International Journal of Computer Applications (0975 - 8887)
Volume 131 - No.17, December 2015

A Comparison on Techniques for Automatic Generation of Presentation Slides

Biju P. Dais
P G Scholar
Department of Computer Science & Engineering
College of Engineering, Perumon(CUSAT), Kerala, India

Smitha C.S.
Assistant Professor in CSE
Department of Computer Science & Engineering
College of Engineering, Perumon(CUSAT), Kerala, India

ABSTRACT

The automatic generation of presentation slides from technical articles is one of the most desired but under-researched area in the field of computing. Automated generation of slide contents from technical articles is much difficult than a typical text summarization process, since it requires the identification of all the crucial contents from the article and their arrangement in a systematic manner, thus making it a non trivial task. The process is considered to be one of the core applications of text mining. Automatic slide generators can be broadly classified based on NLP, Statistical Methods and Machine Learning. A detailed review of some of the most important automatic slide generation techniques from academic articles is presented and a brief comparison among the discussed techniques is given.

General Terms

Text Mining, Text Summarization

Keywords

Natural Language Processing, Machine Learning, Web Mining

1. INTRODUCTION

Presentation slides have been widely used over a very long time for the abstract conveyance of ideas. It helps users to transfer their ideas and theories effectively. Researchers present their hypothesis with the help of presentation slides and hence, presentations have proved themselves to be extremely beneficial in meetings and conferences. Tools like Microsoft PowerPoint, OpenOffice, Libre etc help in building and formatting the slides according to one's adequate needs. The authors themselves have to create the slides from scratch, which is a tedious and time consuming procedure. The self generation of slides is beyond the capabilities of the above softwares, leading to the need of tools that does automatic creation of slides from technical articles, which is extremely difficult and non-trivial. Since automation in almost all aspects has become a key trend today, the current computational capabilities stress on complete replacement of the manual procedure of slide generation.

Research articles have a more or less similar structure. Almost all the technical articles have sections like Abstract, Introduction, Related Work etc. The main strategy behind an automated slide generator is to exploit this similarity and hence map each section to one or more slides in the presentation. For this purpose, different layouts may be used by the generator. One commonly used layout structure is to maintain the logical structure of the research article and offer a summarization for each section in the output slides.

To generate the slide contents, automated slide generators rely heavily on text summarizers. They can be classified based on their working strategies. The summarization strategy of a generator can work based on statistics, which may use the TF-IDF scoring method, Natural Language Processing, which analyses the discourse structure of the article, Machine Learning which depends on domain knowledge acquired through extensive training. The summarizers can be extractive or abstractive. Extractive summarizers build a summary by just picking up the important sentences without modification from an article. Multi-document extractive summarizers adopt a greedy strategy to build summaries from related documents by selecting sentences having high relevance factors. On the other hand, abstractive summarizers build summaries by either shortening or rewording the selected sentences.

It is possible to generate the summaries that may be organized on to the slides using the citation information from research articles. [1] and [2] are summarization schemes that exploit the citation information from scholarly articles. The purpose of using the citation information is to gather important concepts, which can be used to extract the relevant sentences from the article. [3] and [4] discussed schemes wherein the summaries were generated by extracting elements from multiple related articles. It enhances the diversity of the generated summaries. A Query specific summarizer [5] involve the summarization of an article with respect to the keywords specified by a user. Other summarization schemes involve the use of Hidden Markov Models [6] and Conditional Random Fields [7] for document summarization.

Since the inclusion of images and tables to the slides offer an enhanced understandability to the viewers, automated slide generators should be capable of adding graphical elements also to the output slides. The efficiency of a slide generator can be measured as the degree to which all the important information has been mapped from the article to the output slides. The slides that are generated must have the following traits:

(1) All the important concepts must be arranged in a structured and methodical way and must be unbiased.

(2) The interaction from the user part for slide generation must be minimal
(3) The slides must have good content quality, diversity and sufficient length when compared to its corresponding research article.
(4) The slides must be editable at will so that the author can tweak the contents to suit the user's needs.

Even though a number of techniques have been proposed for automatically generating presentations from research articles, their efficiencies and throughputs are to be boosted to a much higher level. The rest of this paper is organized as follows. Section 2 describes some of the methodologies for automatically generating presentation slides. A brief comparison among the various techniques is presented in section 3 and Section 4 concludes review.

