Research Paper
Volume : 5 | Issue : 8 | Special Issue August-2016 • ISSN No 2277 - 8179

Enhanced Powerpoint Slide Generation System:Eppsgen

Engineering

KEYWORDS: Abstracting methods, text mining

S.Kavitha	Asso.Professor, Computer Science and Engineering Vedavyasa Institute of Technology Karad,Malappuram
Ambili C.V	PG Scholar, Computer Science and Engineering Vedavyasa Institute of Technology

ABSTRACT

In this system we propose a method to automatically generate slides from academic papers containing images . The generated slides are effective for researchers and students. The system works on multiple keywords. Here also we apply svr and ilp method on each paper to create a better output .The eppsgen is having evident advantages over existing methods.

I. INTRODUCTION

Presentation slides are very useful for students in academic process.This method is useful to present information in a pictorial way. Many softwares such as the Microsoft powerpoint is available for this.This is useful for formatting of slides not content.

In this system we propose a method to automatically generate slides which can be provided as draft slides for further presentation. This work is useful to elaborate work by introducing sections like abstract, introduction, conclusion, bibliography etc. Each slide contains a section and related sentences.

Also each slide contains images. It is the main advantage of our system. Existing methods are NLP techniques ,machine learning methods etc. These are less effective compared to our method. The slides are required to be much more structured. Here we use keywords also.

Here slides are created using svr and ilp methods. The images are included using apache poi library. Slides contain most sections that user require .Svr method is used to select the main sentences.Then the keywords are also selected. It provides a sentence ranking. Svr method is based on a set of features.

Ilp method is used to distinguish the keywords and sentences to be selected for slide generation. Selecting and aligning the keywords and sentences slides are created.Ilp model is based on objective function and constraints.

The experimental evaluation of Eppsgen is based on a set of paper-slide pair comparison. Rouge scores are taken. Rouge toolkit provides average recall, precision and F measure scores. By analyzing the rouge scores of ppsgen and eppsgen ,obtained that Eppsgen is having improved result.

Eppsgen is a web based system. User creates profile and logs into the system. User logs in if admin adds the user. Then user creates ppts using svr and ilp.Slides are more structured and having sections like Abstract, Introduction, Related Work etc. Ppts are having the name specified by user.System enables keyword search and ppt download.

II. RELATED WORK

Automatic slide generation is a difficult task to implement. Eventhough a lot of efforts have been made.One such effort is by analyzing the semantic structure by Utiyama and Hasida. Using GDA tagset underlying semantic structure and relationships were identified. Important sentences and topics were identified.The disadvantage is lack of effective topic identification.

Next is based on NLP techniques by Tomohide Shibata and Sadao Kurohashi..NLP mainly focuses on analysis of discourse structure,relationship between text unit,grammar analysis etc.Discourse structure is used to summarize contents efficiently.The system extracts topic and non topic parts from paper. The system works based on information extractor and slide generator.System implemented text segmentation and chunking to detect segments as well as noun phrases.

Next is mining of webpages by Shaikh Mostafa Al Masum and Mitsuru Ishizuka.Data gathered using Google and Yahoo.This system adds images also to improve the understandability.The system built presentation based on techniques like web page fetching ,web page parsing and summary extraction. The system works on relevance score of sentence. The system creates HTML scripts.

Algorithms for mining needs to be improved. The system is sensitive to higher loads .Presentations created are user interactive.

Nowadays machine learning techniques become popular. The system is based on extensive initial training. Then it uses svr on the trained data for score prediction. Svr mainly ranks the sentences. LIBSVM with RBF kernel is used to implement svr. Then ILP model with object function is used to select key phrases and sentences. But the main problem is that graphical elements are not added. This is the disadvantage of Yue Hu and Xiaojun Wan's work. So it becomes essential to propose a new system.

III. EPPSGEN

Proposed system is web based Eppsgen. It is a keyword search enabled system. Svr model is used for sentence ranking and ilp for slide generation. Images are selected using Apache poi library. The system works as draft slides for further presentations.It eases the task of researchers.

Svr model aims to learn a function f(u)=w* (u)+b where u is a feature vector, w is the sentence weight, (u) is the vector space. It works on regression score. Sentence's importance score is found as Score(s)=max(sim(s,s_i^*)).Where s_i^* is the set of sentence in the corresponding slides and s is a sentence.

