Course Project Report

Document content extraction for slide generation

Submitted By

Bhagyashri Bhamare (181IT111)
C Sneha(181IT112)
Karthik R(181IT235)

as part of the requirements of the course

Information Retrieval (IT458) [Jul - Nov 2021]

in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology in Information Technology

under the guidance of

Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

undergone at

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL

JUL-NOV 2021

DEPARTMENT OF INFORMATION TECHNOLOGY

National Institute of Technology Karnataka, Surathkal

CERTIFICATE

This is to certify that the Course project Work Report entitled **"Document content extraction for slide generation"** is submitted by the group mentioned below -

Details of Project Group

Name of the Student	Register No.	Signature with Date
1. Bhagyashri Bhamare	181IT111	Bhagyashri Bhamare (21/11/21)
2. C Sneha	181IT112	C Sneha (21/11/21)
3. Karthik R	181IT235	Karthik R (21/11/21)

this report is a record of the work carried out by them as part of the course **Information Retrieval (IT458)** during the semester **Jul - Nov 2021**. It is accepted as the Course Project Report submission in the partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Information Technology.**

(Name and Signature of Course Instructor)
Dr. Sowmya Kamath S

DECLARATION

We hereby declare that the project report entitled **"Document content extraction for slide generation"** submitted by us for the course **Information Retrieval (IT458)** during the semester **Jul-Nov 2021**, as part of the partial course requirements for the award of the degree of Bachelor of Technology in Information Technology at NITK Surathkal is our original work. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Details of Project Group

Name of the Student	Register No.	Signature with Date
1. Bhagyashri Bhamare	181IT111	Bhagyashri Bhamare (21/11/21)
2. C Sneha	181IT112	C Sneha (21/11/21)
3. Karthik R	181IT235	Karthik R (21/11/21)

Place: NITK, Surathkal
Date: 21/11/21

Document content extraction for slide generation

Bhagyashri Bhamare[1], C Sneha[2], Karthik R [3]

Abstract— **Slides are an easy and effective way to present information to an audience, especially in conferences and presentations. It is quite a challenge to read and summarize an academic paper within a short span of time. A lot of effort has to be put in and it consumes a large amount of time to extract the most important and relevant information for the paper. Little research has been done to automate the document to slide generation process. In our work we present a system for extraction of content for slides which can be compiled in an appealing and concise manner in a slide deck. The module for labeling sentences is based on SummaRuNNer, which is a Recurrent Neural Network (RNN) based on a sequence model for extractive summarization. Rather than ranking the sentences in the whole document based on similarities that are semantic, the algorithm used measures the novelty of sentences along with their importance by considering a sentence window and combining lexical and semantic features within it. We then use the ROUGE score to evaluate the output.**

Keywords: **Report to presentation, text summerization**

I. INTRODUCTION

Presentations are used in a variety of settings, from business to education to research, since they are visually successful in summarising and explaining large amounts of information to an audience. Such slides are commonly created to have bullet points that the presenter believes are important which act as reminders for the presenter as well as summerizes the information for the audience for better understanding. Manually preparing the presentation slides speacially when presenting a research document or report is tiresome and time-consuming.

Automation in the process of creating slides when a document with the information exists would save a lot of precious time and effort. Our project can help with the above mentioned problem by selecting the important sentences that should be added to the slides. It reduces the time taken to pick and choose the sentences from the document.

Abstractive and extractive methods are main ways for summarization. Abstractive approaches summarize text using words not necessarily appearing in the original document. Extractive summarization focuses on identifying important constituents usually at the sentence level and connecting them to generate a text snippet, which is usually significantly shorter than the original However, getting the main points from a research paper or report is a big challenge. This is because the methods that exist to encode implicit relation between sentences as well as the semantics of the sentences have limitations. We have used an extractive summarizer that considers a window of consecutive sentences and finds the most useful sentence among them based on their novelty and importance. The nun phrases that are frequent along with the selected sentences are then made into the points for the slide by structuring them in a layered format.

