International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 07
June 2017

Academic Papers Slides Generation Using Nlp & Ilp

SUHANA.K[1] & EBIN.PM[2]

[1]MASTER OF TECHNOLOGY IN COMPUTER SCIENCE COCHIN COLLEGE OF ENGINEERING AND TECHNOLOGY, VALANCHERY, KERALA, INDIA

[2]ASSISTANT PROFESSOR IN COMPUTER SCIENCE COCHIN COLLEGE OF ENGINEERING AND TECHNOLOGY, VALANCHERY, KERALA, INDIA

ABSTRACT

Presentation slides have been a popular and efficacious designates to present and transfer information, especially in academic conferences. So we investigate a very arduous task of automatically engendering presentation slides for academic papers with including both text elements and graphical elements. The engendered presentation slides can be utilized as drafts to avail the presenters prepare their formal slides in a more expeditious way. The proposed system is to address this task. It first employs the regression method to learn the paramountcy scores of the sentences in an academic paper, and then exploits the integer linear programming (ILP) method to engender well-structured slides by culling and aligning key phrases and sentences. Evaluation results on a test set of 200 pairs of papers and slides accumulated on the web demonstrate that our proposed system can engender slides with better quality. A utilizer study is withal illustrated to show that proposed system has a few evident advantages over baseline methods. The automatic presentation slides generation system's Main quandary was images. Earlier system are failed to include images with the slides. The proposed system overcome this quandary.

Key words: - Abstracting methods, NLP, Text Mining, ILP, Slide Generation, Academic paper, Presentation slides, Integer Linear Programming, Support Vector Regression, Web Mining

INTRODUCTION

Presentation slides have been a popular and effective means to present and transfer information, especially in academic conferences. The researchers always make use of slides to present their work in a pictorial way on the conferences. There are many software's such as Microsoft Power-Point and Open Office to help researchers prepare their slides. However, these tools only help them in the formatting of the slides, but not in the content. It still takes presenters much time to write the slides from scratch. In this work, we propose a method of automatically generating presentation slides for academic papers. We aim to automatically generate well-structured slides and provide such draft slides as a basis to reduce the presenters' time and effort when preparing their final presentation slides. Academic papers always have a similar structure. They generally contain several sections like abstract, introduction, related work, proposed method, experiments and conclusions. Although presentation slides can be written in various ways by different presenters, a presenter, especially a beginner, always aligns slides sequentially with the paper sections when preparing the slides. Each section is aligned to one or more slides and one slide usually has a title and several sentences. These sentences may be included in some bullet points. Our method attempts to generate draft slides of the typical type mentioned above and helps people to prepare their final slides. We propose a novel system called PPSGen to generate well-structured presentation slides for academic papers with including both text elements and graphical elements. In our system, the importance of each sentence in a paper is learned by using the support vector regression (SVR) model with a number of useful features, and then the presentation slides for the paper are generated by using the integer linear programming (ILP) model with elaborately designed objective function and constraints to select and align key phrases and sentences

2. RELEGATED WORK

PPSGen investigated the method of automatically engendering presentation slides for academic papers. This method first

employs the regression method to learn the paramountcy scores of the sentences in an academic paper, and then exploits the integer linear programming (ILP) method to engender well-structured slides by culling and aligning key phrases and sentences. Evaluation results on a test set of 200 pairs of papers and slides amassed on the web demonstrated that PPSGen system can engender slides with better quality. Sentence assessment is done predicated on Support Vector Regression (SVR) . Then by utilizing extracted sentence from the paper slide is engendered. Slide generation is predicated on Integer Linear Programming (ILP). Then after post processing engendered output is the presentation slide. The disadvantage of the system is that it considers only text elements into the slide. Graph elements such as tables and figures are not considered. Multi document summarization involves multiple aspects of content cull and surface entelechy. Multi summarization method is utilized to engender summaries from multiple papers predicated only on integer linear programming (ILP). The summaries must be informative, succinct, grammatical, and comply with stylistic inditing conventions . It learns individual aspects but optimized jointly utilizing an integer linear programme. The ILP framework sanctions amalgamating the decisions of the expert learners and to cull and re-indite source content through a cumulation of objective setting, soft and hard constraints. The disadvantage of this system is that it fixates on summarization of the content. It does not engender slides and it does not consider graph elements. Sentence paramountcy assessment is not up to quality standards. SlidesGen investigated automatic generation of presentation slides from technical papers in LATEX. A query categorical extractive summarizer QueSTS is utilized to extract sentences from the text in the paper to engender slides. QueSTS transfers the input text to an integrated graph (IG) where a sentence represents a node and edges subsist between the nodes that the sentences corresponding to them are homogeneous. The weights of the edges are calculated as cosine homogeneous attribute between the sentences . SlidesGen framework includes steps such as Preprocessing, Configuration file generation, Extracting key phrases and QueSTS Summarizer . Conclusively slides are engendered for each section in the paper and graphics are rendered in presentation.

