CONVERTING PAPERS TO SLIDES USING PARSING

Dr.Kiran P[1], Sanchitha S D[2]

[1] Associate Professor, Department of CSE, RNS Institute of Technology, Bengaluru, Karnataka, India.
[2]PG Scholar, Department of CSE, RNS Institute of Technology, Bengaluru, Karnataka, India.

ABSTRACT

This paper investigates a very challenging task of automatically generating presentation slides for academic papers. The generated presentation slides can be used as drafts to help the presenters prepare their formal slides in a quicker way. A novel system called PPSGen is proposed to address this task. It first employs the regression method to learn the importance scores of the sentences in an academic paper, and then exploits the integer linear programming (ILP) method to generate well-structured slides by selecting and aligning key phrases and sentences.

Key words- *Abstracting methods, text mining*

1. 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 softwares such as Microsoft PowerPoint and OpenOffice 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, it proposes a method of automatically generating presentation slides for academic papers. This paper aims 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. This method attempts to generate draft slides of the typical type mentioned above and helps people to prepare their final slides.

Automatic slides generation for academic papers is a very challenging task. Current methods generally extract objects like sentences from the paper to construct the slides. In contrast to the short summary extracted by a summarization system, the slides are required to be much more structured and much longer. Slides can be divided into an ordered sequence of parts. Each part addresses a specific topic and these topics are also relevant to each other. Generally speaking, automatic slide generation is much more difficult than summarization. Slides usually not only have text elements but also graph elements such as figures and tables.

In this study, a novel system called PPSGen is proposed to generate well-structured presentation slides for academic papers. In this 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. RELATED WORK

Automatic slides generation for academic papers remains far under-investigated nowadays. Few studies directly research on the topic of automatic slides generation. M. Utiyama and K. Hasida et al., [1] discusses automatic generation of presentation slides from semantically annotated documents. The reported system is also domain independent and easy to adapt to different languages.

Y. Yasumura, M. Takeichi, and K. Nitta et al., [2] introduced a support system for making slides from technical papers. The inputs of the system are academic papers in LATEX format. T. Shibata and S. Kurohashi et al., [3] proposed a method to automatically generate slides from raw texts. Clauses and sentences are considered as discourse units and coherence relations between the units such as list, contrast, topic chaining and cause are identified. Some of the clauses are detected as topic parts and others are regarded as non-topic parts.

T. Hayama, H. Nanba, and S. Kunifuji et al., [4] studied the problem of aligning technical papers and presentation slides and used variation of the Hidden Markov Model (HMM) to align the text in the slides to the most likely section in the paper, which also used the additional information of titles and position gaps. M.Y. Kan [5] applied a modified maximum similarity method to do the monotonic alignments and trained a classifier to detect slides which should not be aligned.

B. Beamer and R. Girju et al., [6] compared and evaluated four different alignment methods that were combined by methods such as TF-IDF term weighting and query expansion. M. Sravanthi, C. R. Chowdary, and P. S. Kumar et al., [7] investigated automatic generation of presentation slides from technical papers in LATEX. A query specific extractive summarizer QueSTS is used to extract sentences from the text in the paper to generate slides. D. Galanis, G. Lampouras, and I. Androutsopoulos et al., [8] used a method based on both SVR and ILP to deal with multi-document summarization which is most relevant to the work.

3. PROBLEM DEFINITION AND CORPUS

3.1 Problem Definition

This work aims to automatically generate presentation slides for academic papers. This method needs to generate well-structured slides as the draft slides for a presenter to prepare the final slides. A beginner usually prepares slides which are sequentially aligned with the paper. One section in the paper is generally aligned to one or more slides. One slide usually includes several bullet points and sentences that explain the corresponding bullet points. It is reasonable to use that style of slides that beginners always use to make draft slides and we regard it well-structured because it uses pairs of bullet points and sentences to address important points and makes it easy for the reader to handle the points. In this work, it only considers the text elements in the paper. Other elements such as tables and figures are not included in the generated slides. It focus on the generation of the text elements.

3.2 Corpus and Preprocessing

To learn how humans generate slides from academic papers, a corpus is build that contains pairs of academic papers and their corresponding slides. Many researchers in the computer science field place their papers and the corresponding slides together in their homepages. The homepages' URLs are obtained by crawling Arnetminer. After downloading the homepages, we use several strict patterns to extract the links of the papers and the associated slides and download the files to build the dataset.