Multiple bullet points are frequently structured in a presentation to create a multi-level hierarchical format on presentation slides. This is done with points describing the high-level subjects first and points at the next levels which provide more explanation or specifics on the initial point. According to statistical analysis of the training dataset, over 93 percent of the points in the presentation are present in the initial two levels layers, with only 7 percent in the later layers. As a result, we have only used two levels of bullet points for the presentation slides.

Our project is a system that creates the content for the presentation slides for research papers or reports by considering the high scoring sentences. We have used a corpus which contains around 5000 pairs of papers and slides for the purpose of training and testing our model.

II. LITERATURE REVIEW

A. Automation of Report-To-Presentation Generation

Early attempts to automatically generate presentation slides extend back to more than 2 decades and focus on heuristic rule-based ways of information processing from web searches to create contents of a slide for a user-entered topic employed web sources, established schemas, and rule-based heuristics to produce decks that are random based on a specific theme as one recent example. Different sorts of rules were utilised in this collection of works, but all of them mainly relied on handcrafted characteristics or heuristics .

Researchers have lately begun to use machine learning algorithms to lean the significance of key phrases and sentences. Regression, random forest, and deep neural networks are some of the methods used to rank sentence importance. In addition, they use integer linear programming for sentence selection.

B. Text Summarization

In paper[1], the authors have implemented a 2 step approach to convert documents to slides. They used slide titles to retrieve text, figures and tables. In they next step they summarize the content into bullets. They used long form Question Answering for this purpose. They use a hierarchical model to extract slide header from the documents and correspondingly the body of the slide is extracted from the paper based on the level of importance and summarised to make it both conextually understandable and evenly divided among multiple slides instead of assuming a one to one mapping between heading and slide as done by earlier methods.

The standard way of text summarization[3] concentrates on constructing a paragraph with the summary of the text out of a longer document, is not the same as summarising research papers or reports for creating the slides of a presentation. Automation of the generation of slides can be accomplished by extracting the important sentences in a hierarchical order and then organising them into slides that are coordinated with the document (i.e. the research paper or report taken) in a sequential sequence.

Presentation slides have been created from scientific articles with the PPS[6] framework[6]. To pick key sentences and rank them, they used Support Vector Regressor and Integer Linear Programming (ILP). To create the structure of bullet points, Wang et al. extracted some phrases from the texts and used it to understand the hierarchical link that exist between pairs of words.

SummaRuNNer is a neural extractive summarizer (Nallapati et al., 2017) that approaches the summarising challenge as a sequence labelling problem. The CNN/Daily Mail corpus, which contains news pieces that are shorter than academic papers, was used to test SummaRuNNer. For the summary of scientific papers, we have used an improved version of the SummaRuNNer model.

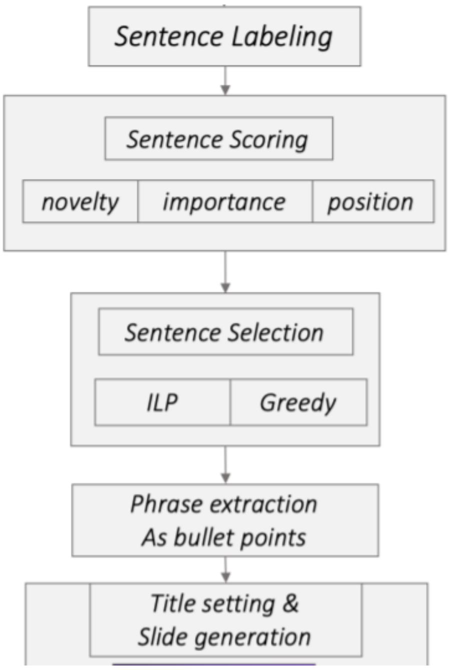


Fig. 1: System architecture

III. Methodology

The process of creating slides consists of there key parts that begin with recognising the salient sentences from the document or report taken. The initial step is to find the sentences that are most important from the document that are comparable to the content in the slide. The next step is to train the model so that it ranks the sentences. The last step is to choose the most important sentences based on the scores that are predicted, the sentence length and the summary size. The hierarchical structure of the bullet points can then be shaped by getting the noun phrases which are frequent from the texts selected.