The disadvantage of the system is that it takes input only as latex format, wherein all documents cannot be considered as input. Most of the documents are in PDF format so many documents cannot be processed.

3. AUTOMATIC SLIDE GENERATION BASED ON DISCOURSE STRUCTURE ANALYSIS

3.1 The GDA Tagset:

GDA is a project to make WWW texts machine understandable on the substratum of a. linguistic tag set, and to develop applications such as content-predicated presentation, retrieval, question-answering, summarization, and translation with much higher quality than afore. GDA thus proposes an integrated ecumenical platform for electronic content authoring, presentation, and reuse. The GDA tagset 1 is predicated on XML, and designed as compatible as possible with HTML, and TEI 2, etc.

3.2 Parse-Tree Bracketing

GDA tagging is to encode semantic structure; syntactic annotation is exploited only as far as it contributes to semantic encoding. Withal, syntactic tags are designed to simplify syntactic annotation by minimizing the number of tags and accordingly the depth of embedding among them.

3.3 Topic Detection

Topics are often represented by paramount words and/or phrases in the documents. A traditional method for topic identification is to utilize word/phrase-occurrence frequencies to extract such

expressions. Such a method is not adequate for extracting topics, however, due to the following reasons:

1. A word is often too short to plenarily represent a topic.
2. A topic is often represented by a variety of expressions.

3.4 Slide Generation

A slide show is engendered by composing a slide for each topic culled. In the current implementation of the slide presentation system, each slide is fundamentally an itemized summary of the segment concerning the topic. The initial slide may be a table of contents of the whole slide show, which is compiled by listing the topics. Each slide in the main body of the presentation is composed by following the steps below. Here a topical element is a GDA element linked with the topic via the eq, ctp, sub, or sup cognation. A topical

element which is the subject of a whole sentence is called a topical subject.

1. Let the topic be the heading of the slide.
2. Extract consequential sentences which contain topical subjects.
3. Abstract redundant sentences, such as one elaborated by another extracted sentence, where elaboration is encoded by the ela cognation.
4. Itemize the remaining sentences

4. INVESTIGATING AUTOMATIC ALIGNMENT METHODS FOR SLIDE GENERATION FROM ACADEMIC PAPERS

4.1 Alignment Methods

Discovering how humans generate slide presentations from papers starts with observing where slide regions originate from. We make the general assumption that a slide region either a) is a summarization (excerpt or abstract) from the associated paper, or b) comes from other sources including but not limited to the author's personal (world and/or specific) knowledge. A complete alignment module would thus need to be able to discern if the information in a region comes from the target paper or if it does not. , the task of the aligner is then to choose the region in the paper that is summarized or from which the excerpt is taken. Original hypothesis was that the vast majority of the data in a given slide presentation would come from the target paper and concluded that a reasonable first attempt at building an aligner could be made under this assumption.

4.2 Scoring Methods

There are two scoring methods, which will refer to as scoring method 1 and scoring method 2. Scoring method 1 is implemented by aligners A and B and is equivalent to the average TFIDF score of the search terms relative to the target region. I.e. to calculate the score for a slide region relative to a target paper region with method 1, the TF-IDF scores of all the search terms are added and the sum is divided by the number of terms, and the 113 target region with the highest average score wins. Scoring method 2 is implemented by aligners C and D and is based on the quantity of matched terms, reverting to scoring method 1 only in the case of a tie. Thus, to calculate the score for a slide region relative to a target paper region with method 2, the number of search terms with non-zero TF-IDF scores for the paper region is counted and the region with the largest number of such search terms

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 07
June 2017

wins. In the case of a tie, the average score is calculated as it is in method 1 and the region with the highest average score wins the tie. With either scoring method, a zero score results in the system predicting that the slide region is not derived from any paper region.

