ISSN:2348-2079

Volume-5 Issue-1

International Journal of Intellectual Advancements and Research in Engineering Computations

Automatic generation of presentation slides for academic papers

N.Yogapreethi, PG Scholar[1] & S.Maheswari, Associate Professor[2]
Department of Computer Science & Engineering,
Nandha Engineering College,
Erode-638052, India.
yogapreethi94@gmail.com, maheswari.s@nandhaengg.org

Abstract— Generally to get the slides from begin, it takes plenty of your time. These slides contains information concerning base paper objective from abstract details of system i.e. introduction and to boot approach utilized, writing audit from connected work section, alpha results and conclusions from base paper. The generated slides will be used as rough plan for any preparation. This helps presenters in making ready their formal slides in quicker manner. Some rough structure for slide shows from papers capable to save lots of the author abundant time once organizing shows. during this paper we tend to investigate totally different perspective for educational papers slide generation. to put in writing the slides from scratch takes plenty of your time of presenter. they typically contain many sections like abstract, introduction, connected work, projected methodology, experiments and conclusions. to keep up individuation in making ready slides this concept is crucial and distinctive. every section from the tutorial paper is known and is aligned to 1 or a lot of slides. each bullet purpose are mapped with the slide heading purpose. Out of the many sentences below that within that heading sentences importance is calculated thus on keep those because it is within the slides.

Keywords— Abstracting methods, Integer Linear Programming, Support Vector Regression model, text mining.

I.INTRODUCTION

Presentation slides are the one in every of necessary approach of learning. during this approach of learning get the response from the listeners they'll be students, staff or client etc. Presenters produce their go along victimization the code tools like Microsoft Power- purpose, Open workplace etc.All these code presenters ought to sort the content into the slide then it will be long task. during this projected methodology automatic slides are generated consistent with the sections within the paper i.e, titles in tutorial paper and corresponding relevant sentences from identical paper. It helps users in obtaining a rough structure of the tutorial paper.

It continually have the same structure. they typically contain many sections like abstract, introduction, connected work, projected methodology, experiments and conclusions. though presentation slides will be written in numerous ways in which by totally different presenters, a presenter, particularly for a beginner, continually aligns slides consecutive with the paper sections once making ready the slides. every section is aligned to 1 or a lot of slides and one slide sometimes contains a title and a number of other sentences. These sentences is also enclosed in some bullet points. Our methodology tries to get draft slides of the standard sort mentioned higher than and helps individuals to arrange their final slides. Automatic slides generation for educational.

papers could be a terribly difficult task. Current strategies typically extract objects like sentences from the paper to construct the slides. In distinction to the short outline extracted by a summarization system, the slides are needed to be rather more structured and far longer. Slides will be divided into an ordered sequence of components.

Each part addresses a particular topic and these topics also are relevant to every different. typically speaking, automatic slide generation is far harder than summarization. Slides sometimes not solely have text components however additionally graph components like figures and tables. however our work focuses on the text components solely. during this study, we tend to propose a novel system referred to as PPSGen to get well-structured presentation slides for educational papers. In our system, the importance of every sentence in a very paper is learned by victimization the Support Vector Regression (SVR) model, and so the presentation slides for the paper area unit generated by victimization the integer linear programming (ILP) model to pick out and align key phrases and sentences.

The rest of this paper is organized as follows. connected work is introduced in section 2. we tend to describe our methodology well in section 3. we tend to show the experiment leads to section 4 and conclude our add section 5.

II.RELATED WORK

Automatic slides generation for tutorial papers remains way under-investigated today. Few studies directly analysis on the

subject of automatic slides generation. In M. Sravanthiet. al. [1] introduces the answer for reducing the hassle of the presenter and facilitate them in making a structured outline of paper. It helps in making slides for presentation with vital purpose and al necessary figure etc. The very important points of the paper can mention. The Latex document is provided because the input and born-again into xml format. The xml file can break down and extract the knowledge. a question specific extractors can facilitate to summarize and generate the slides.