A. *Sentence Labeling*

It's possible that the text in manually prepared slides was not retrieved straight from the source publication. Text can be reduced, summarised, or rewritten instead.

As a result, extractive labels for sentences in the input document are required. The phrase tagging process tries to find important sentences that are semantically related to the slides. This produces an extracted summary that will serve as the basis for training and evaluation. Every research paper will be represented as a list of sentences and each will be assigned a label 0 or 1. The system will predict the label as 1 when the sentence is to be included in the slide. Summarisation is treated as a sequence labeling problem by SummaRuNNer; if when we add a sentence to the summary there is an increase in the ROUGE score, it is labelled with a 1(one), else it is labelled with a 0(zero). A hierarchical structure of the sections is common in articles relted to research and education. As a part of the overall summary of the document, each section should have its own summary.

As a result, the labelling procedure should be tweaked to evenly distribute positive labels throughout the paper. It is, however, difficult to accurately pars sections of open-access scholarly papers. As a result, we suggest a windowed labelling strategy in which rating is done exclusively within a set of non-overlapping text windows, each containing w consecutive phrases. If adding the current sentence enhances the ROUGE-1 index, the sentence is designated as 1. By experimenting with different window widths and calculating the ROUGE score between selected sentences and the presentation slides, the ideal window size can be established empirically.

B. *Sentence and Document Embedding*

Sentences are ranked according to their importance, novelty, and content resemblance to the ground truth. A document is represented as a vector to quantify these features. We look at the following approach of Simple Document Embedding.

Calculating the average of the sentence encodings that are generated by using a Bidirectional Long Short-Term Memory(BiLSTM) yields a simple document embedding. We use the forward and backward hidden states of the last token in every sentence to encode it. The embedding for the whole document would be the mean of all sentence embedding with RELU as activation function.

C. *Ranking the Sentences*

The novelty, position and importance of a sentence in relation to the previously picked sentences determine its rank in the document. A sentence's salience is a measure of its importance. The uniqueness of a sentence in relation to the current summery is represented by the novelty of the

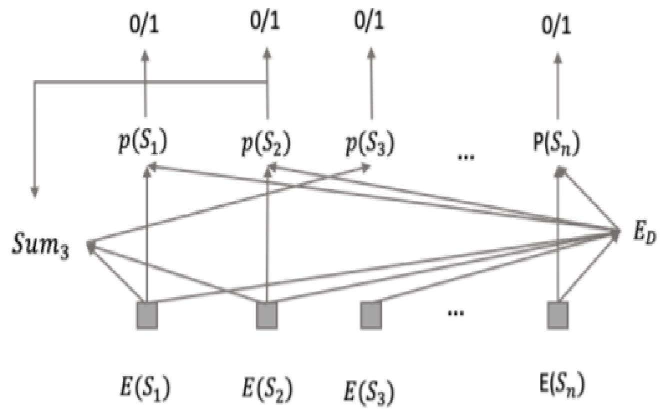


Fig. 2: Score Prediction

sentence. The summary embedding is the weighted sum of the previous sentences added to summary until the current sentence. The greater the chance of adding the sentence to the summary will give a larger portion in the summary embedding.

IV. RESULTS

When we're dealing with languages, measuring the outcomes of our model's outputs becomes a challenge. For Gisting Evaluation, we employ Recall-Oriented Understudy (ROUGE). Rather than being a single metric, ROUGE is a collection of metrics. ROUGE-N counts the number of 'n-grams' that match our model-generated text and a 'reference'.

An n-gram is a collection of tokens or words. A single word constitutes a unigram (1-gram). Two successive words make up a bigram (2-gram). We now need to determine whether we want to calculate the ROUGE recall, precision, or F1 score once we've decided which N to use. Here we use the F1 score for computing our results. To calculate F1 score we need the precision and recall values.

