Automatic Generation of Introduction and Abstract for Research Papers

219354V - R.P.D. Kumarasinghe

Supervisor: Dr. Nisansa de Silva

Overview

1. Introduction
2. Research Problem
3. Research Objectives
4. Literature Survey
5. Methodology
6. References

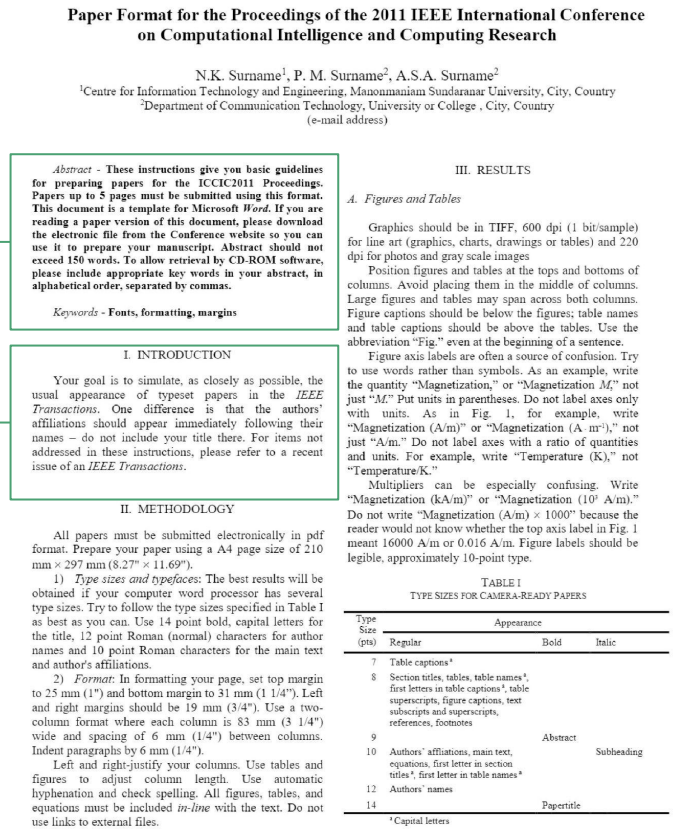
Introduction

Introduction

- The abstract of a research paper provides a quick summary of the entire paper from problem to solution to the result
- The Introduction section provides a primer to the rest of the paper by summarising the goals and the setting of the research while expanding on the basis established by the abstract

Abstract

Introduction



Paper Format for the Proceedings of the 2011 IEEE International Conference on Computational Intelligence and Computing Research

N.K. Surname[1], P. M. Surname[2], A.S.A. Surname[2]
[1]Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, City, Country
[2]Department of Communication Technology, University or College , City, Country
(e-mail address)

Abstract - **These instructions give you basic guidelines for preparing papers for the ICCIC2011 Proceedings. Papers up to 5 pages must be submitted using this format. This document is a template for Microsoft *Word*. If you are reading a paper version of this document, please download the electronic file from the Conference website so you can use it to prepare your manuscript. Abstract should not exceed 150 words. To allow retrieval by CD-ROM software, please include appropriate key words in your abstract, in alphabetical order, separated by commas.**

Keywords - **Fonts, formatting, margins**

I. INTRODUCTION

Your goal is to simulate, as closely as possible, the usual appearance of typeset papers in the *IEEE Transactions*. One difference is that the authors' affiliations should appear immediately following their names – do not include your title there. For items not addressed in these instructions, please refer to a recent issue of an *IEEE Transactions*.

II. METHODOLOGY

All papers must be submitted electronically in pdf format. Prepare your paper using a A4 page size of 210 mm × 297 mm (8.27" × 11.69").

1) *Type sizes and typefaces*: The best results will be obtained if your computer word processor has several type sizes. Try to follow the type sizes specified in Table I as best as you can. Use 14 point bold, capital letters for the title, 12 point Roman (normal) characters for author names and 10 point Roman characters for the main text and author's affiliations.

