods of constructing slides simply extract the sentences from the papers. For that they required more structured form of the paper. PPSGen is a very challenging task which supports vector regression (SVR) model with a number of useful features and also integer linear programming is used to elaborate objective functions, align key phrases and sentences. SVR-based sentence scoring model is used to assign an importance score for each sentence in the academic paper. PPSGen system address the problem of existing system and it has lots of advantages over baseline methods as it generates the slides with better quality than baseline methods.

Abu-Jbara and D. Radev [2]This paper introduced citation-based summarization approach for scientific papers to produced readable summarization. This approach filter out the irrelevant sentences, extracts the set of sentences that are representative and finally refined them to improve readability. In this paper, author name the sentence that contains an explicit reference to other paper citation sentence. It is important task to highlight the most important aspects of the cited paper. Coherence and readability aspects of the problem are considered in this paper. The proposed approach produces citation-based summaries in three steps such as, pre-processing, extraction, and post processing. Therefore it produces better summaries than other several baseline summarizations for functional Category Classification SVM with linear kernel is used.

V. Qazvinian, D. R. Radev, S. M. Mohammad, B. J. Dorr, D. M.Zajic, M. Whidby, and T. Moon [3] In this paper authors proposed C-LexRank, a graph-based summarization model. This model automatically summarized 30 single scientific articles selected from 6 different topics in the ACL Anthology Network (AAN). Authors were mainly focusing to generate extractive summaries of a set of Question Answering (QA) and Dependency Parsing (DP) papers. Bibliometric lexical link mining that exploits the structure of citations and summarization techniques are combining. From multiple research papers in automatically generating a technical summary on a given topic authors compare and contrast the usefulness of abstracts and of citations. The proposed approach in this paper, which is C-LexRank is effective in producing a summary of a paper's contributions. In this to find a lower-bound on the pyramid scores they compare C-LexRank with random summarization. C-LexRank is used to generate summarization of 30 papers, six different topics in ACL Anthology Network (AAN) and also generated summaries of Question Answering (QA) set and Dependency Parsing (DP) papers.

V. Qazvinian and D. R. Radev [4] This paper identifies the background details of the scientific papers. Afterthat, Markov Random Field tuned to detect the patterns. These

ISSN: 2278 – 7798

patterns are used to context data create, and employ a Belief Propagation mechanism. This methodology is based on MRF. This paper refers implicit citations which contain information about a specific secondary. In implicit citations patterns that such sentences create and observe that context sentences of explicit citations are examined. The proposed model in this paper is based on the probabilistic inference of the random variables with help of graphical models. Before on Information Retrieval tasks author's uses Graphical models which having a number of properties and corresponding techniques. In digital libraries of classic texts, Conditional Random Fields (CRF) is used to extract references from unstructured text. Proposed method extracts the context information of a cited paper. MRF model is build hidden nodes are corresponds to each other are observed.

V. Qazvinian and D. R. Radev[5] In this paper, the proposed model summarizes single topic from the article and this summarized topic is further used to summarize the entire topic of the specified article. Clustering approach is used in this paper. The main contributions of this is to use citation summaries and network analysis techniques which produce a summary of a single scientific article as a framework for future research on topic summarization. Corpus is built by extracting small clusters from the AAN data. Dependency Parsing (DP), Phrased Based Machine Translation (PBMT), Text Summarization (Summ), Question Answering (QA), and Textual Entailment (TE) these clusters are collected in this paper. Non-overlapping contribution (fact) is used to perceive each item on the list. In this paper, for article summarization graph clustering method is used. Experimental result outperforms the current state-of-art multi-document summarizing.

Q. Mei and C.Zhai[6] This paper proposed language modelling methods to incorporate features such as authority and proximity to accurately estimate the impact language model. For the evaluation of greater impact summarization test set is based on ACM SIGIR papers. To exploit both the citation context and original content of a paper to generate impact-based summary authors proposed a language models. This paper made the study about study to incorporate features such as authority and proximity into the estimation of language models. An impact-based summary is used for facilitating the exploration of literature, it also helps to generate query.

M. A. Whidby [7]This paper represented issues arises with parsers and summarization systems on documents which contains citations in scientific literature. As a solution, constituent and parenthetical citations are proposed in this paper. In this approach, constituent citations are replaced with filter text containing a unique identifier, and parenthetical citations are removed.

3 System Design

This process automatically generates presenta-tion slides with content fetchedfrom academic PDF document. Using PDFlib software, content is fetchedfrom document. For analysis purpose, this content should be in normalized form and analysis should be done very fast.