The recall divides the total number of n-grams in the reference by the number of overlapping n-grams identified in both the model output and the reference. Precision metric is calculated in almost the exact same way, but rather than dividing by the reference n-gram count, we divide by the model n-gram count.Finally the formula to calculate recall is:

$$f1 = \frac{2 * precision * recall}{precision + recall} \quad (1)$$

TABLE I: Various ROUGE average metrics

Metric	Recall	Precision	F-score
ROUGE-1	0.63	0.80	0.67
ROUGE-2	0.47	0.60	0.50
ROUGE-3	0.41	0.54	0.45
ROUGE-4	0.39	0.51	0.43

TABLE II: ROUGE score with respect to the window sizes

Size	ROUGE-1	ROUGE-2
7	44.43	11.37
10	45.62	11.83

V. CONCLUSIONS

In our work we have presented our methodology for extracting text and summarisizng to generate a slide given a research or academic paper. We used SummaRuNNer as our base model for the question answering task and RELU as our activation function filter out the final output. In future we look to explore other models and expand our dataset to other languages such as Chinese and Russian as most academic works and published in these languages.

REFERENCES

[1]. Edward Sun, Yufang Hou, Dakuo Wang, Yufeng Zhang, Nancy X.R. Wang. Document to slide generation using query based text-summarization. (arXiv: 2021).

[2]. K. Gokul Prasad, H. Mathivanan, T. V. Greetha and M. Jayaprakasam, Document Summarization and Information Extraction for Generation of Presentation Slides,2009, pp. 126-128, doi: 10.1109/ARTCom.2009.74

[3]. Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 3009–3019. Association for Computational Linguistics.

[4]. Yue Hu and Xiaojun Wan. 2014. Ppsgen: Learningbased presentation slides generation for academic papers. IEEE transactions on knowledge and data engineering, 27(4):1085–1097

[5]. Ather Seifid, Jian Wu, Lee Giles "Automatic slide generation for scientific papers." SciKnow '19, November 19–22, 2019, Marina del Rey, CA

[6]. VWu, Tiantian Yu, Hongzhi Wan, Fucheng Yang, Fangtao. (2019). Research on Answer Extraction for Automatic Question Answering System. 10.2991/iccia-19.2019.34

