International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 2, February 2016

LSG: A Regression Based Approach to Automate LaTeX Slides Generation for Technical Articles

[1]Biju P Dais, [2]Smitha C S
[1] PG Scholar, [2] Assistant Professor in CSE
College of Engineering Perumon (CUSAT)
[1] bijupdais@gmail.com, [2] smithacscs@gmail.com

Abstract- **A data mining approach for automating the generation of presentation slides from an academic article is presented in this paper. Initially, the system is trained using a large dataset to learn the intricacies involved in how humans do the task of slide generation. The input to the proposed scheme is a technical article. The proposed method operates in 2 stages - Scoring and Selection. During the phase of scoring, the sentences in the input are extracted and their importance is analyzed by calculating a relevance score for each, by using a trained Support Vector Regression model. During the phase of selection, an Integer Linear Programming model with a robust objective function and well defined constraints selects important key phrases and the sentences which best summarizes them from the document. The proposed system can include graphical elements as well to the slides. The resultant slides are output in either TeX or PPT editable formats based on user preference. The sentences are also compressed optionally by the system so as to resemble humanly generated slides to a much higher level.**

***Index Terms*— Support Vector Regression, Integer Linear Programming, Text Mining, Summarization**

I. INTRODUCTION

Presentation slides have been widely used over a very long period for the intellectual conveyance of ideas and theories. Most of us rely on the use of presentation of slides because they immensely help us transfer information to a large audience with ease. Researchers propose their theories and hypothesis by writing research manuscripts and present their works using presentation slides. Authors use applications like Microsoft PowerPoint, Open Office, Libre etc to create presentations. But these softwares help us just in creating and formatting the presentations according to our needs. The automatic generation of presentations is beyond the capabilities of these tools.

Researchers very much desire a system that can automatically create presentations slides from their research manuscripts so that they can avoid the tedious process of creation of presentations from scratch. The authors can use the automatically generated presentations as a starting pointing and tune them so as to suit their tastes.

The key idea behind the automatic generation of presentation slides from research articles is to maximally exploit the similarity in their structure. Research articles normally contain some common sections like Abstract, Introduction, and Related Works and so on. So, after the inference of the logical structure of the article, an automatic slide generator can map each section to one or more output slides just like a research person would do.

For automatically generating the presentation slides, the most important concepts and components in the article needs to be identified. The amount of information that has been conveyed on to the slide determines the efficiency of such a slide generation mechanism. It is desired to have a high amount of information in a research manuscript; most slide generators can use a summarization scheme. Different summarization schemes exist in which summarizers either can be extractive or abstractive. Extractive summarizers summarize the article at hand by picking the important sentences as such while abstractive summarizers on the other hand work by rephrasing the selected sentences.

In this study, a new system for automatically generating presentations from a research manuscript is proposed. The system works on 2 stages. In the first stage, a relevance score is predicted for each individual sentence in the article using the model of Support Vector Regression. The score is indicative of how important a sentence is in the article. In the second stage, an Integer Linear Programming system backed up strictly defined constraints are used to construct the contents that would eventually be used to create the slides. For each section, using the ILP model, the system identifies the important key phrases and selects out the sentences that best summarize that section. Since the inclusion of graphical elements greatly enhances the understandability of the presentation, the proposed system extracts images from the manuscript and adds them to the appropriate sections in the output presentation.

Moreover, to match humanly generated slides to a greater level, the system compresses the selected sentences. The presentations can be generated either in LaTeX or Powerpoint formats based on user preference.

The rest of the paper is organized as follows: Section-2 discusses some of the related works in the area of slide generation, Section-3 describes the proposed methodology and finally Section-4 concludes the article.

II. RELATED WORKS

Various schemes have been proposed to automate the generation of presentation slides from an article source. In this section, a survey of some of the techniques for automatically generating presentations from articles is discussed.

Utiyama et al., [1] proposed a new scheme which used the GDA tagset annotated version of an article to learn the semantic structure of the article. This information was used for finding the important topics. Sentences corresponding to these topics are extracted which were then organized on to the output slides.

Yasamura et al., [2] implemented a solution to generate presentation from the LaTeX manuscript of a technical article. The TF-IDF scoring scheme was used to calculate the weights of all terms in the article so as to find out relevance score for all document objects. The term weights were used to determine the size of each section summary. The output slides were customizable by the user.

Sravanti et al., [3] elaborated a system to auto generate presentation slides from a research manuscript. Here also, the starting point of slide generation is from the raw LaTeX source of the research manuscript. After the inference of the logical structure from the article, each section was categorized to fall under Introduction, Related Works, Model, Experiments and Conclusion respectively. The process of automatic slide generation in this technique involved the use of QueSTS summarizer [4]. Graphical elements could also be extracted from the article by the system and the slides were built.