The standard cosine measure is taken as the similarity function.Each sentence in paper is represented by a set of features. Such features are similarity with title, word overlap with title,sentence parse tree information, stop word percentage. Based on the importance scores of sentences in the training data we can learn an svr model and then apply it to predict the importance score of sentences of other papers. Ilp model is used for creating well structured slides by selecting and aligning the key phrases and sentences.To extract the key phrases chunking implemented by OpenNLP library is used. Two types of phrases considered are global and local phrases. An object function is the vital part of the ILP model described as,

$$\max \lambda 1 \sum_{i=1} li / L \max + \lambda 2 \sum_{i=1} Cbibi / |B^*| + \lambda 3 \sum_{i=1} wi / n^* yi$$

Where wi is the importance weight of sentence si, n is the number of sentence, li is the length of sentence ,xi is variable indicates whether si is included in slide, bm is the variable that indicates whether bigram bm is included in slides, Lmax is the Maximum length of slides , Cbi is the count of occurrence of bigram.

Object function is having 3 parts. The first part is for importance score maximization by summing up the score. The second part maximizes count of bigrams in paper which also appear in slide. This part maximizes slide diversity. The last part aims to maximize the weighted coverage of key phrases selected and determine which key phrases should be selected.

The object function is set to some constraints.
The constraints are,

(7) $\sum_{i=1}^{n} xi \leq Lm$

It guarantees that total word count of slides doesn't exceed Lmax.Constraint (8) is given as,

$$\sum_{lpj \in Lpi} lpj \geq xi \; for i = 1 \ldots n.$$

Lpi is a set of local phrases that sentence si contains.This constraint is used to elaborate the relationship between sentence and local phrases.

Constraint (9) $\sum_{si \in Sj}^{n} xi \geq$ for j=1.....|LP|.

Sj is the set of sentences that contains local phrase lpj. If a local phrase lpj is selected (lpj= 1), then atleast one sentence in Sj must be selected in slide generation process.

$$\sum_{lpj \in LPk} lpj \geq yk \; for k = 1; \ldots n.$$

It is the tenth constraint . LPk is the set of local phrases that sentence sk contains. If at least one local phrase in LPk is selected the value of yk must be set to 1.This is because the ILP solution should maximize the object function .If no local phrase in LPk is selected yk must be set to 0. It ensures that yk is set to 1 only when at least one local phrase in LPk is selected.

By applying these constraints to object function ,it is possible to maximize the output obtained. Another key feature of our system is that it is enabled with keyword search .Eppsgen provides graphical element with the ppt. The proposed architecture is given by,

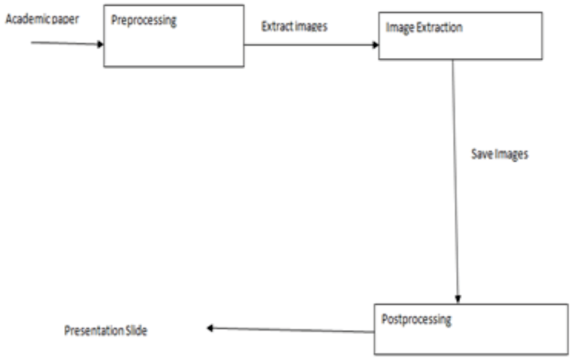


Fig 1. Architecture of Eppsgen

This architecture shows image extraction using Eppsgen system. First we input academic paper to the system.Then the next process is a preprocessing to remove unwanted images.Next step is to extract images using Apache poi library.Apache poi deals with powerpoint elements. Apache poi deals with extracting and editing powerpoint elements.In this system images are stored to a folder and finally displayed.

Apache POI is a popular API that allows programmers to create, modify, and display MS Office files using Java programs.

It is an open source library developed and distributed by Apache Software Foundation to design or modify Microsoft Office files using Java program. It contains classes and methods to decode the user input data or a file into MS Office documents.

Apache POI contains classes and methods to work on all OLE2 Compound documents of MS Office. The list of components of this API is given below.

- POIFS (Poor Obfuscation Implementation File System) : This component is the basic factor of all other POI elements. It is used to read different files explicitly.
- HSSF (Horrible Spreadsheet Format) : It is used to read and write xls format of MS-Excel files.
- XSSF (XML Spreadsheet Format) : It is used for xlsx file format of MS-Excel.
- HPSF (Horrible Property Set Format) : It is used to extract property sets of the MS-Office file.

IV. RESULT ANALYSIS

When analyzing the generated slides, the word count of slides is 15% of the word count of paper. Here we compare Eppsgen with Ppsen.Tfidf scoring method is taken as the basic method and compared with svr.Sentence having larger score is used to generate slides.

The methods are compared using the rouge toolkit. It is run in command prompt. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans.

The measures count the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans. Four different ROUGE measures: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S included in the ROUGE summarization evaluation package and their evaluations. Three of them have been used in the Document Understanding Conference (DUC) 2004, a large-scale summarization evaluation sponsored by NIST.