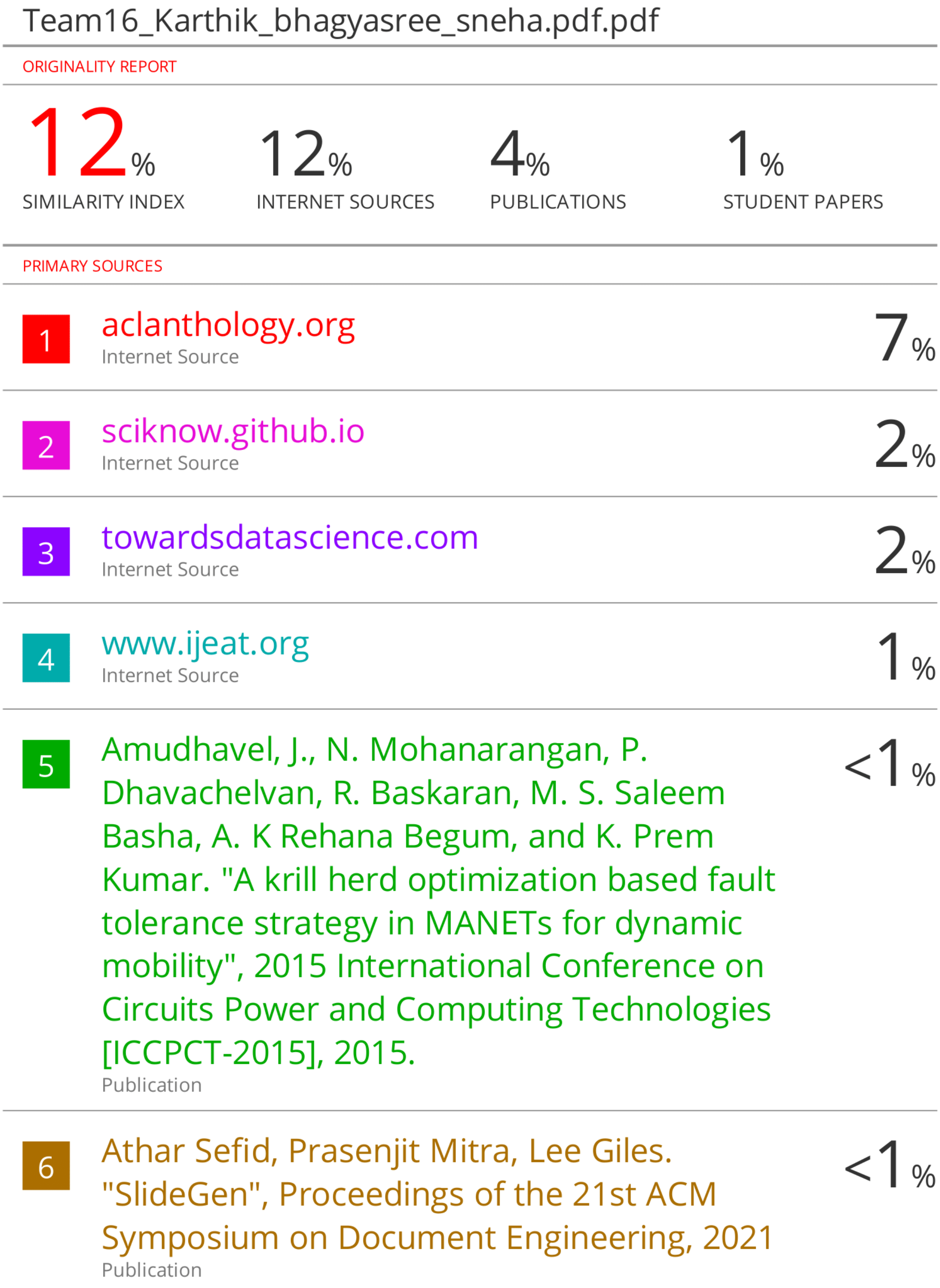
[7]. Sharma, Lokesh Mittal, Namita. (2018). Answer Extraction in Question Answering using Structure Features and Dependency Principles.

[8]. Raju Barskar, Gulfishan Firdose Ahmed, Nepal Barskar,'An Approach for Extracting Exact Answers to Question Answering (QA) System for English Sentences',Procedia Engineering, Volume 30, 2012, ISSN 1877-7058,

[9]. Kamal Al-Sabahi, Zhang Zuping, and Mohammed Nadher. 2018. A hierarchical structured selfattentive model for extractive document summarization (hssas). IEEE Access, 6:24205–24212.

[10]. Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. The SIGNLL Conference on Computational Natural Language Learning (CoNLL), 2016.

APPENDIX



Team16_Karthik_bhagyasree_sneha.pdf.pdf

ORIGINALITY REPORT

12%	12%	4%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	aclanthology.org Internet Source	7%
2	sciknow.github.io Internet Source	2%
3	towardsdatascience.com Internet Source	2%
4	www.ijeat.org Internet Source	1%
5	Amudhavel, J., N. Mohanarangan, P. Dhavachelvan, R. Baskaran, M. S. Saleem Basha, A. K Rehana Begum, and K. Prem Kumar. "A krill herd optimization based fault tolerance strategy in MANETs for dynamic mobility", 2015 International Conference on Circuits Power and Computing Technologies [ICCPCT-2015], 2015. Publication	<1%
6	Athar Sefid, Prasenjit Mitra, Lee Giles. "SlideGen", Proceedings of the 21st ACM Symposium on Document Engineering, 2021 Publication	<1%