4.3 Query Expansion

One common problem with rudimentary TF-IDF based information retrieval systems is that matching tokens must have a form identical to the search terms. Hence, synonyms and other semantically related words that probably should match do not. Query expansion is one way to consider terms which are semantically near, but orthographically different from the search terms. The general principle of query expansion is that, via an external knowledge base, semantic neighbors of search terms are added to the search query before the score is calculated.

5 .EXPEIMENTAL RESULTS

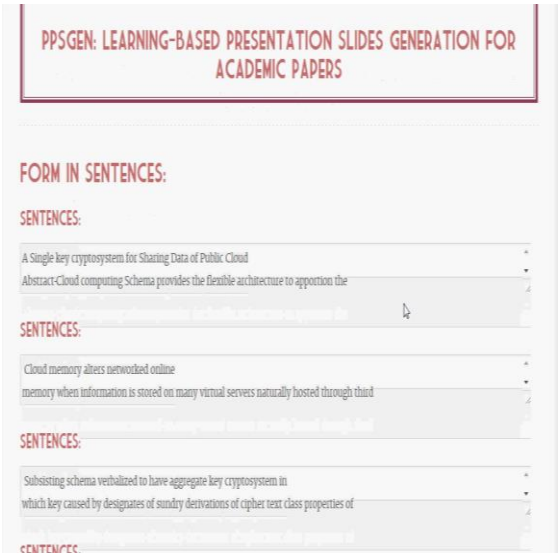

Fig 1: Sentence Formation

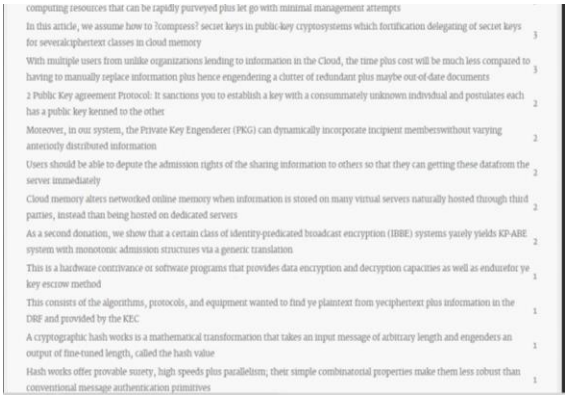

Fig 2: Sentence Count

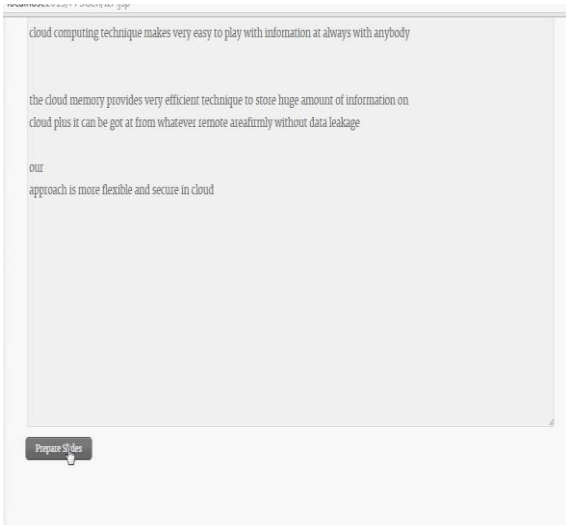

Fig 3: Pre processing Slide

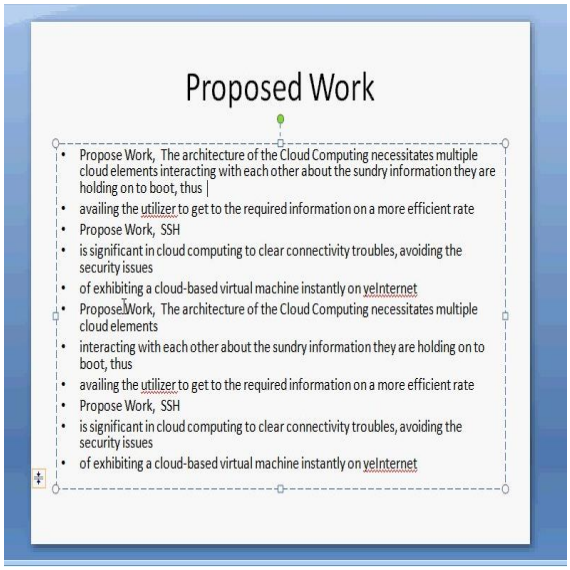

Fig 4: Sample Slide with Text Data

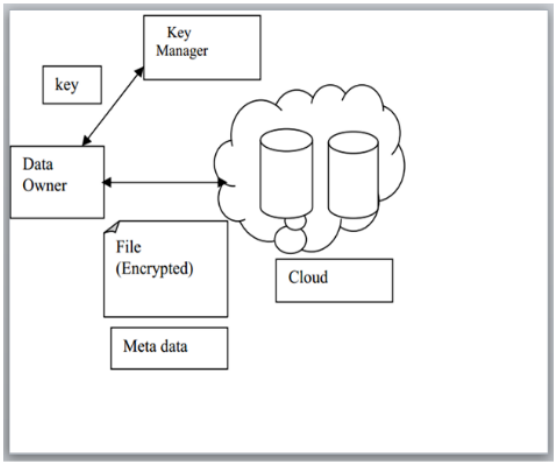

Fig 5: Sample Slide with Image

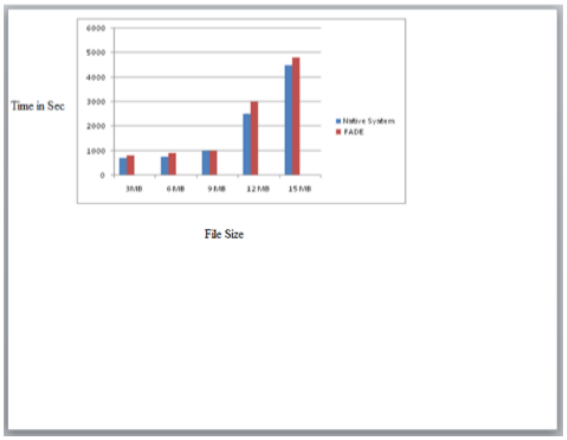

Fig 5: Sample Slide with Graph

6. CONCLUSION

Presentation slides have been a popular and efficacious denotes to present and transfer information. The main issue for engendering slides by utilizing software's such as Microsoft Power- Point and Open Office, which only avail them in the formatting of the slides, but not in the content. It still takes presenters much time to indite the slides from