Traditionally evaluation of summarization involves human judgments of different quality metrics, for example, coherence, conciseness, grammaticality, readability, and content.Rouge toolkit provides average recall,precision and F- measure scores.In svr method we consider global and local phrases.

Mainly 4 slides are taken as reference and rouge scores are calculated of reference slides and also the generated slides.The generated slides are having better rouge scores in ppsgen.In Eppsgen also generated slides are having better rouge scores.But comparing to ppsgen Eppsgen is having better rouge scores.

Rouge scores show the content quality and structure.Rouge scores of Ppsgen is shown in table.The measurement is done using rouge 1 toolkit.Average recall,precision and the F-measure scores are taken.Rouge 1 is showing that svr is better method compared to tfidf.

First two systems are taken. System 1 is our system and system 2 is reference slides. Text documents are provided as input to the rouge toolkit.The rouge scores are obtained are small precision values.Rouge scores of Ppsgen are,

Volume : 5 | Issue : 8 | Special Issue August-2016 • ISSN No 2277 - 8179

Table 1.Ppsgen

ROUGE-Type	System Name	Avg_ Recall	Avg_ Precision	Avg_F-Score
ROUGE-1	SYSTEM2.TXT	0.20562	0.52857	0.29606
ROUGE-1	SYSTEM1.TXT	0.29241	0.53571	0.37832
ROUGE-1	SYSTEM4.TXT	0.29965	0.75833	0.42956
ROUGE-1	SYSTEM3.TXT	0.30361	0.7733	0.44061
ROUGE-1	SYSTEM2.TXT	0.26464	0.1887	0.18107
ROUGE-1	SYSTEM1.TXT	0.33877	0.21528	0.26326
ROUGE-1	SYSTEM4.TXT	0.13524	0.2045	0.22916
ROUGE-1	SYSTEM3.TXT	0.49169	0.25909	0.33936

Rouge scores of Eppsgen are,

Table 2.Eppsgen

ROUGE-Type	System Name	Avg_ Recall	Avg_ Precision	Avg_F-Score
ROUGE-1	SYSTEM2.TXT	0.20564	0.52856	0.29606
ROUGE-1	SYSTEM1.TXT	0.29243	0.53572	0.37832
ROUGE-1	SYSTEM4.TXT	0.29964	0.75833	0.42955
ROUGE-1	SYSTEM3.TXT	0.30362	0.7732	0.44061
ROUGE-1	SYSTEM2.TXT	0.26463	0.1886	0.18106
ROUGE-1	SYSTEM1.TXT	0.33876	0.21528	0.26326
ROUGE-1	SYSTEM4.TXT	0.13524	0.2044	0.22916
ROUGE-1	SYSTEM3.TXT	0.49168	0.25909	0.33936

When compare the values shown in the above tables ,it is obtained that Eppsgen is having better rouge scores.So we can say that Eppsgen works well compared to Ppsgen.From the above two tables it is obtained that Eppsgen creates more structured and better ppts.

Another feature of Eppsgen is that it contains images and also enables keyword based search. Next step is to compare the values obtained from ppsgen and Eppsgen. Bar charts are used for comparison. The values are plotted among X axis and Y axis.

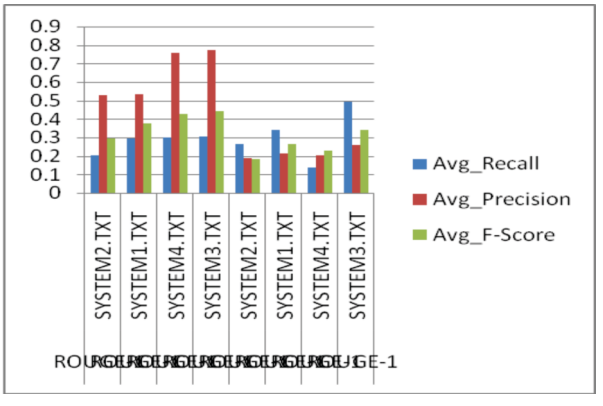


Fig 1:Rouge scores of ppsgen

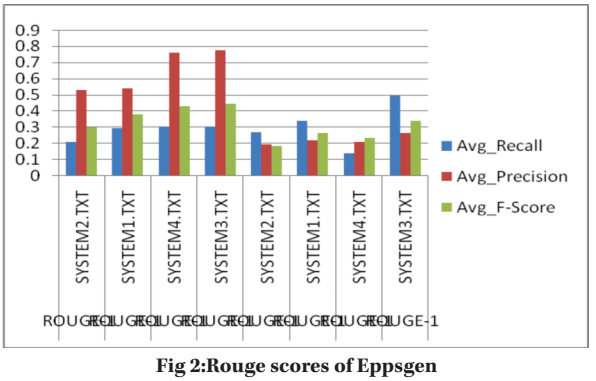


Fig 2:Rouge scores of Eppsgen

The above given charts compare the ppsgen and Eppsgen systems. System 1 is author written slides and system 2 is generated slides.Generated slides are better in content quality and also structure.