Shibata et al., [5] described a method for the generation of presentation slides from an article by the analysis of the discourse structure of the article. A clause and sentence was considered as a discourse unit by the system and the important coherence relations such as contrast, list, additive, elaboration etc were extracted and analyzed. Topic and non topic parts were identified using the discourse structure of the text. The output slides were produced by having proper intends to the contents so as to enhance readability. The sentences are connected to the most similar preceding sentences. Parts having less importance are pruned based on some heuristic measures.

K. Gokul Prasad et al., [6] proposed a new scheme to create presentation slides for seminars and lectures. The 2 modules – Information Extractor and Slide Generator extracts the text contents from the article and hence uses common NLP operations of text segmentation and chunking to identify the noun phrases and segments. The system constructed an ontology tree for each noun phrase detected using a chunker system. The ontology and weight values calculated were used for aligning key phrases and contents for bullet points and hence, the presentations were generated.

Tulasi Prasad Sariki et al., [7] presented a novel scheme to generate presentation slides by initially fetching the document for which the slides were to be generated. The system then implemented various basic preprocessing techniques such as sentence division, case folding, stop word removal, stemming and lemmatization to the document. Individual sentences were considered and a combination of popular baseline summarizers was used to find relevance score for each sentence. The system is capable of accepting keyword queries and building a presentation specific to the input query.

ShaikhMostha Al Masum et al., [8] elaborated a new agent based scheme where in the user could give queries as input. In the background the system collected information about the query by searching the internet. Images could also be added to the output slides by the system. The system worked on various techniques like web data fetching, web page parsing and summary extraction. Each operation was done by agents. Specific algorithms were used for web data fetching and parsing. In the post processing phase, MPML scripts were generated and the output slides were formed in HTML and Javascript formats and the topics were explained to different headings by agent characters.

Mistsuru Ishizuka et al., [9] discussed a new scheme similar to [8] by accepting keywords from the user and generating a concise report and presentation by querying the internet. The system worked on different steps, each of which completed by software agents. If the keywords were ambiguous, the disambiguated senses were also added to the search keys. The summarization scheme used a vector distance for measuring the closeness between sentences. The system generated a report specific to each topic and from each report, a presentation was built.

Yue Hu et al., [10] approached the task of automatic slides generation by elaborating a scheme that followed a corpus based machine learning approach. The system worked in 2 phases to generate slides from a research article. The system initially predicted an importance score for each of the individual sentences in the article. The score lies in the range [0, 1]. Now, based on the predicted scores, an Integer Linear Programming Model is exploited so as to select the contents that would be arranged on to the output slides. For each section, important keyphrases and corresponding sentences are extracted in creating the slides. A phrase is likely to be selected if it occurred in many sections. The slides that were produced acted as draft slides to the user and reduced the difficulty in creating the slides from scratch.

III. THE PROPOSED SCHEME

In this section, a new system is proposed for automatically generating presentation slides from technical articles. After the system is successfully trained, it works on the basis of mainly 2 core stages – Scoring and Selection.

A. The Training Stage

Corpus Setup: A very large dataset consisting of pairs of research articles and their corresponding slides are downloaded from the internet. For this purpose, www.aminer.org is crawled to get the homepages of different authors. From the homepages, research articles and slides are downloaded and added to the corpus. To construct the training dataset, for each pair, the system does the following steps:

➢ Extract the sentences from the paper and slide separately.

➢ Calculate the maximum value of cosine similarity value
for each article sentence with slide sentences to get the training label for that sentence.

Each sentence in the article is represented using various features as follows:

➢ Cosine similarity values with the article title, section, subsection and subsubsection headings.

➢ Word overlap with the article title, section, subsection
and subsubsection headings.

➢ Sentence Position – The position of a sentence in a Section

➢ The number of Verb Phrases and Noun Phrases in a sentence.

➢ The percentage of stop words in percentage.

➢ The length of the sentence, number of words after Removing the stop words.

After scaling the training dataset to [-1, 1], the SVR Model[13] for score prediction is trained and generated. LibSVM with RBF Kernel[12] is used for implementing the Support Vector Regression Model.

➢ Support Vector Regression

The support vector regression follows the basic working methodologies of the Support Vector Machine. The main difference between SVR and SVM lies in the output values - in the case of SVR, the output is a real number, making the prediction process very difficult. The SVR algorithm is also a much more complex version of SVM. The key idea behind SVR and SVM remains the same - to minimize the error, individualizing the hyperplane that is used to maximize the margin, keeping in mind that part of the error is tolerated.