Utiyama et al., [2] planned a brand new theme that used the GDA tagset annotated version of an editorial to find out the linguistics structure of the article. This data was used for locating the vital topics. Sentences similar to these topics are extracted that were then organized on to the output slides. Yasamura et al., [3] enforced an answer to get presentation from the LaTeX manuscript of a technical article. The TF-IDF evaluation theme was accustomed calculate the weights of all terms within the article thus on establish relevancy score for all document objects. The term weights were accustomed confirm the dimensions of every section outline. The output slides were customizable by the user.

Sravanti et al., [4] detailed a system to machine generate presentation slides from a quest manuscript. Here also, the start line of slide generation is from the raw LaTeX supply of the analysis manuscript. once the logical thinking of the logical structure from the article, every section was classified to be Introduction, connected Works, Model, Experiments and Conclusion severally. the method of automatic slide generation during this technique concerned the utilization of QueSTS summarizer [5]. Graphical parts may even be extracted from the article by the system and also the slides were engineered.

Shibata et al., [6] represented a technique for the generation of presentation slides from an editorial by the analysis of the discourse structure of the article. A clause and sentence was thought of as a discourse unit by the system and also the vital coherence relations like distinction, list, additive, elaboration etc were extracted and analyzed. Topic and non topic components were known victimization the discourse structure of the text. The output slides were created by having correct intends to the contents thus on enhance readability. The sentences are connected to the foremost similar preceding sentences. components having less importance are cropped supported some heuristic measures. K.

Gokul Prasad et al., [7] planned a brand new theme to form presentation slides for seminars and lectures. the two modules – data Extractor and Slide Generator extracts the text contents from the article and thence uses common IP operations of text segmentation and constellation to spot the noun phrases and segments. The system made an ontology tree for every phrase detected employing a chunker system. The metaphysics and weight values calculated were used for positioning key phrases and contents for bullet points and thence, the shows were generated.

Tulasi Prasad Sariki et al., [8] given a completely unique theme to get presentation slides by ab initio taking the document that the slides were to be generated. The system then enforced numerous basic preprocessing techniques like sentence division, case folding, stop word removal, stemming and lemmatization to the document. Individual sentences were thought of and a mix of in style baseline summarizers was accustomed realize relevancy score for every sentence. The system is capable of acceptive keyword queries and building a presentation specific to the input question.

ShaikhMostha Al Masum et al., [9] detailed a brand new agent based mostly theme wherever within the user may offer queries as input. within the background the system collected data regarding the question by looking the net. pictures may even be additional to the output slides by the system. The system worked on numerous techniques like net information taking, website parsing and outline extraction.

Mistsuru Ishizuka et al., [10] mentioned a brand new theme by acceptive keywords from the user and generating a compact report and presentation by querying the net. The system worked on completely different steps, every of that completed by code agents. If the keywords were ambiguous, the disambiguated senses were additionally additional to the search keys. The summarisation theme used a vector distance for measurement the closeness between sentences. The system generated a report specific {to every|to every} topic and from each report, a presentation was engineered.

Yue Hu et al., [11] approached the task of automatic slides generation by elaborating a theme that followed a corpus based mostly machine learning approach. The system worked in two phases to get slides from a quest article. There are some limitations and challenges in previous strategies, e.g. Extraction of image and tables from the bottom paper. to beat those we have a tendency to planned a technique that selects range of vital sentences, images, tables and also the phrases from the corresponding base paper. Calculation of the importance of sentence with image reference is difficult task.

.

III.METHODOLOGY

In this paper, we have a tendency to propose a system to mechanically generate slides that have sensible structure and content quality from educational papers. The design of our system is shown in Figure 1. we tend to use the international intelligence agency based mostly sentence marking model to assign associate degree importance score for every sentence within the given paper, wherever the international intelligence agency model is trained on a corpus collected on the online. Then, we tend to generate slides from the given paper by victimization ILP. additional details of every half are going to be mentioned within the following sections.

