International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

A Review on Learning Based Automatic PPT Generation Using Machine Learning

Abhineet Ranjan[1], Akash Gangadhare[1] , S. V. Shinde[2]
[1]BE, Department of Computer Engineering, AISSMS College of Engineering, Pune, Maharashtra, India
[2]Assistant Professor, Department of Computer Engineering, AISSMS College of Engineering, Pune, Maharashtra, India

ABSTRACT

Presentation slides are widely used to communicate information to the audience. There are various tools available in the market which only deals with the formatting of the slides but not the content. However, this traditional way of preparing slides is labor-intensive in nature and leaves scope for human errors. Also, for lengthy documents, there is a chance of some important information being missed out. The drawbacks of the traditional way lead to a need for an intelligent system. The intelligent system needs to be capable of generating slides with minimum human interference. In this paper, we are enforcing the automated PPT creation from multi-documents of different extensions based on input query or title that formulate extraction of valuable information source and model a presentation view to automating slide creation using integer linear programming (ILP) method to generate well-structured slides by selecting and aligning key phrases and sentences. This will eventually help in reducing a great amount of the presenter's time and efforts. The proposed system works on natural language processing (NLP) rules to classify data for the desired slides.

Keywords: Classification, NLP, Support Vector Regression (SVR), ILP, Slide Generation, NLTK, Feature extraction.

I. INTRODUCTION

Microsoft PowerPoint, virtual presentation software developed by Robert Gaskins and Dennis Austin for the American computer software company Forethought, Inc. PowerPoint was designed to facilitate visual demonstrations for group presentations in the business environment. Presentations are arranged as a series of individually designed slides that contain images, text, or other objects. The presenter has numerous programming tools to assist him in setting up the slides, including Microsoft Power-Point, Open Office, and Libre Office. Such tools are helpful in setting up the theme and outline of the presentation; however, they do not help presenters in selecting the content for the slides.

The traditional tools thus require a lot of investment, in terms of time and efforts. Collectively, a group of slides may be known as a slide deck [4]. The main focus of this project is to develop a system that helps to generate powerpoint presentation based on user query thus, preventing users time and increase performance. Support Vector Regression is used in maintaining all the main features which characterize the maximal margin of the algorithm. The Support Vector (SV) algorithm is a nonlinear generalization of the Generalized Portrait algorithm developed in Russia in.

The key idea is to construct a Language function from the objective function which is called as the primal objective function and the corresponding constraints, by introducing a dual set of variables. Figure 4.1 Automatic slides generation for multiple documents is a very challenging task. Current methods generally extract objects like sentences from the file to construct the slides. In contrast to the short summary extracted by a summarization system, the slides are required to be much more structured and much longer. Slides can be divided into an ordered sequence of parts. Each part addresses a specific topic and these topics are also relevant to each other. Generally speaking, automatic slide generation is much more difficult than summarization. Slides usually not only have text elements but also graph elements such as figures and tables.

Documents always have a similar structure. They generally contain several sections. Although presentation slides can be written in various ways by different presenters, a presenter, especially for a beginner, always aligns slides sequentially with the paper sections when preparing the slides. Each section is aligned to one or more slides and one slide usually has a title and several sentences. These sentences may be included in some bullet points. Our method attempts to generate draft slides of the typical type mentioned above and helps people to prepare their final slides.

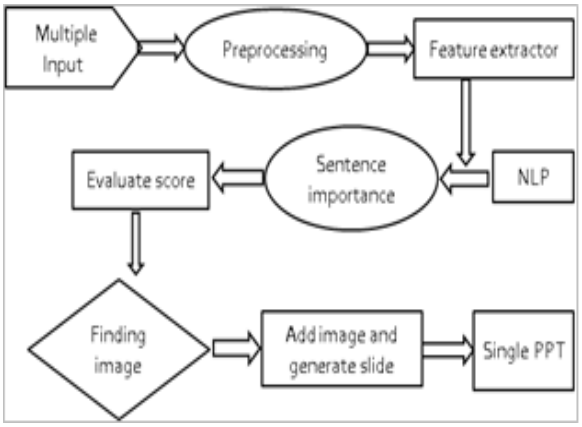

Figure 1. System Architecture (ASAR)

A strategy is proposed given the name as ASAR in Fig 1 for making presentation slides from multi-documents of different extensions based on input query or title that formulate extraction of valuable information and generate slide. A system is proposed which automatically generates slides containing graphical elements as well as text data. we propose a system to automatically generate slides that have good structure and content quality from multiple documents. We use the SVR based sentence scoring model to assign an importance score for each sentence in the given paper, where the SVR model is trained on a corpus collected on the Web. Then, we generate slides from the given text by using ILP. ASAR 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.

In this study, we propose a novel system to generate well-structured presentation slides from multiple documents with different extensions. In our system, the importance of each sentence in a paper is learned by using the Support Vector Regression (SVR) model, and then the presentation slides for the paper are generated by using the Integer Linear Programming (ILP) model to select and align key phrases and sentences.

