ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

A provision system for making presentation Slides from document

M. Vigneshwaran B.E.,
PG Student,
Department of Computer Science & Engineering,
Priyadarshini Engineering College,
Vaniyambadi,
Vignesh2cse@gmail.com

S.Suresh M.E.,
Assistant Professor,
Department of Computer Science & Engineering
Priyadarshini Engineering college,
Vaniyambadi,
sureshcseit@gmail.com

ABSTRACT

Presentations are one of the most common and effective way of communicating the overview of the work to audience. Given a technical paper, automatic generation of presentation slides reduces the efforts of the presenters and it also helps in creating a structured summary of the paper. We propose the framework of a novel system that does this task. Any paper that has an abstract and whose section can be categorized under instruction, related works, model, experiments and conclusions can be given as input. As documents in latex are rich in structural and semantics information we used them as input to our system. These documents are initially converted to xml format. This xml file is parsed and its information is extracted. All graphical elements in the paper are placed them at appropriate locations in the slides.

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 much software 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. 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. But our work focuses on the text elements only.

Experiments on a test set of 200 paper-slides pairs indicate our method can generate slides with better quality than the baseline methods. Using the ROUGE toolkit and the pyramid evaluation, the slides generated by our method can get better ROUGE scores and pyramid scores. More- over, based on a user study, our slides can get higher rating scores by human judges in both content and structure aspects. Therefore, our slides are considered a better basis for preparing the final slides.

RELATED WORK

DOCUMENTING SUMMARIZATION

The task of document summarization aims to generate a very short summary for a given document or document set. Various methods have been proposed for document summarization, including rule-based methods [3], graph-based methods [5], learning-based methods [2], ILP-based methods [11], etc.

Recently support vector regression and ILP have been used widely in the task of summarization. Ouyang et al. [14] and Galanis and Malakasiotis [2] used SVR to train and learn the sentence importance score. McDonald [11] proposed the first ILP method for summarization. It constructed summaries by maximizing the importance of the selected sentences and minimizing their pairwise similarity.

SLIDE GENERATION

Automatic slides generation for academic papers remains far under-investigated nowadays. Few studies directly research on the topic of automatic slides generation. Utiyama and Hasida [7] attempted to automatically generate slides from input documents annotated with the GDA tagset.1 GDA tagging can be used to encode semantic structure. The semantic relations include grammatical relations such as subject, the semantic relations such as agent, patient, and rhetorical relations such as cause and elaboration. They first detect topics in the input documents and then extract important sentences relevant to the topics to generate slides.

SCIENTIFIC ARTICLE SUMMARIZATION

The goal of scientific article summarization is to generate a short summary for a given scientific article or article set. Early works including [10] tried to use various features specific to scientific text (e.g., rhetorical clues features). Citation information has already shown its effectiveness for summarization of the scientific articles. Various works including [12] employed citation information for the scientific article summarization. Earlier work indicated that citation sentences may contain important concepts that can give useful descriptions of a paper. Agarwal et al. [8] introduced an unsupervised approach to the problem of multi-document scientific article summarization. The input is a list of papers cited together within the same source article. The key point of this approach is a topic based clustering of fragments extracted from each cocited article. Yeloglu et al. [9] compared four different approaches for multi-document scientific articles summarization: MEAD, MEAD with corpus specific vocabulary, LexRank and W3SS.

PROBLEM DEFINITION

In our work, we aim to automatically generate presentation slides for academic papers. We need to generate well-structured slides as the draft slides for a presenter to prepare the final slides. There are various kinds of slides which are made by Microsoft PowerPoint and OpenOffice.

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, we not only consider the text elements in the paper. Other elements such as tables and figures are also included in the generated slides.

PROPOSED FRAMEWORK

Overview

we propose a system to automatically generate slides that have good structure and content quality from academic papers.

The architecture of our system is shown in Fig. 2. We use the SVR-based sentence scoring model to assign an importance score for each sentence in the given paper.Then, we generate slides from the given paper by using ILP.

Sentence Importance

In our 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.we introduce a few useful features and propose to use the support vector regression model to achieve this goal.