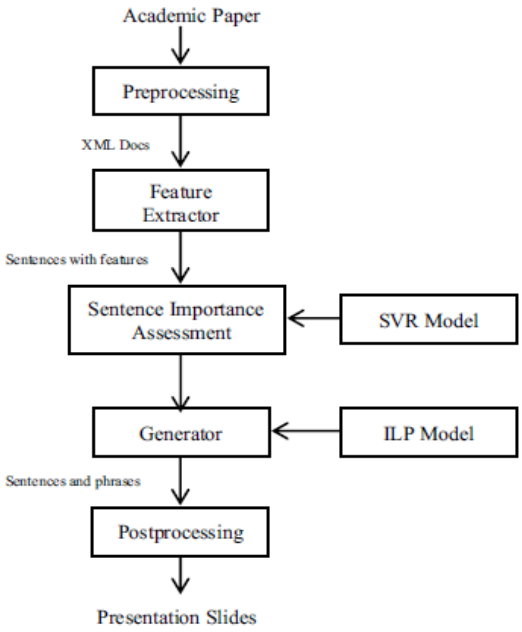

Fig 1:Proposed System Architecture Corpus And Preprocessing

To learn however humans generate slides from educational papers, we tend to build a corpus that contains pairs of educational papers and their corresponding slides. several researchers within the technology field place their papers and also the corresponding slides along in their homepages. The homepages' URLs are obtained by creeping Arnetminer2 . when downloading the homepages, we tend to use many strict patterns to extract the links of the papers and also the associated slides and transfer the files to create the dataset. we have a tendency to collect over two thousand pairs.

After cleanup up the inaccurate pairs, we've got 1200 paper-slides pairs. The papers are beat PDF format and also the slides are in either PDF or PowerPoint format. For the papers, we tend to extract their texts by victimization PDFlib3 and find their physical structures of paragraphs, subsections and sections by victimization ParsCit4 . A custom XML format is employed to explain this structure. For the slides, we tend to conjointly extract their texts and physical structures like sentences, titles, bullet points, etc. we have a tendency to use xpdf5 and also the API provided by Microsoft workplace to subsume the slides in PDF and PowerPoint formats, severally. The slides are remodeled to a predefined XML format in addition.

We utilize the Stanford NLP library for sentence parsing. for creating slides it's important to identify significance of every sentence from the created archive. Support machine regression (SVR) model is used to calculate the importance of every sentence within the document. Sentence importance score is calculated and matrix is generated for every sentence. SVR is used to arrange and absorb the sentence imperative score. we have a tendency to need foreseeing the importance score of every sentence to form slide for tutorial paper presentation. SVR model is best than classification model as a result of regression score is far finer to use for vital sentence choice strategy for generating slide than a classification which supplies coarse score.

By regarding generated matrix sentence significance score is enumerable that later provided as an input to the integer linear programming (ILP) summarizer. ILP summarizer module plays very important role in summarizing these sentences thus on find solely important topics connected with each sentence. In our system, Support Vector Regression (SVR) model is employed to find out the importance of every sentence in an exceedingly paper, so integer linear programming (ILP) model is employed to pick out and align key phrases and sentences for generated slides. Here we have a tendency to are victimization two algorithms for process input file to get slide are SVR model and ILP technique.

A.Svr

1) Sentence Position: Here position of sentence is figured exploitation condition $SP(s) = position(s, doc(s))/|d(s)|$, wherever position(s, doc(s)) is that the sentence request of sentence s in its record doc(s), and $|d(s)|$ is that the assortment of sentences in doc(s).

2) Similarity:

Where Sij = similitude (featurei , featurej). On the off chance that there is no closeness between components (Sii = 1, Sij = 0 for $i \neq j$), the given condition is identical to the customary cosine similarityformula.

3) Word Overlap: It is variety of words shared by the input question sentence q with sentence s take into account finding matching. this is often accomplished by removing stop words and redundant words from each q and s.