II. METHODS AND MATERIAL

A primary challenge of the proposed methods is to generate presentation slides automatically, users have no choice about the structure of the presentation, and cannot participate in the contents and the layouts of the slides. In this paper, we investigate the problem of presentation design support and propose a design support system. Users may select topics to design the presentation structures first, and input or search the contents for the topics, then allow these components into different pages, finally decide the layouts of the slides. Different methods and material used for this approach shown below.

A. Sentence Extraction

Sentence Extraction is done based on the topics generated during topic modeling. Sentence Extraction is done to create a summary report for the given documents. Summaries are used to provide an overview of the given documents. Sentence Extraction is based on Support Vector Regression [2].

B. Stop Word Removal

Stop words are language specific functional words which carry no information. It may be of the following types such as pronouns, prepositions, conjunctions. Stopword list is contained in a list which contains the maximum of stop words. Once the file is read it checks the uploaded file with the stop word dataset. If the words in the document and dataset are matched, then the corresponding word is removed from the document [3].

C. Feature Extraction

Feature extraction is used to extract important features such as word frequency, sentence position, word overlap with the title and sentence parse tree information. In this system only, word frequent is mined from the document. Here the maximum number of repeated words in the file is considered for analyzing and is used to know which topics are explained in the paper. In this module, the system gets the non-stop words as input and calculates the count of words and finds the repeated occurrence of each and every word from the non-stop words [7].

D. Support Vector Regression (SVR)

Support Vector Regression is used in maintaining all the main features which characterize the maximal margin of the algorithm. The Support Vector (SV) algorithm is a nonlinear generalization of the Generalized Portrait algorithm developed in Russia in the sixties.

E. Integer Linear Programming (ILP)

Integer Linear Programming is used generate slide from the text which is extracted using feature extractor. It is also used for putting graphical elements such as bullets, points, numbering and putting some images based on user customization.

III. LITERATURE SURVEY

A. PPSGen: Learning-Based Presentation Slides Generation for Academic Papers [1]

In 2014, Yue Hu and Xiaojun Wan proposed "PPSGen: Learning-Based Presentation Slides Generation for Academic Papers". In this paper, we investigate a very challenging task of automatically generating presentation slides for academic papers. The generated presentation slides can be used as drafts to help the presenters prepare their formal slides in a quicker way. A novel system called PPSGen is proposed to address this task. It first employs regression methods to learn the importance of the sentences in an academic paper and then exploits the integer linear programming (ILP) method to generate well-structured slides by selecting and aligning key phrases and sentences. Evaluation results on a test set of 200 pairs of papers and slides collected on the web demonstrate that our proposed PPSGen system can generate slides with better quality. A user study is also illustrated to show that PPSGen has a few evident advantages over baseline methods.

B. A System for Summarizing Scientific Topics Starting from Keywords [2]

In 2014, R. Jha, A. Abu-Jbara and D. Radev proposed "A System for Summarizing Scientific Topics Starting from Keywords". In this paper, we investigate the problem of automatic generation of scientific surveys starting from keywords provided by a user. We present a system that can take a topic query as input and generate a survey of the topic by first selecting a set of relevant documents, and then selecting relevant sentences from those documents. We discuss the issues of robust evaluation of such systems and describe an evaluation corpus we generated by manually extracting factoids, or information units,

from 47 gold standard documents (surveys and tutorials) on seven topics in Natural Language Processing. We have manually annotated 2,625 sentences with these factoids (around 375 sentences per topic) to build an evaluation corpus for this task. We present evaluation results for the performance of our system using this annotated data.

C. Coherent Citation-Based Summarization of Scientific Papers [3]

In 2011, Abu-Jbara and D. Radev proposed "Coherent citation-based summarization of scientific papers" In citation-based summarization, text written by several researchers is leveraged to identify the important aspects of a target paper. Previous work on this problem focused almost exclusively on its extraction aspect (i.e. selecting a representative set of citation sentences that highlight the contribution of the target paper). Meanwhile, the fluency of the produced summaries has been mostly ignored. For example, diversity, readability, cohesion, and ordering of the sentences included in the summary have not been thoroughly considered. This resulted in noisy and confusing summaries. In this work, we present an approach for producing readable and cohesive citation-based summaries. Our experiments show that the proposed approach outperforms several baselines in terms of both extraction quality and fluency.

D. SlidesGen: Automatic generation of slides [4]

In 2009, M. Sravanthi, C. R. Chowdary and P. S. Kumar proposed "SlidesGen: Automatic generation of presentation slides for a technical paper using Summarization". Presentations are one of the most common and effective ways of communicating the overview of a work to the audience. Given a technical paper, automatic generation and helps in creating a structured summary of the paper. In this paper, we propose the framework of a novel system that does this task. Any paper that has an abstract and whose sections can be categorized under introduction, related work, model, experiments and conclusions can be given as input. As documents in LATEX are rich in the structural and semantic information we used them as input to our system. These documents are initially converted to XML format. This XML is parsed and information in it is extracted. A query specific extractive summarizer has been used to generate slides. All graphical elements from the paper are made well use of by placing them at appropriate locations in the slides. These slides are presented in the document order.