scratch. so we are utilizing a PPSGen system to engender presentation slides from academic papers. We train a sentence scoring model predicated on SVR and utilize the ILP method to align and extract key phrases and sentences for engendering the slides. Experimental results show that our method can engender much better slides than traditional methods. The consequential features of proposed system are the presentation slides with including utilizing

both text and graphical elements in the paper and make slides more comprehensible and vivid. Presence of graphical elements amends the quality of slides.

7. REFERENCE

1. M. Utiyama and K. Hasida(1999), Automatic slide presentation from semantically annotated with the GDA tag set, in Proc. ACL Workshop Conf. , pp. 25–30.

2. Y. Yasumura, M. Takeichi, and K. Nitta(2003) , A support system for making presentation slides, Trans. Japanese Soc. Artif. Intell.,vol. 18, 212–220.

3. T.Shibata and S. Kurohashi,(2005),Automatic slide generation based on discourse structure analysis, in Proc. Int. Joint Conf. Natural Lang. Process., 754–766.

4. B. Beamer and R. Girju(2009),Investigating automatic alignment methods for slide generation from academic papers, in Proc. 13th Conf. Comput. Natural Lang. Learn, 111–119.

5. S. M. A. Masum, M. Ishizuka, and M. T. Islam(2005), 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., 246–249.