The papers are all in PDF format and the slides are in either PDF or PowerPoint format. For the papers, we extract their texts by using PDFlib and detect their physical structures of paragraphs, sections and sections by using ParsCit. A custom XML format is used to describe this structure. For the slides, it also extract their texts and physical structures like sentences, titles, bullet points, etc. It uses xpdf and the API provided by Microsoft Office to deal with the slides in PDF and PowerPoint formats, respectively. The slides are transformed to a predefined XML format as well.

4. PROPOSED METHOD

4.1 Overview

This paper proposes a system to automatically generate slides that have good structure and content quality from academic papers. The architecture of the system is shown in Fig. 1. It uses the SVR-based sentence scoring model to assign an importance score for each sentence in the given paper, where the SVR model is trained on a corpus collected on the web. Then, generate slides from the given paper by using ILP.

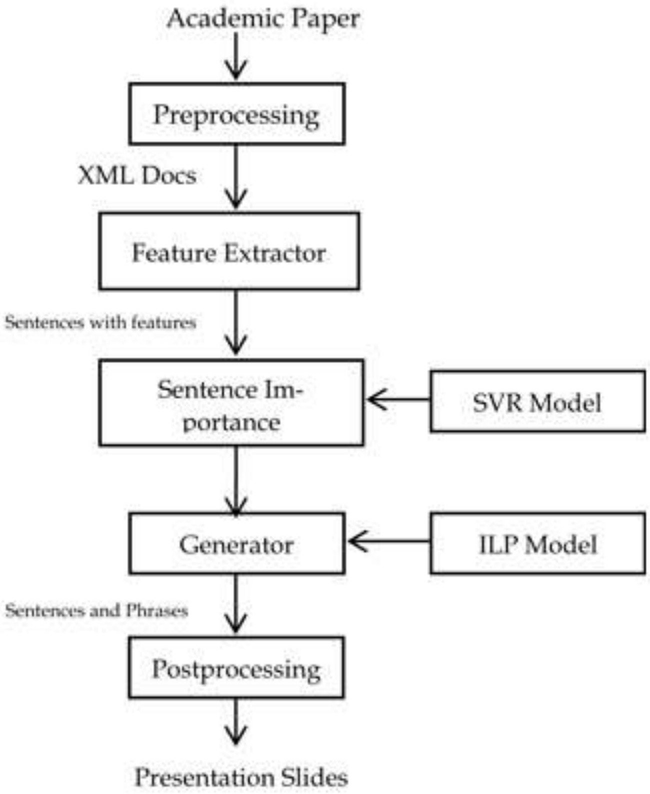

Fig-1: System architecture

4.2 Sentence Importance Assessment

In the proposed PPSGen system, sentence importance assessment is one of the two key steps, which aims to assign an importance score to each sentence in the given paper. The score of each sentence will be used in the slides generation process. In this study, a few useful features are introduced and propose to use the support vector regression model to achieve this goal.

4.2.1 Support Vector Regression Model

Initially predict the importance score of each sentence for sentence selection in slides generation. Based on the score the sentences are selected for slides. The reason why the SVR model is used instead of the classification model is that the regression score is finer to be used for sentence selection than the coarse binary category.

4.2.2 Training Data Construction and Model Learning

To construct training data based on the paper-slides pairs, apply a similarity scoring method to assign the importance scores to the sentences in a paper. The main hypothesis is that the sentences in the slides should represent the substance of the corresponding paper. The sentences in the paper which are more similar to the sentences in the slides should be considered more important and higher scores should be assigned to them using the scoring method. The sentence's importance score is set as the maximum similarity between the sentence and any sentence in the corresponding slides.

Intuitively, a sentence with a higher maximum similarity is closer to one sentence in the author-written slides. Since the author-written slides contain the sentences that human authors considered most important, a sentence with a higher score is most likely to be important, too. It adopts the maximum similarity instead of the overall similarity with all the sentences in the slides or the average similarity with each sentence in the slides. The motivation is that slides can be generally divided into several parts and each part may be relevant to one section in the paper. The sentences in a specific section should be more similar to the corresponding part in the slides and less similar to the other parts. Therefore, it is more reasonable to use the maximum similarity to assign the importance scores of the sentences.