B. The Working Stage

The working stage of the system is the phase where we automatically generate a presentation for a research manuscript. The input to the system is a research article in Portable Document Format. The output presentations can either be in LaTeX of PPT formats.

The overall architecture of the system is shown in Fig-1:

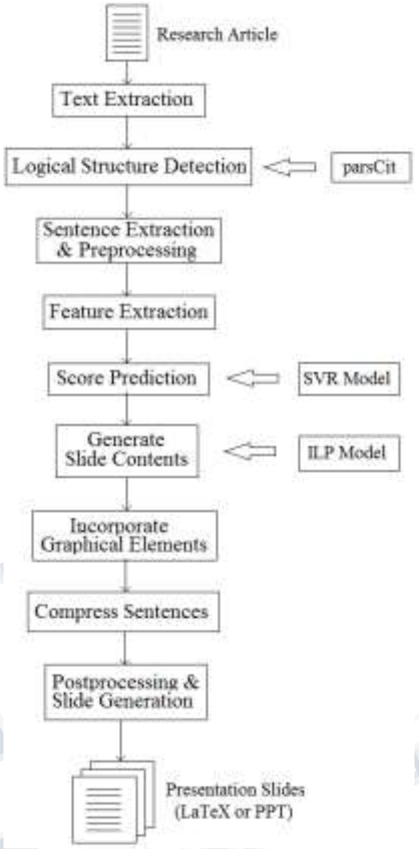
Research Article
Text Extraction
Logical Structure Detection
parsCit
Sentence Extraction & Preprocessing
Feature Extraction
Score Prediction
SVR Model
Generate Slide Contents
ILP Model
Incorporate Graphical Elements
Compress Sentences
Postprocessing & Slide Generation
Presentation Slides (LaTeX or PPT)

Fig: 1 Architecture of the system

The working stage of the proposed scheme works sequentially by the following steps:

➢ ***Input and Text Extraction:*** In this step, the input is specified as a Portable Document Format file. The system extracts the text contents of the research manuscript and is saved as a text file.

➢ ***Logical Structure Detection***: The system infers the logical hierarchy of sections, subsections, subsubsections by using parsCit[13] . The input to parsCit is the text contents of the manuscript. The output of parsCit is an XML file with information on the logical structure of the article.

➢ ***Sentence Extraction and Preprocessing***: With the help of the inferred logical structure, the system extracts the individual sentences with information on their positional details i.e., to which section/ subsection/ subsubsection they belong. After this process, common preprocessing techniques like stemming and stop word removal are done on all sentences.

➢ ***Feature Extraction***: The system considers each individual sentence and extracts its features. This forms the test dataset and is analyzed in the next phase.

➢ ***Score Prediction***: The test dataset is analyzed by the SVR Model and an importance score is predicted for each sentence. The score indicates the level of importance of a sentence. The score value lies between 0 and 1. Higher the score, higher the probability for that sentence to be selected to the slides.

- ***Content Selection***: For selecting the contents to be included to the slides, the system uses an Integer Linear Programming Model with a robust objective function having well defined constraints. In this module, the system collects global phrase in a particular section is known as a local phrase. Instead of just placing some extracted sentences as a summary for a section/ subsection, the system identifies the most important key phrases using the ILP Model since it would have more emphasis on the actual concept portrayed by the particular section.

The ILP Model Objective function and the constraints are as follows:

$$\max_{lp,x} \lambda_1 \sum_{i=1}^{n} \frac{l_i}{L_{max}} w_i x_i + \lambda_2 \sum_{i=1}^{|B|} \frac{c_{b_i} b_i}{|B^*|} + \lambda_3 \sum_{i=1}^{n} \frac{w_i}{n} y_i$$

The first part maximizes the importance score of the slides, the second part ensures the diversity of the selected sentences and makes the slides unbiased and the third part of the function ensures that the selected key phrases have maximum coverage.

The constraints of the ILP Model are as listed as follows:

$$\sum_{i=1}^{n} l_i x_i \leq L_{max}, \qquad (1)$$

$$\sum_{lp_j \in LP_i} lp_j \geq x_i, \; for\; i = 1, \ldots n, \qquad (2)$$

$$\sum_{s_i \in S_j} x_i \geq lp_j, \; for\; j = 1, \ldots |LP|, \qquad (3)$$

$$\sum_{lp_j \in LP_k} lp_j \geq y_k, \; for\; k = 1, \ldots n, \qquad (4)$$

$$\sum_{b_m \in B_i} b_m \geq |B_i| x_i, \; for\; i = 1, \ldots n, \qquad (5)$$

$$\sum_{s_i \in S_m} x_i \geq b_m, \; for\; m = 1, \ldots |B|, \qquad (6)$$

$$\sum_{lp_j \in GP_t} lp_j \geq gp_t, \; for\; t = 1, \ldots |GP|, \qquad (7)$$

$$gp_t \geq lp_j, \; for\; \forall lp_j \in GP_t; t = 1, \ldots |GP| \qquad (8)$$