2) *Format*: In formatting your page, set top margin to 25 mm (1") and bottom margin to 31 mm (1 1/4"). Left and right margins should be 19 mm (3/4"). Use a two-column format where each column is 83 mm (3 1/4") wide and spacing of 6 mm (1/4") between columns. Indent paragraphs by 6 mm (1/4").

Left and right-justify your columns. Use tables and figures to adjust column length. Use automatic hyphenation and check spelling. All figures, tables, and equations must be included *in-line* with the text. Do not use links to external files.

III. RESULTS

A. Figures and Tables

Graphics should be in TIFF, 600 dpi (1 bit/sample) for line art (graphics, charts, drawings or tables) and 220 dpi for photos and gray scale images

Position figures and tables at the tops and bottoms of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table names and table captions should be above the tables. Use the abbreviation "Fig." even at the beginning of a sentence.

Figure axis labels are often a source of confusion. Try to use words rather than symbols. As an example, write the quantity "Magnetization," or "Magnetization *M*," not just "*M*." Put units in parentheses. Do not label axes only with units. As in Fig. 1, for example, write "Magnetization (A/m)" or "Magnetization (A·m^{-1})," not just "A/m." Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)," not "Temperature/K."

Multipliers can be especially confusing. Write "Magnetization (kA/m)" or "Magnetization (10^3 A/m)." Do not write "Magnetization (A/m) × 1000" because the reader would not know whether the top axis label in Fig. 1 meant 16000 A/m or 0.016 A/m. Figure labels should be legible, approximately 10-point type.

TABLE I
TYPE SIZES FOR CAMERA-READY PAPERS

Type Size (pts)	Appearance		
	Regular	Bold	Italic
7	Table captions[a]		
8	Section titles, tables, table names[a], first letters in table captions[a], table superscripts, figure captions, text subscripts and superscripts, references, footnotes		
9		Abstract	
10	Authors' affiliations, main text, equations, first letter in section titles[a], first letter in table names[a]		Subheading
12	Authors' names		
14		Papertitle	

[a] Capital letters

Research Problem

Research Problem

- Abstract and Introduction are expected to be concise and informative.
- But generating them manually is difficult and time consuming.
- Summarization has domain specific training approaches which are perform well.
    - But summarization for research papers in “Computational Linguistics” domain is not yet addressed

Research Objectives

Research Objectives

1. Creating a sufficient data set for the task of Abstract and Introduction generation in the computational linguistic domain
2. Evaluating existing state-of-the art solutions of text summarization technologies on the above data set and other comparable data sets.
3. Creating automatic summarization models capable of Abstract And Introduction generation in the computational linguistic domain.
4. Creating an online application which, when given the LATEX source sans the Abstract and Introduction, generates these sections automatically.

Literature Survey

Literature Survey

According to the “Modeling document summarization as multi-objective optimization [1]”

Main objective of summarization,

1. Coverage of information
2. Information significance
3. Redundancy in information
4. Cohesion in text

Literature Survey: Validation Mechanisms

1. BLEU [2]

The closer a machine translation is to a professional human translation, the better it is

$$\text{BP} = \begin{cases} 1 & if c > r \\ e^{1-r/c} & if c \leq r \end{cases}$$

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

2. ROUGE [3]

- ROUGE-N
- ROUGE-L
- ROUGE-W
- ROUGE-S
- ROUGE-SU

Literature Survey: Related Review Papers

Paper Title	Authors	Year	Citations
A survey on automatic text summarization [4]	Nazari and Mahdavi	2019	20
Automatic text summarization: A comprehensive survey [5]	El-Kassas et al.	2021	55
A survey on extractive text summarization [6]	Moratanch and Gopalan	2017	97
Text Summarization Techniques: A Brief Survey [7]	Allahyari et al.	2017	326
Recent automatic text summarization techniques: a survey [8]	Gambhir and Gupta	2017	466
A survey on abstractive text summarization [9]	Moratanch and Chitrakala	2016	69

Literature Survey: Building Blocks

According to Jones et al. [10] and Hovy et al. [11]