When we compare the average recall, precision and F measure scores of ppsgen and Eppsgen ,it is obtained that Eppsgen is having better recall score. Eppsgen is having better precision score .Also Eppsgen is having better F measure score. The scores of generated slides are mainly compared.So by all this data we can say that Eppsgen is the better system compared to ppsgen.

In pattern recognition and information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

Suppose a computer program for recognizing dogs in scenes from a video identifies 7 dogs in a scene containing 9 dogs and some cats. If 4 of the identifications are correct, but 3 are actually cats, the program's precision is 4/7 while its recall is 4/9. When a search engine returns 30 pages only 20 of which were relevant while failing to return 40 additional relevant pages, its precision is 20/30 = 2/3 while its recall is 20/60 = 1/3.

In statistics, if the null hypothesis is that all and only the relevant items are retrieved, absence of type I and type II errors corresponds respectively to maximum precision (no false positive) and maximum recall (no false negative). The above pattern recognition example contained 7 − 4 = 3 type I errors and 9 − 4 = 5 type II errors. Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity.

In the statistical analysis of binary classification, the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results, and r is the number of correct positive results divided by the number of positive results that should have been returned.

The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0.The traditional F-measure or balanced F-score (F1 score) is the harmonic mean of precision and recall:

F1=2.(precision.recall)/(precision+recall) The F-score is often used in the field of information retrieval for measuring search, document classification, and query classification performance.

Earlier works focused primarily on the F_1 score, but with the proliferation of large scale search engines, performance goals changed to place more emphasis on either precision or recall and so is seen in wide application.

The F-score is also used in machine learning. Note, however, that the F-measures do not take the true negatives into account, and that measures such as the Phi coefficient ,Matthews correlation coefficient, Informedness or Cohen's kappa may be preferable to assess the performance of a binary classifier.

The F-score has been widely used in the natural language processing literature, such as the evaluation of named entity recognition and word segmentation.Here we analyse the above mentioned scores.

V. CONCLUSION AND FUTURE WORK

This system proposes a system called Eppsgen for academic papers. Here we train a scoring model based on svr and use ILP method to generate slides by sentence and keyword selection. The analysis shows that our system (Eppsgen) works well compared to Ppsgen.

Rouge scores are compared. Eppsgen is having structure and content quality compared to ppsgen.Eppsgen extracts images also.But images are taken as a whole.

In future work,we mainly consider the alignment of graphical elements,thus making slides more understandable.We mainly identify and extract graphical elements in Eppsgen.For an improved system,the relationship between text elements and graphical elements should be considered.

In that case ,we have to identify which graphical elements should be selected to make slides and which sentences are relevant to a graphical element. This enables us to attach tables and graph to text elements.

In future we can improve system by adding variety of templates also.It is also possible to improve system by adding templates from the world wide web.

Now we focus on web based keyword search enabled Eppsgen.This system can also be improved to include more academic papers and prepare ppts based on all academic papers.These all issues will be solved in future.

The above mentioned features are implemented using the machine learning techniques or rule based methods.Thereby we can attach figures and tables to the text elements.

ACKNOWLEDGMENT

We are indebted to HOD Mrs.Kavitha Murugesan of our department, our tutor Ms.Ginnu George and the workshop reviewers for their comments on our work

REFERENCES

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

[2.] Automatic slide generation based on discourse structure analysis, in Proc. Int. Joint Conf. NaturalLang. Process., 2005, pp. 754–766, T. Shibata and S. Kurohashi.

[3.] Investigating automatic alignment methods for slide generation from academic papers, in Proc. 13th Conf.Comput. Natural Lang. Learn., Jun. 2009, pp. 111–119, B. Beamer and R. Girju.

[4.] 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, S. M. A. Masum, M. Ishizuka, and M. T. Islam.

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

[6.] A framework for monitoring agent-based normative systems, in Proc. 8th Int. Conf. Auton. Agents Multiagent Syst.,2009, pp. 153–160, S. Modgil, N. Faci, F. Meneguzzi, N. Oren, S. Miles, and M.Luck.

[7.] 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 , S. M. A. Masum and M. Ishizuka.

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

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

[10.] Developing learning strategies for topic-based summarization, Proc. 16th ACM Conf. Conf. Inf.Knowl. Manage., Nov. 2007, pp. 79–86, Y. Ouyang, S. Li, and W. Li.