2. REVIEW OF AUTOMATIC SLIDE GENERATION TECHNIQUES

Most slide generators require the identification of important topics and sentences from the article. A wide variety of techniques have been used to identify the relevant parts that form the substance of an article. A review of the techniques for slide generation is presented to throw light on how the methods identify and organize the important parts of the technical document to form a presentation slide.

2.1 Generation of Slides Based on Inference of Underlying Semantic Structure of Articles

The inference of semantic structure of the article helps in understanding the relations between sentences, identification of important topics and co-references very easily.

Utiyama et al., [8] developed a system that could automatically generate presentation slides from an article and dynamically customize the presentation based on the queries issued by the audience. The articles were initially annotated by the GDA Tagset. Using the annotation, the system extracted the underlying semantic structure and relationships in the article. The system used the knowledge of semantic dependencies and co-references to extract out the topic parts from the article. From the list of topics, the system filtered out the most relevant topics, which later becomes the initial presentation. The system prepared a slide for each topic using the summarizing sentences extracted from the article. Based on the queries, the slides were updated on the fly and if required, the system referred the article again in case the document contained additional information relevant to the query at hand. Even though the system was language/ domain/ style independent, it could not add graphical elements to the output slides. Moreover, effective topic identification and audience interaction schemes were not efficiently made use of in the system.

2.2 Generation of Slides from the LaTeX Manuscript of an Article

LaTeX manuscript of a research article offers high structural information about the document. It offers an easy extraction facility of text and graphical contents from the document. The LaTeX source of an article is considered as the starting point for slide generation.

Y. Yasumura et al., [9] implemented a solution to the problem of automatically generating presentation slides from technical articles. The system required a user to input the LaTeX manuscript of the technical document. The method involved the calculation of weights of all the terms in the document using TF-IDF scoring technique. The term weights served as a strategy to determine a relevance score for all the objects in the document. The size of the summary to be extracted for each section could be determined by using the term weights. The solution employed the use of slide layouts so as to generate presentation slides. The output slides could be customized by the user. The implementation required the user to specify keywords to the system.

Sravanthi et al., [10] contributed a new solution to self-generate presentation slides of a technical document. The process began with the inference of the detailed logical structure of the article from the LaTeX manuscript of the document. Each section was categorized to fall under Introduction, Related Work, Model, Experiments or Conclusion. The system extracted important key phrases and the QueSTS Summarizer [11] was used to summarize the Model, Experiments and the Conclusion sections. The QueSTS summarizer considered the section to be summarized as an integrated graph consisting of vertices representing sentences. Node weights and edge weights were calculated and were used for picking the summary sentences corresponding to each section. The system extracted the graphical elements from the LaTeX manuscript and appended them to the output slides as required. It is evident that the efficiency of the system could attain a much higher level by incorporating NLP based techniques to increase the quality of the generated presentation slides.

2.3 Generation of Slides Using Natural Language Processing

Natural Language Processing, one of the pioneering area of computation, focuses on the analysis of the discourse structure, relationships between text units, grammar analysis etc. NLP based summarizers[12] are capable of studying the underlying discourse structure of the documents and to use this knowledge to summarize the contents efficiently. Slide generators can create the slides by using the extracted contents. Discussed below are techniques that generate slides based on NLP strategies.

Shibata et al., [13] described a method to generate presentation slides from text by analysing the discourse structure of the article. The system considered a clause and a sentence as a discourse unit and coherence relations such as contrast, list, additive, elaboration etc were extracted and analysed. The system extracted topic and non topic parts from the article based on the underlying discourse structure. The extracted topic and non-topic parts were placed on the slides by providing proper indents based on the analysis of their syntactic structure. The system built the slides by connecting relevant sentences to the most similar preceding sentences. The system also pruned the non-topic parts based on some heuristic measures to provide an enhanced readability.

K. Gokul Prasad et al., [14] implemented a system that was focused on the educational domain to create presentation slide for seminars and lectures. The system worked on the basis of 2 modules - Information Extractor & Slide Generator respectively and involved the initial extraction of text contents from the input. The system employed the use of core NLP operations like text segmentation and chunking to detect segments as well as noun phrases. The implementation allotted segments and their component sentences with weightage values. Slides were built on the basis of phrases having high relevance factors. An ontology tree was built for each noun phrase detected using a chunker. The system used the ontology

tree to infer the semantic relations between sentences. Based on the generated ontology and weightage value, the system detected the important key phrases, which were used for bullet point identification and finally, the construction of presentations. The system's accuracy could be boosted by using a domain-specific ontology.