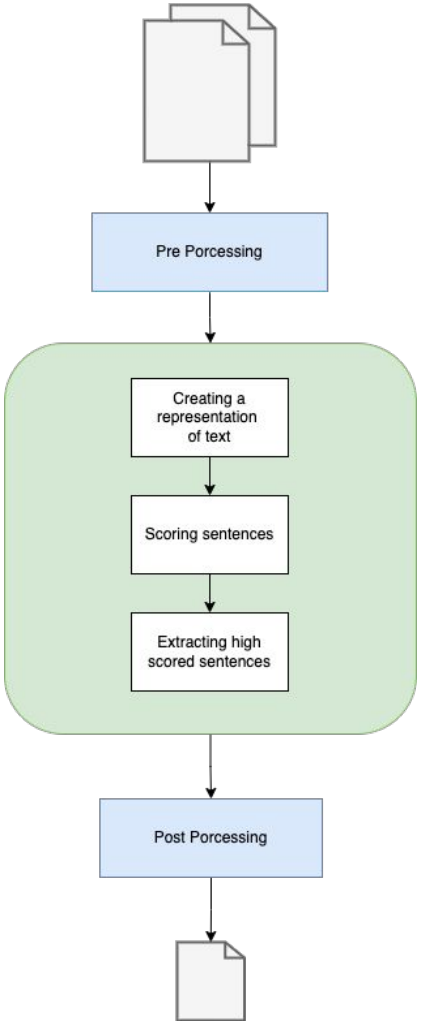
Three steps break down of the building blocks of the structural components in automatic text summarization,

1. Identification
2. Interpretation
3. Generation

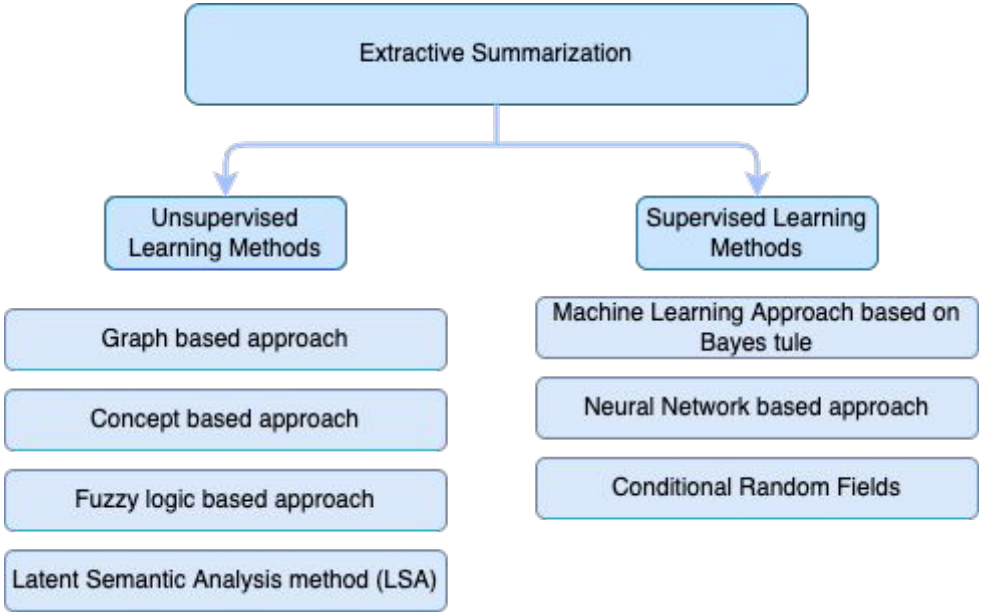
Literature Survey: Building Blocks

El-Kassas et al. [5] have classified the summarization systems considering different aspects