E. Slideseer: A Digital Library of Aligned Document and Presentation Pairs [5]

In 2009, M.Y. Kan proposed "SlideSeer: A digital library of aligned document and presentation pairs". Research findings are often transmitted both as written documents and narrated slide presentations. As these two forms of media contain both unique and replicated information, it is useful to combine and align these two views to create a single synchronized medium. We introduce SlideSeer, a digital library that discovers, aligns and presents such presentation and document pairs. We discuss the three major system components of the SlideSeer DL: 1) the resource discovery, 2) the _ne-grained alignment and 3) the user interface. For resource discovery, we have bootstrapped collection building using metadata from DBLP and CiteSeer. For alignment, we modify maximum similarity alignment to favor monotonic alignments and incorporate a classifier to handle slides which should not be aligned. For the user interface, we allow the user to seamlessly switch between four carefully motivated views of the resulting synchronized media pairs.

F. Automatic Slide Presentation from Semantically Annotated Documents [6]

In 2000, M. Utiyama and K. Hasida proposed "Automatic slide presentation from semantically annotated documents". This paper discusses how to automatically generate slide shows. The reported presentation system inputs documents annotated with the GDA tag set, an XML tag set which allows

machines to automatically infer the semantic structure underlying the raw documents. The system picks up important topics in the input document on the basis of the semantic dependencies and coreferences identified from the tags. This topic selection depends also on interactions with the audience, leading to dynamic adaptation of the presentation. A slide is composed of each topic by extracting relevant sentences and paraphrasing them to an itemized summary. Some heuristics are employed here for paraphrasing and layout. Since the GDA tag set is independent of the domain and style of documents and applicable to diverse natural languages, the reported system is also domain/ style independent and easy to adapt to different languages.

G. Identification of Keywords and Phrases: First Review [7]

In 2014, S.V. Shinde and Prof. S.Z. Gawali proposed "Identification of keywords and phrases in text Document and sensing a word for document retrieval and ranking: First Review". In this paper Keywords, phrases, sentences are atomic subatomic molecular levels of a document. A keyword justifies a phrase. A phrase justifies a sentence. Sentences justify paragraph. Single value to group value identif5es a document Relevance of information. Relevance counters the precision of information retrieved by search engines. In this paper, an Analysis search on concept extraction, sensing word and phrases is being done, with an investigation in Document retrieval engines functioning, text mining, ranking algorithm and machine learning methods for acumen information classification. The paper gives a conceptual overview of 20 latest research papers with best methodologies to incorporate in our proposed system. This is investigation report which gives directions of research in GUI development which incorporates NLP understanding to search engine.

H. Technique for Generating Automatic Slides on the Basis of Paper Structure Analysis [8]

In 2016, Ektaa Meshram and D. A. Phalke proposed "Technique for Generating Automatic Slides on the basis of Paper Structure Analysis". In this paper, 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 to lead 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 to the presentation according to the label of an image. An evaluation result shows that the system is more reliable than the existing system.

IV. CONCLUSION

Automatic slide generation for multiple input files is an interesting technique. This method is used to automatically generate slides considering the important point from the docs. Slides provide a better way of understanding rather than document. This system can be further enhanced by using training data in which the slide generated will be effective when compared with traditional methods. It can be further extended by creating slides using

the approach of machine learning. Multiple PPT generations would gain popularity and advantage over the recent future.

We have discussed various automatic ppt generation techniques in detail. There is no single technique that can analyze all type of information. Different parameters are responsible for the accuracy of given technique or solution. Most of the techniques discussed can only be used for the information which is in the English language. For other languages, these techniques cannot be used.

V. REFERENCES

[1]. Yue Hu and Xiaojun Wan, PPSGen, "Learning-Based Presentation Slides Generation for Academic Papers", IEEE Transactions on knowledge and data engineering, vol. 27, no. 4, April 2015.

[2]. R. Jha, A. Abu-Jbara, and D. Radev, "A system for summarizing scientific topics starting from keywords", ACM Compute. Surv, vol. 40, no. 3, p. 8, 2013.

[3]. 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. 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]. M. Utiyama and K. Hasida, "Automatic slide presentation from semantically annotated documents", in Proc. ACL Workshop Conf.Its Appl., 1999, pp. 2530.

[7]. MTech Scholar S. V Shinde, Prof. S.Z. Gawali, "Identification of keywords and phrases in text Document and sensing a word for document retrieval and ranking: First Review", in IJAIEM Volume 3, Issue 5, May 2014, ISSN 2319 4847.

[8]. Ektaa Meshram, D. A. Phalke, "Technique for Generating Automatic Slides on the basis of Paper Structure Analysis", in IJIRSET, Vol. 5, Issue 6, June 2016