2.4 Generation of Slides Using Text Summarization

Text summarization can be defined as the extraction of the most relevant representative parts of a text block. Summarizers uses strategies like ranking, random selection, location based selection, HMM etc to select sentences to be included in the summary.

Tulasi Prasad Sariki et al., [15] presented a novel scheme to summarize a document and hence use the summarized version to build presentation slides. The system initially fetched and created the document to be summarized and applied a number of basic text preprocessing techniques such as sentence division, case folding, stop word removal, stemming and lemmatization. The system then used a scoring scheme which was a combination of some of the baseline scoring methodologies such as cue phrase, key, title and location based scoring methods to allot a relevance score for each sentence. Using a sentence ranking scheme, the summary was created and from this summary, presentation slides were generated. The system was capable of making summaries specific to query keywords and hence construct a query specific presentation.

2.5 Generation of Slides Using Web Mining

The internet can be seen as a huge repository containing information which can be mined constructively to suit the requirements. Web mining usually involves the fetching of web pages containing the required information, extraction of data from them and finally, the application of mining techniques. A few techniques that project web mining to retrieve topic specific information and their organization as presentation slides are discussed.

ShaikhMostha Al Masum et al., [16] elaborated a new agent based scheme to build presentation slides by mining query based information from the internet. The data was gathered from Wikipedia or by using popular search engines like Google, Yahoo and Alta Vista based on the availability of required information. The system added images also to the output slides for providing better clarity & understanding on the topic. The method built the presentation data using a combination of techniques like web data fetching, web page parsing and summary extraction, each of which were performed by agents. The selection of a presentation template was done based on the choice of the information repository. The proposed scheme used the statistical method to find out a relevance score for each sentence during summarization. The method also used dedicated algorithms for webpage parsing and presentation generation. The system created MPML scripts and finally generated slides in HTML and Javascript and the topics were explained to different headings by agent characters. The algorithms for mining needed to be improved so as to increase the overall performance of the system.

Mistsuru Ishizuka et al., [17] worked on a new scheme that generated a concise report and a presentation by mining the web resources based on the query issued by a user. Each step in the creation of slides was completed by software agents. The system used 6 different software agents. Even though the above method is a similar technique, it did not generate a report corresponding to the issued query. As an initial step, the input queries were preprocessed by the system. Ambiguities if present were removed by adding to the topic, its top disambiguated senses. Based on the query topic, the system employed system agents to use techniques of web crawling and data extraction to download the web pages and hence extract the headings as well as the text contents from the downloaded papers. From the extracted data, concise summaries were generated using a summarization unit. For summarization, the closeness between texts was analysed by the calculation of a vector distance. For each topic, a report was generated and from each report, the proposed system generated the presentation scenes. The scenes are then converted to MPML scripts and finally to a Javascript code based presentation. The downside of the system was that it could not handle multiple user interactions and was sensitive to higher loads.

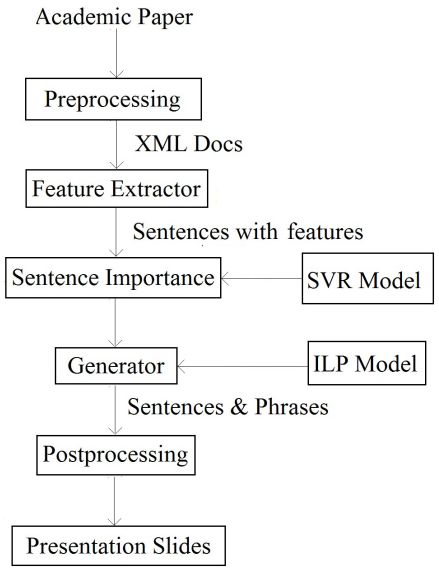

Fig. 1. System flow of machine learning based slide generation method

2.6 Generation of Slides Based on Machine Learning

Nowadays, machine learning has been widely used in many areas of computing nowadays. Almost all machine learning systems involve a training phase used to learn a model, which will be used for tasks like classification and prediction purposes for a set of test data.