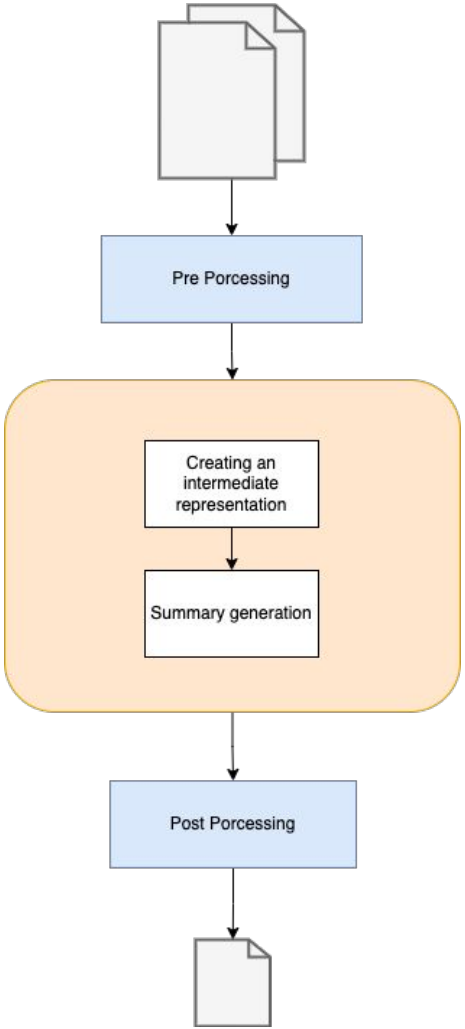
Literature Survey: Approach > Extractive



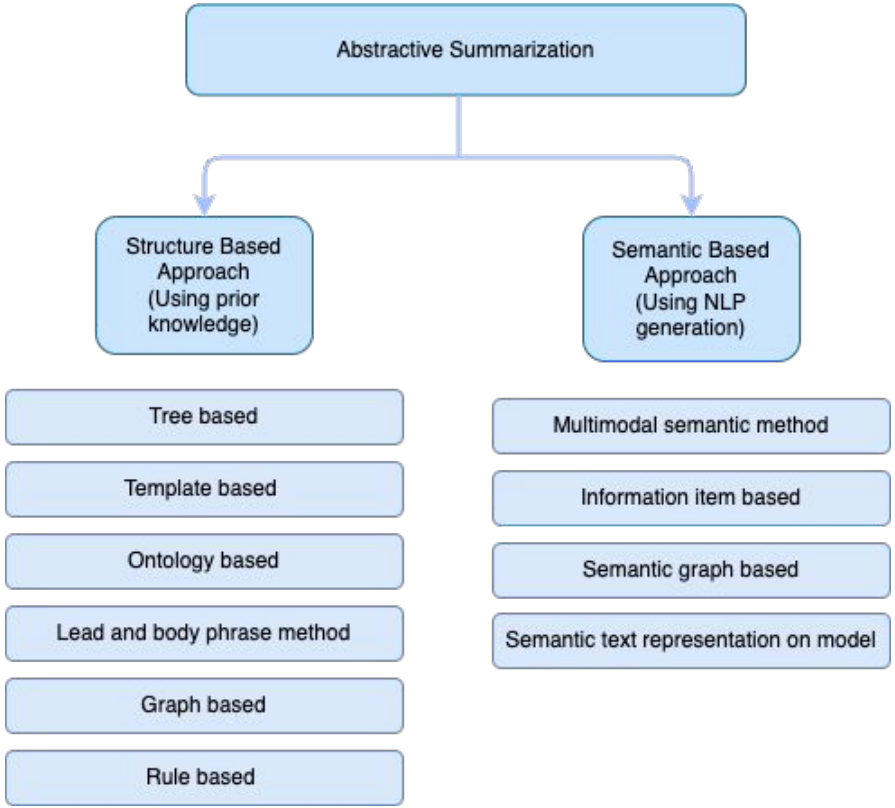
Literature Survey: Approach > Extractive