ISSN: 2278 – 7798

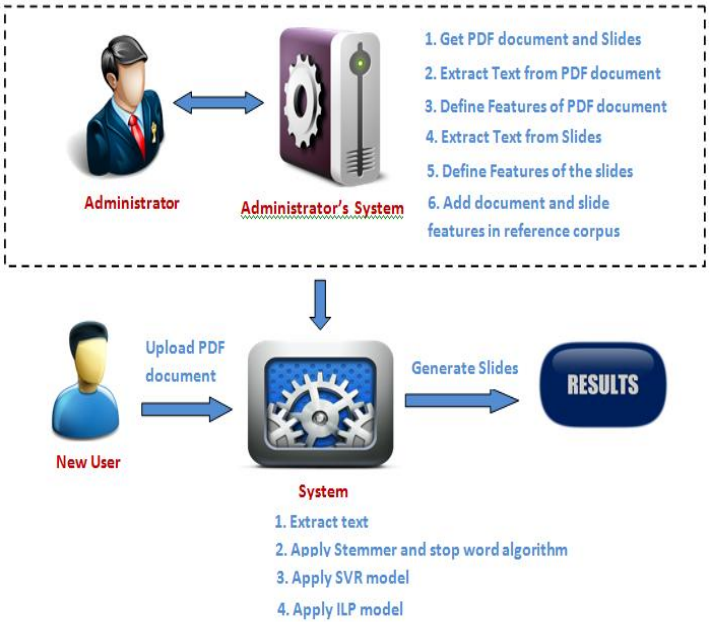

Fig: System Design

Methodology for end user:

A] Upload academic PDF file:

User selects the PDF file and upload to generate presentation slides of it.

B] View slides:

System works out on the PDF file and generates presentation slides with meaningful content. This content is fetched from the PDF file uploaded by the user.

Methodology for System:

A] Load PDF document:

System will get the PDF file uploaded by the user. System first load it into buffer to work on it.

B] Extract Text:

System will use PDFlib software to extract text from the PDF document.

C] Pre-processing:

On the extracted text, system apply Stop word algorithm and stemmer algorithm to get original form of words. After that features traced with the help of ready corpus designed explicitly. Finally XML document of feature is created and used further for slide generation.

D] SVR Model:

System use Support Vector regression model for sentence importance assessment by calculating importance score of each statement. Highest score terms, sentences, phrases will be used as key statements, phrases for slide generation.

E] ILP model:

After getting importance score for each sentence from academic PDF document with the help of SVR model, ILP is used to generate fine draft slides with content as key phrases and sentences.

F] Slide Generation:

After finalizing the key phrases, key sentences and slide patterns using Microsoft API automatic slides are generated. Selected content is added in the slides as per predefined formatted pattern.

Methodology for administrator:

Administrator is responsible to enrich the corpus which is required for slide generation having features and patterns of PDF document and slides belongs to particular academic documents. Hence following are the methodologies of administrator

A] Visit the URL:

Administrator visit the URL where datasets are present. These datasets will be having pair of PDF document and its respective presentation slides.

B] Download the dataset:

After selecting pair of PDF document and presentation slides, administrator downloads the particular dataset.

C] Extracting text:

For analysis purpose text must be extracted from the PDF document. PDFlib software used for this. Also ParsCit is used to get physical structure of paragraphs and sections. To get text from the slides XPDF or Microsoft API is used.

ISSN: 2278 – 7798

After structure analysis respective XML is created to describe the patterns for both, document as well as slides. These patterns are preserved in corpus.

4 Conclusion

Based on the literature survey it is clear that academic slides generation from documents is under research. Existing systems had worked on tagging document, LATEX document, raw document, technical document etc. to generate slides. All these existing systems have flaws like input format and non-alignment problem of key phrases and key sentences that creates slides in imperfect manner. PDF document as input is rarely used in existing system which is globally accepted form of technical papers. Hence there must be system that work on slide creation from academic documents available in PDF format. Also slide content should be more accurate and relevant.

References

1. Yue Hu and Xiaojun Wan "PPSGen: Learning-Based Presentation Slides Generation for Academic Papers", IEEE Transaction on knowledge and data engineering VOL. 27, NO. 4, pp 1085-1097. APRIL 2015

2. 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.

3. V. Qazvinian, D. R. Radev, S. M. Mohammad, B. J. Dorr, D. M.Zajic, M. Whidby, and T. Moon, "Generating extractive summaries of scientific paradigms," J. Artif. Intell. Res., vol. 46, pp. 165–201, 2013.

4. 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.

5. V. Qazvinian and D. R. Radev, "Scientific paper summarization using citation summary networks," in Proc. 22nd Int. Conf. Comput.Linguistics-Volume 1, Aug. 2008, pp. 689–696.

6. Q. Mei and C.Zhai, "Generating impact-based summaries for scientific literature," in Proc. ACL, vol. 8, pp. 816–824, 2008.

7. M. A. Whidby, "Citation handling: Processing citation texts in scientific documents," Doctoral dissertation, Dept. Comput. Sci.Univ. Maryland, College Park, MD, USA, 2012.

ISSN: 2278 – 7798