Yue Hu et al., [18] established a new scheme to automatically generate presentation slides from research articles. A corpus based machine learning approach was followed where in the system was subject to an extensive initial training by making use of a very large corpus of paper-slide pairs. The system was trained in order to learn a Support Vector Regression Model to be used for score prediction. The system worked in 2 phases to create slides from a research article. In the first phase, the system worked on extraction of individual sentences from the article and representing each sentence using a list of features. The overall work flow is shown in Figure-1. The system then used the trained SVR model to predict an importance score for each sentence. A higher score signifies a higher relevance factor for the sentence. LIBSVM [19] with the RBF kernel was used to implement the SVR model. The Important key phrases were identified from the article using entity extraction. For identifying highly important key phrases, the concept of global and local phrases were used. A phrase was considered to be important if it was found to have a high frequency of occurrence. In the second phase, the system used an elaborate Integer Linear Programming model with a robust objective function and well defined

constraints to select key phrases and sentences such that the value of the objective function was maximized. The output slides served as a draft presentation and could be edited by the author. For each section, the important keywords and their corresponding sentences were arranged as the slide contents. Graphical elements were not added and the system considered only a single style of presentation format.

3. COMPARISON OF TECHNIQUES

The differences in the working strategies of all the discussed methods need to be understood effectively so as to gain an insight on to how slides can be automatically generated from research articles. For that purpose, a comparative explanation among the studied techniques is given in Table-1. The comparison discussed is effective and provides good information on the differences between the techniques.

Table 1. Comparison of automatic slide generation techniques

Basis of Procedure	Type
Inference of semantic dependencies in the article, identification of topics, dynamic customization of presentation.	Annotation based
Calculation of weights for the terms in the document, summarization of each section and generation of slides.	Statistical Method
Extraction of contents from LaTeX Manuscript, categorization of sections, usage of TF-IDF method for summarization.	Statistical Method
Analysis of discourse structure, detection of topic/non-topic parts, slide generation based on intending, pruning of non-topic parts.	NLP, Discourse Analysis
Text segmentation, chunking and ontology creation to identify important phrases for building the presentation.	NLP, Ontology
Preprocessing of text contents, usage of a combination of scoring schemes and finally, an adaptation of a ranking method to select important sentences to be added to the slides.	Statistical Method.
Gathering topic specific data from the internet, summarize them and generate a query specific presentation.	Statistical Method, Agent based
Gathering topic specific data from the internet, summarize them and generate a query specific report & presentation.	Agent based
Prediction of importance scores for each sentence using SVR, selection of the most important contents using ILP.	Machine Learning

As evident from the above comparison, slide generators can work on the basis of a variety of techniques. The comparison includes almost all important techniques of slide generation.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Experimental Results of Semantic Structure Inference Based Method

Eventhough the system could dynamically adapt the presentations based on queries, the experimental results proved that:

—Heuristic measures adopted by the system could fail in some scenarios.

4.2 Experimental Results of Methods using LaTeX Manuscript of an article

The system was given 8 articles and the corresponding authors were selected for rating the system based on various quality parameters. The results put forward the following points:

—Coherence of the slides was good.

—Coverage of the slide was great.

—Better quality would be achieved if sentences were compressed.

4.3 Experimental Results of NLP Based Methods

The evaluation results of [13] revealed the following areas:

—Errors could happen due to word-chain mis-identification.

—Recognition errors of contrast relations between clauses and sentences were encountered.

—The inference from the results was to transform original texts to multimodal presentations to enhance the presentation quality.

The experimental analysis of [14] proved the following results. Evaluation parameters of Precision, Recall and F1-Score were used for this purpose.

—The system had higher efficiencies on working with non-technical documents.

—It was inferred from the results that a domain specific ontology could be made use of to elevate the throughput of the system for technical documents.