Literature Survey: Approach > Abstractive



Literature Survey: Approach > Abstractive

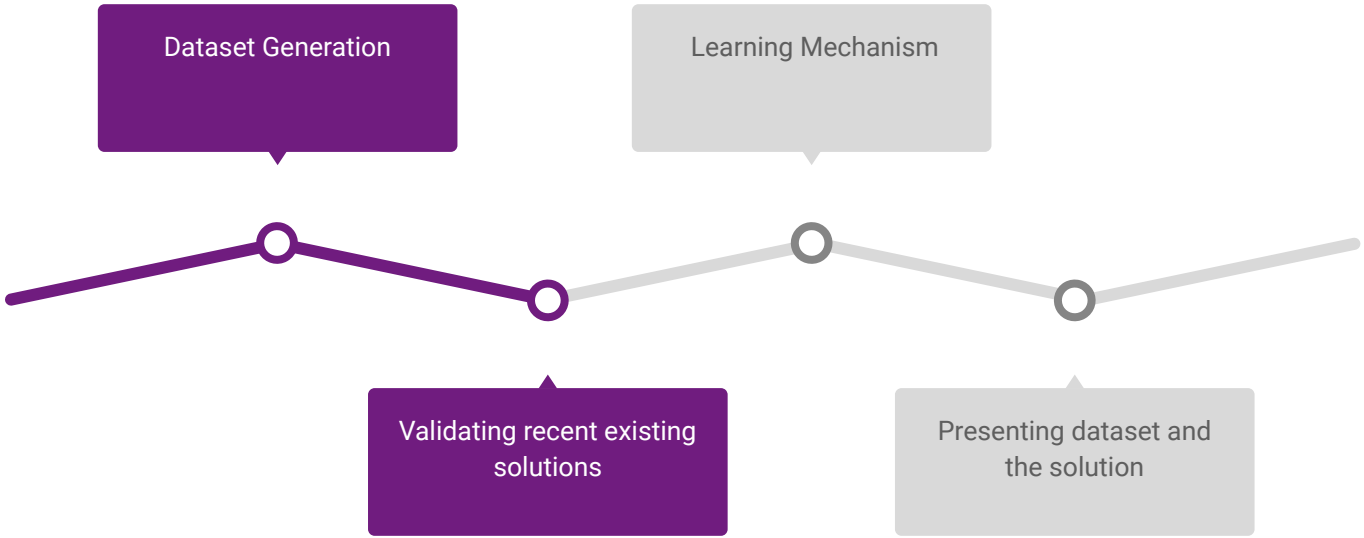


Literature Survey: Datasets

Dataset	No of articles	Details
CNN/Daily Mail dataset	More than 300k	Average summary: ● 53 words ● 3.72 sentences
NYT	More than 1.8 million	Articles from 1987 to 2007
XSum	More than 220k	BBC articles with single sentence summaries
Newsroom dataset	1.3 million (1.2 million publicly available)	From 38 major publications
Bytecup dataset	1.3 million (1.1 million released for training)	2018 Bytecup ML contest dataset
ArXiv Dataset	215K	Average: ● Doc words: 4938 ● Summary words: 220
PubMed Dataset	133K	Average: ● Doc words: 2016 ● Summary words: 203