$$\sum_{i=1}^{|GP|} gp_i * 2 \leq \sum_{i=1}^{n} x_i, \qquad (9)$$

$$x_i, lp_j, y_k, b_m, gp_t \in \{0, 1\}, \forall i, j, k, m. \qquad (10)$$

Where,

w_i: Importance score of sentence s_i

n: Number of sentences

l_i: Length of sentence s_i

x_i: Whether sentence s_i is included in slides

lp_j: If local phrase lp_j is included in slides

gp_t: Whether global phrase gp_t is included in slides

GP_t: The set of local phrases relevant to global phrase gp_t

Y_k: If sentence s_k contains atleast one selected local phrase

B_m: If bigram b_m is included in slides

L_{max}: Maximum length of slides

C_{bi}: Count of bigram bi in the article

B:* Total set of bigrams in the paper

B: Set of unique bigrams

LP: Set of local phrases

GP: Set of global phrases

Constraint - 1: Ensures total word count of the slides is less than L_{max} .

Constraint - 2 : If sentence s_i is selected, atleast one local phrase in it must be selected.

Constraint - 3: If a local phrase lp_j is selected, at least one sentence relevant to phrase lp_j must be selected.

Constraint - 4: y_k must be set to 1 if at least one local phrase in Lp_j is selected. Else, *yk* must be set to 0.

Constraints - 5, 6: If sentence s_i is selected, all its bigrams must be selected. If s_i is not selected, some bigrams in s_i may still be selected. If b_m is selected, at least one sentence in S_m(sentences containing bigram b_m) is selected.

Constraint - 7: If a global phrase is selected, at least one relevant local phrase should be selected. If a global phrase is not selected, no corresponding local phrases should be selected.

Constraint - 8: If a local phrase is selected, its corresponding global phrase must be selected. If a local phrase is not selected, the corresponding global phrase can still be selected due to its other local phrases.

Constraint - 9: It ensures that the total number of global phrases selected is less than half the number of local phrases.

Constraint - 10: Ensures that all the variables involved have integer values.

IV. CONCLUSION

This paper proposed a new system called LSG to automatically generate presentation slides from a research article. The system used a trained SVR model to predict the importance of a sentence and uses an ILP Model to select the keyphrases and sentences, Graphical elements can also be incorporated to the output slides and the final presentation is produced in either TeX or PPT formats. The future works of this scheme includes the consideration of drawings in the article also while preparing the slides. Multiple similar documents can also be used as input for generating slides so as to make the slides richer in information, thereby boosting the quality of the output.

REFERENCES

[1] M. Utiyama and K. Hasida, ”Automatic slide presentation from semantically annotated documents”, in Proc. ACLWorkshop 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] 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. 284-289.

[4] M. Sravanthi, C. R. Chowdary, and P. S. Kumar, ”QueSTS: A query specific text summarization approach”, in Proc. 21st Int. FLAIRS Conf., 2008, pp. 219-224.

[5] 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.

[6] Gokul Prasad, K., Mathivanan, H., Jayaprakasam, M., and Geetha, T. V., ”Document summarization and information extraction for generation of presentation slides”, Advances in Recent Technologies in Communication and Computing, 2009. ARTCom’09. International Conference on. IEEE, 2009.

[7] Sariki, Tulasi Prasad, Bharadwaja Kumar, and Ramesh Ragala. ”Effective Classroom Presentation Generation Using Text Summarization”.

[8] S. M. A. Masum, M. Ishizuka, and M. T. Islam, ”Autopresentation: 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.

[9] 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.

[10] Hu, Yue, and Xiaojun Wan. ”Ppsgen: learning to generate resentation slides for academic papers”, Proceedings of the Twenty-Third international joint conference on Artificial Inelligence. AAAI Press, 2013

[11] V. Vapnik, Statistical Learning Theory. Hoboken, NJ, USA: Wiley, 1998.

[12] C. C. Chang and C. J. Lin. (2001), LIBSVM: A library for support vector machines, [Online]. Available http://www.csie.ntu.edu. tw/ cjlin/libsvm

[13] Minh-Thang Luong, Thuy Dung Nguyen and Min-Yen Kan (2010) Logical Structure Recovery in Scholarly Articles with Rich Document Features. International Journal of Digital Library Systems (IJDLS), 1(4), 1-23.