B. Ilp

Important points known by ILP are currently required to be highlight and convert in to slides format. Post process of the document is finished supported the desired slide format for all the vital topics and slides are generated. It constructs summaries by minimizing their combine wise similarity and increasing the importance of the chosen sentences, as shown below that of the shape.

$$\max x, y \sum_{i=1}^{n} imp(si).xi - \sum_{i=1}^{n} \sum_{j=i+1}^{n} sim(si, sj).yi, j \quad (1)$$

$$\sum_{i=1}^{n} li.xi \leq L_{\max}$$

For (i = 1, …. n & j = i + 1, … n) Where

yi,j – xi ≤ 0

yi,j – xj ≤ 0

yi +xi – yi,j ≤ 1, where n is the Number of sentences in the input documents, imp(Si) is the Importance score of sentence Si, li is the Length of Si sim(Si ,Sj) is the Similarity of sentences Si and Sj and Lmax is the Maximum allowed length The Xi variables are represented in binary. This variable indicates whether or not the corresponding sentences Si is an element of the outline. The yi,j variables, in binary kind provides plan whether or not each Si and Sj are a part of the outline. The slides created by our technique clearly have most popular general quality over the baseline methods. Finally, complete content and structure redaction before data formatting. Please note of the subsequent things once proofreading orthography and descriptive linguistics.

The object operate contains 3 elements. the reason of every half is below: - the primary half maximizes the importance score of the generated slides. It sums the importance a lot of the chosen sentences. instead of merely shrewd the total of the scores, we have a tendency to add the sentence length as a multiplication consider order to penalise the terribly short sentences. - The second half maximizes the entire weights of the bigrams within the paper that conjointly seem within the slides. The intuition is that once additional bigrams are gift within the slides, the sentences within the slides are less redundant. may be thought to be the burden of the written word. we tend to try and embody additional vital bigrams within the slides. - The last half aims to maximise the weighted coverage of the key phrases chosen.

A sentence is covered by a phrase once this sentence contains the phrase. High-quality slides ought to cover the content within the paper the maximum amount as potential. we have a tendency to describe this sort of coverage by victimization the total of the a lot of the sentences that contains the chosen key phrases.

The ILP model is applied to the full paper once and also the technique doesn't assign the amount of slides for every section expressly. victimization the ILP model, we will obtain the aligned key phrases and sentences to be enclosed within the slides. The titles of slides area unit set by victimization the titles of the corresponding sections. we tend to solve the on top of improvement downside by victimization the IBM CPLEX optimizer seven. It usually takes concerning 10 seconds to resolve the matter. Then the draft slides are generated by victimization the API provided by Microsoft office.

IV. Result and Discussions

In order to line up our experiments, we tend to divide our dataset that contains 1200 pairs of paper and slides into 2 parts: 1000 pairs for coaching and also the different two hundred pairs for check. A SVR regression model with the RBF kernel in LIBSVM is trained on the coaching information and applied to the check information. Then the ILP model is employed to get the slides. the utmost word count of the slides is just set to fifteen proportion of the entire word count of the paper. The parameter values of our technique are by trial and error set to 0.3, 0.4 and 0.3, severally. The comparison results over ROUGE metrics are bestowed in Table one and Table a pair of. Table one shows that our projected technique will improve ROUGE scores, i.e., higher content quality within the 1st analysis approach. It means our slides are richer in content and far additional kind of like the human-written slides than those of the baselines as an entire. Table a pair of shows that our technique will principally bounce back ROUGE scores in every half comparison. It means our slides also can come through higher content quality after they are divided into continuous elements and corresponding elements are compared, i.e., the content texts in our technique are additional effectively distributed into totally different sections. we have a tendency to acquire the importance score of a sentence by learning from the human slides. Therefore, the importance scores are additional credible. Moreover, the alignment between key phrases and sentences is additionally helpful to boost the content, as a result of the sentences relevant to such key phrases will be thought-about to be additional vital. we

have a tendency to distinguish international phrase from native phrase. Sentences that contain additional vital international phrases will be relevance be additional vital.