Each sentence in a paper is represented by a set of features. The following features for each sentence are:

1. *Similarity with the titles*. It consider three types of titles: paper title, section titles and section titles. Only the titles of the section and section which contain the sentence are used. It uses the cosine similarity values between the sentence and different types of titles as different features. Stop words are removed and all the words are stemmed in the similarity calculation. Intuitively the sentences that have higher similarity with the titles should be more likely to be selected.
2. *Word overlap with the titles*. It is the number of words shared by the sentence and the set of words of all titles, including all three types of titles mentioned above.
3. *Sentence position*. Position of sentence and number of sentences are computed.
4. *Sentence's parse tree information*. The features are extracted from the sentence's parse tree. It includes the number of noun phrases and verb phrases, the number of sub-sentences and the depth of the parse tree. OpenNLP library is used for sentence parsing.
5. *Stop words percentage*. It is the percentage of the stop words in the total word set of the sentence s. Intuitively the sentences that have high percentage of the stop words are less likely to be important.
6. Other features including the length of sentence s, the number of words after removing stop words and the average length of sentences of the section, section or paragraph that contains the sentence.

All the features mentioned above are scaled into [–1, 1]. Based on the features and importance scores of the sentences in the training data, we can learn an SVR model, and then apply the model to predict an importance score for each sentence in any paper in the test set. The score indicates the possibility of a sentence to be selected for making slides.

4.3 Slides Generation

After getting the predicted importance score for each sentence in the given paper, it exploits the integer linear programming method to generate well-structured slides by selecting and aligning key phrases and sentences. Unlike those methods [1], [2],[9] that generate slides by simply selecting important sentences and placing sentences on the slides, it select both key phrases and sentences to construct well-structured slides. This method use key phrases as the bullet points and sentences relevant to the phrases are placed below the bullet points.

In order to extract the key phrases, chunking implemented by the OpenNLP library is applied to the sentences and noun phrases are extracted as the candidate key phrases. It define two kinds of phrases: global phrases and local phrases. Any unique phrase in an article is a global phrase, and a local phrase means a global phrase in a particular section. A global phrase that appears in different sections can correspond to a few local phrases. Since an important phrase is always used in many different sections, a global phrase that corresponds to more local phrases should be regarded to be more important and more likely to be selected. Thus, it uses the local phrases to generate the bullet points directly for different sections and use the global phrases to address the importance differences between different unique phrases. All the phrases are stemmed and stop words are removed. Moreover, the noun phrases that appear only once in the paper are discarded.

5. ACKNOWLEDGMENT

I would like to make use of this wonderful opportunity to thank my parents for their constant support and motivation. I would like to thank my internal guide Professor Dr. Kiran P, Department of Computer Science at RNS Institute of Technology for their guidance in successfully undertaking the project. I would also like to thank our beloved Dr. G T Raju, who is the professor, dean and HOD of Department of Computer Science for the encouragement and support. Finally I would also like to thank my teaching and non-teaching staff for providing us wonderful teaching and their blessing to move in a right path.

6. CONCLUSION

This paper proposes a novel system called PPSGen to generate presentation slides from academic papers. This system is trained on a sentence scoring model based on SVR and use the ILP method to align and extract key phrases and sentences for generating the slides. In this paper, it only considers one typical style of slides that beginners usually use. In the future, more complicated styles of slides such as styles that slides are not aligned sequentially with the paper and styles that slides have more hierarchies can be considered. Furthermore, this system generates slides based on only one given paper. Additional information such as other relevant papers and the citation information can be used to improve the generated slides.

7. REFERENCES

[1] M. Utiyama and K. Hasida, “Automatic slide presentation from semantically annotated documents,” in Proc. ACL Workshop 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] 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.

[4] T. Hayama, H. Nanba, and S. Kunifuji, “Alignment between a technical paper and presentation sheets using hidden Markov model,” in Proc. Int. Conf. Active Media Technol., 2005, pp. 102–106.

[5] M.Y. Kan, “SlideSeer: A digital library of aligned document and presentation pairs,” in Proc. 7th ACM/IEEE-CS Joint Conf. Digit. Libraries, Jun. 2006, pp. 81–90.

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

[7] 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.

[8] D. Galanis, G. Lampouras, and I. Androutsopoulos, “Extractive multi-document summarization with integer linear programming and support vector regression,” in Proc. COLING, 2012, pp. 911–926.