4.4 Experimental Results of Text Summarization Based Method

The system was compared with baseline summarization schemes and the efficiency of the scheme can be visualized in Figure-2 as follows:

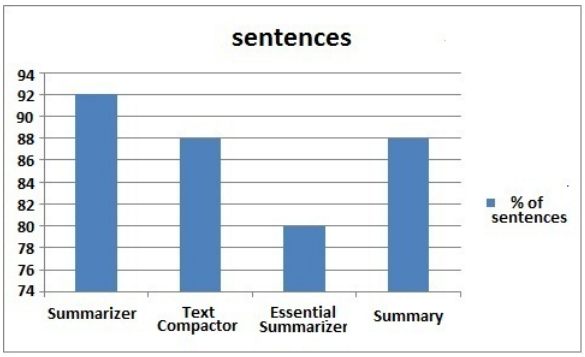

Fig. 2. Efficiency comparison of Summarization based slide generation technique

4.5 Experimental Analysis of Web Mining Based Method

The system was analysed by 25 students wherein the quality of presented data and overall quality were evaluated. The results were mostly positive. A subset of the experimental results of the system is shown as under in Table-2.

Table 2. Experimental evaluation of Web Mining based slide generation approach

Question Asked	Did it Work	Quality of Data Presented	Overall Quality
What is Big Bang?	Yes	Very Good	Acceptable
Tell me about Formula 1	Yes	Very Good	Acceptable
Tell me about F22	No	Not Good	Not Good
What is J2EE	Yes	Good	Good
What do you know about AI	Yes	Very Good	Fine

4.6 Experimental Analysis of Machine Learning Based Method

The system was evaluated based on ROUGE [22] scores obtained for summarization methods and Sentence Importance Computation Methods. The overall experimental results proved that the system offered a much more promising solution to the problem of automatic slide generation. The experimental results are visualized in Table-3.

Table 3. Experimental evaluation of Machine learning based slide generation approach

Method	Rouge-1	Rouge-1	Rouge-SU4
TF-IDF [9]	0.38859	0.11624	0.16424
Random Walk [20]	0.39421	0.11555	0.16463
Mead [21]	0.38778	0.11803	0.16239
C-Lexrank [2]	0.38722	0.11223	0.15858
PPSGen [18]	0.41342	0.13067	0.17502

5. CONCLUSION

Automatic generation of presentation slides from an article involves the identification and organization of all the relevant information in an ordered and methodical syntax. The investigation of some of the recent trends in the field of automatic slide generation is done that comes from greenhorn procedures to the methods that use the current capabilities of computation. A brief comparison of the discussed methodologies is also made and almost all the key techniques for automatic generation of slides have been surveyed. It is evident from the study that there is enough room for improvement in each of the discussed techniques. It would be more efficient if combinations of strategies are used to build the output slides. The goal of the slide generation process should be to learn the complexities behind how researchers build slides from their articles, and use this knowledge to generate slides that are much user friendly and highly customizable.

6. REFERENCES

[1] 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.

[2] 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.

[3] N. Agarwal, K. Gvr, R. S. Reddy, and C. P. Rose, "Towards multidocument summarization of scientific articles: Making interesting comparisons with SciSumm", in Proc. Workshop Autom. Summarization Different Genres, Media, Lang., 2011, pp. 8-15.

[4] O. Yeloglu, M. Evangelos, and Z.-H. Nur, "Multi-document summarization of scientific corpora", in Proc. ACM Symp. Appl. Comput., 2011, pp. 252-258.

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

[6] M. J. Conroy and D. P. O'leary, "Text summarization via hidden Markov models", in Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2001, pp. 406-407.

[7] D. Shen, J. T. Sun, H. Li, Q. Yang, and Z. Chen, "Document summarization using conditional random fields", in Proc. 20th Int. Joint Conf. Artif. Intell., 2007, vol. 7, pp. 2862-2867.

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

[9] 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.

[10] 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.

[11] 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.

[12] D. Marcu, "From discourse structures to text summaries", in Proc. ACL Workshop Intell. Scalable Text Summarization., 1997, vol. 97, pp. 82-88.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

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

[19] 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

[20] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford Digital Libraries, Stanford, CA, USA, Tech. Report: SIDL-WP-1999-0120, 1999.

[21] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Celebi, S.Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang,"MEAD - A platform for multidocument multilingual text summarization," in Proc. 4th Int. Conf. Lang. Resources Eval., 2004, pp. 14.

[22] C. Y. Lin "ROUGE: A package for automatic evaluation of summaries" in Proc. Workshop Text Summarization Branches Out, Post- Conf. Workshop ACL, 2004, pp. 2526.