Using international phrase will cause a far better choice of key sentences, that conjointly end in higher content quality. Table three proves our method's enhancements are statistically vital, and also the T-Test values we tend to get are all way smaller than zero.05. Figure a pair of presents the influences of ROUGE scores once calibration the parameters , and . we have a tendency to set the total of the 3 parameters to at least one, and therefore we have a tendency to really want to vary 2 of the 3 parameters. {we can|we will|we are able to} see that once the parameters are set in an exceedingly big selection of values, our technique can do high ROUGE scores. Our system usually performs higher than those baselines. {we can|we will|we are able to} conjointly see that each one the 3 elements in our ILP model are useful to induce a far better content quality for the generated slides. with none a part of them, the results can deteriorate. Table four shows the typical scores rated by human judgers for every technique. The slides generated by our technique clearly have higher overall quality than those of different strategies. Being in keeping with the automated analysis results, our slides are thought of to possess higher content quality in line with human judges. Moreover, thanks to the indent structure and also the alignment between phrases and sentences, the structure of our slides is additionally judged to be far better than the baselines' slides. Overall, the experimental results indicate that our technique will generate far better slides than the baselines in each automatic and human evaluations [12].

TABLE I

ROUGE F-MEASURE SCORES OBTAINED IN THE FIRST WAY

Method	Rouge-1	Rouge-2	Rouge-SU4
Yoshiaki *et al.*	0.38859	0.11624	0.16424
Random Walk	0.39421	0.11555	0.16463
Mead	0.38778	0.11803	0.16239
Our Method	**0.41342**	**0.13067**	**0.17502**

TABLE II

ROUGE-1 F-MEASURE SCORES OBTAINED IN THE SECOND WAY

Method	Rouge 1			
	First 30%	Mid 40%	Last 30%	Avg
Yoshiaki *et al.*	0.28220	0.30732	0.28411	0.29121
Random Walk	0.29241	0.30661	0.28443	0.29448
Mead	0.31132	0.28063	0.25481	0.28225
Our Method	**0.30235**	**0.32662**	**0.29911**	**0.30936**

TABLE III

T-TEST P-VALUES BETWEEN EACH BASELINE METHOD AND OUR METHOD

System (tested by our method)	T-Test		
	Rouge-1	Rouge-2	Rouge-SU4
Yoshiaki *et al.*	2.492E-13	2.749E-08	1.015E-06
Random Walk	5.329E-06	6.976E-11	1.125E-06
Mead	2.176E-11	1.601E-11	6.331E-08

TABLE IV
AVERAGE RATING SCORES OF JUDGES

Judge	Method	Structure	Content	Overall
avg	Yoshiaki *et al.*	3	2.68	2.71
	Mead	2.7	2.53	2.39
	Our Method	**3.69**	**3.45**	**3.58**

V. CONCLUSION

This paper proposes a novel framework called PPSGen to generate presentation slides from academic papers. We train a sentence scoring model based on support vector regression and use the integer linear programming method to adjust and concentrate key expressions and sentences for producing the slides. Test comes about demonstrate that our strategy can generate much better slides than traditional methods. In future work, we will enhance our framework by utilizing both content and graphical components in the paper and make slides more understandable and clear. When managing the graphical components, we have to recognize the graphical elements in the paper first. The relationship between the content components and the graphical components likewise should be identified. We have to know which sentences are most pertinent to a graphical component and which graphical components ought to be chosen to create the slides. We can utilize govern based strategies or machine learning based techniques to take care of the above issues. At that point we can essentially join the tables and figures. we select to the most important sentences in the slides.

REFERENCES

[1].M. Sravanthi, C. RavindranathChowdary and P. Sreenivasa Kumar,"SlidesGen: Automatic Generation of Presentation Slides for a Technical Paper Using

Summarization",in Proceedings of the TwentySecond International FLAIRS Conference [2009]

[2].M. Utiyama and K. Hasida, "Automatic slide presentation from semantically annotated documents", in Proc. ACLWorkshop Conf. Its Appl., [1999], pp. [25]-[30].

[3].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].

[4].M. Sravanthi, C. R. Chowdary, and P. S. Kumar, "SlidesGen: Automatic generation of presentation slides for a technical paper using summarization", in Proc. [22]nd Int. FLAIRS Conf., [2009], pp. [284]-[289].

[5].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].

[6].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].

[7].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].

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

[9].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].

[10].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].

[11].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]

[12].Sravanthi, M., C. Ravindranath Chowdary, and P. Sreenivasa Kumar. "SlidesGen: Automatic Generation of Presentation Slides for a Technical Paper Using Summarization." In FLAIRS Conference. [2009]