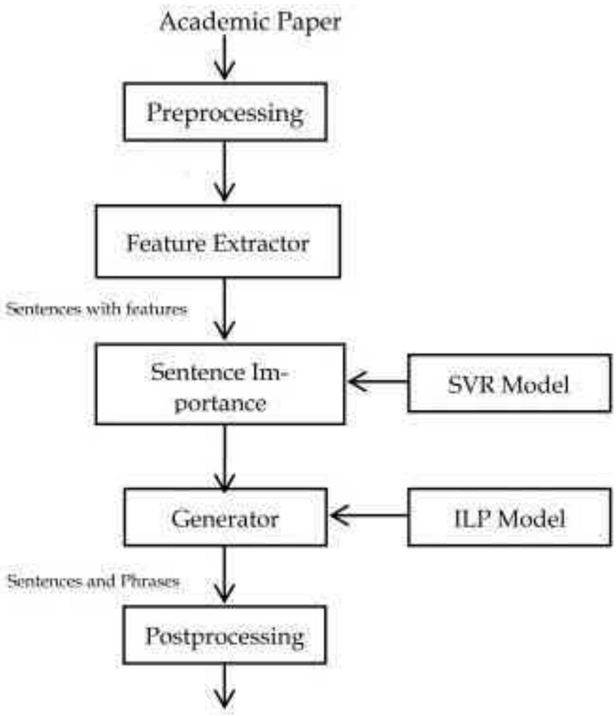

Fig 1.System Architecture

Support Vector Regression Model

Here we briefly introduce the SVR model [4]. We need to predict the importance score of each sentence for sentence selection in slides generation.

To construct training data based on the paper-slides pairs, we 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.

we define the sentence's importance score in the paper as follows:

$$Score(s) =max(sim(s,s_i*))$$

Where,

s is a sentence in the paper,

S* is the set of the sentences in the corresponding slides

s_i* is a sentence in S*.

Each sentence in a paper is represented by a set of features. we make use of the following features for each sentences:

1) ***Similarity with the titles***. We 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. We use 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'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.
4) ***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.
5) ***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.

Slide Generation

After getting the predicted importance score for each sentence in the given paper, we exploit the integer linear programming method to generate well-structured slides by selecting and aligning key phrases and sentences.

Unlike those methods [7], [13], that generate slides by simply selecting important sentences and placing sentences on the slides, we select both key phrases and sentences to construct well-structured slides. We 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.

We 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. we use 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.

We implement four baseline methods for comparison with our proposed method:

TF-IDF based method*:* This method is used by [13] based upon the TF-IDF scores, which extracts senten- ces for each section or section. The IDF scores are cal- culated on the corpus we collect. The sentences that have larger TF-IDF scores are selected to generate the slides.

MEAD based method: MEAD8 [4] is the most elabo- rate and publicly available platform for multi-lingual summarization. It implements multiple summariza- tion methods including position-based, centroid- based, length-based and query-based. A combina- tion of these methods is adopted to extract the slides. MEAD is applied to the whole paper instead of each section or section.

***Random Walk based method*:** In the Random Walk [6] method, sentences are regarded as nodes, the cosine similarities between sentences are assigned to be the weights of edges and the random walk method is employed to assign scores to the sentences. Here we apply the random walk method to each section, i.e., the sentences in one section and their relationship make up the graph that the random walk method applies to. The sentences that have larger scores are selected to generate the slides.

C-lexrank. C-Lexrank [12] is a clustering-based model in which the cosine similarities of sentences pairs are used to build a network of sentences. Based on the similarity network, C-Lexrank employs [1], a hier- archical agglomeration algorithm which works by greedily optimizing the modularity for sparse graphs.

CONCLUSION

We train a sentence scoring model based on SVR and use the ILP method to align and extract key phrases and sentences for generating the slides. Experimental results show that our method can generate much better slides than traditional methods. The proposed system consider one typical style of slides that beginners usually use. In the future, more complicated styles of slides such as the styles that are not aligned sequentially with the paper and styles that slides have more hierarchies. We will also try to extract the slide skeletons from the human-written slides and apply these slide skeletons to the automatic generated slides.

References

[1] A. Clauset, M. E. Newman, and C. Moore, “Finding community structure in very large networks.” Phys. Rev. E, vol. 70, no. 6, p. 066111, 2004.

[2] D. Galanis and P. Malakasiotis, “AUEB at TAC 2008,” in Proc. Text Anal. Conf., 2008.

[3] D. Marcu, “From discourse structures to text summaries,” in Proc. ACL Workshop Intell. Scalable Text Summarization., 1997, vol. 97, pp. 82–88.

[4] 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. 1–4.

[5] G. Erkan and D. R. Radev, “LexPageRank: Prestige in multi-docu- ment text summarization,” in Proc. EMNLP, 2004, pp. 365–371.

[6] 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.

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

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

[9] O. Yeloglu, M. Evangelos, and Z.-H. Nur, "Multi-document sum- marization of scientific corpora," in Proc. ACM Symp. Appl. Com- put., 2011, pp. 252–258.

[10] P. B. Baxendale, "Machine-made index for technical literature: an experiment," IBM J. Res. Develop., vol. 2, no. 4, pp. 354–361, 1958.

[11] R. McDonald, "A study of global inference algorithms in multi- document summarization," in Proc. Eur. Conf. Inf. Retrieval, 2007, pp. 557–564.

[12] 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.

[13] 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.

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