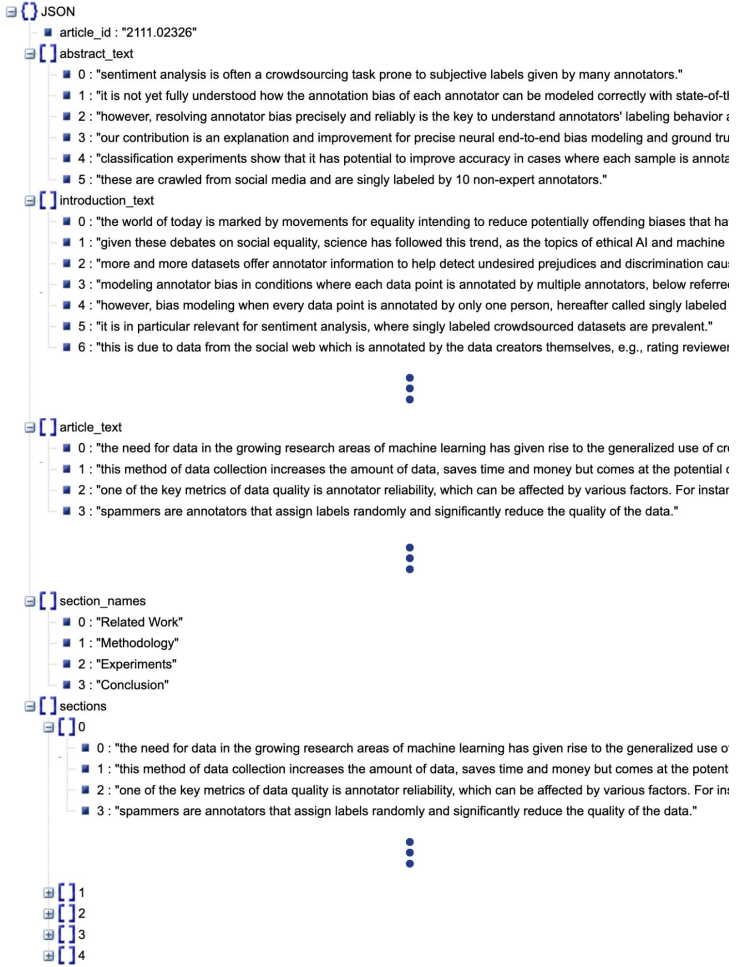
Methodology

Methodology: Backlog



Methodology: Dataset Generation

- Data from arxiv
- JSON Lines files
  - Fields
    - article_id
    - abstract_text
    - introduction_text
    - article_text
    - section_names
    - sections



Methodology: Validating state of the art solutions against the dataset

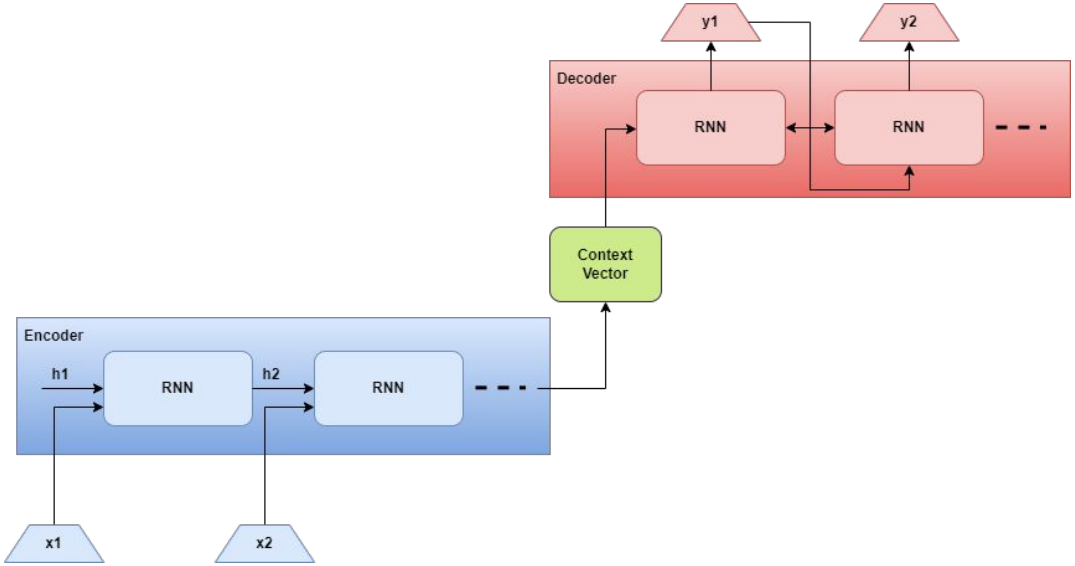
Main focus is Cohan et al. [12]

> discourse aware model for abstractive summarization

	RG-1	**RG-2**	**RG-3**	**RG-L**
arXiv dataset	35.80	11.05	3.62	31.80
PubMed dataset	38.93	15.37	9.97	35.21

Methodology: Learning Mechanism

- Sequence to Sequence (seq2seq) encoder decoder architecture



References

References

[1] L. Huang, Y. He, F. Wei, and W. Li, “Modeling document summarization as multi-objective optimization,” 04 2010, pp. 382–386

[2] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[3] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in Text summarization branches out, 2004, pp. 74–81.

[4] N. Nazari and M. Mahdavi, “A survey on automatic text summarization,” Journal of AI and Data Mining, vol. 7, no. 1, pp. 121–135, 2019. [Online]. Available: http://jad.shahroodut.ac.ir/article_1189_162.html

[5] W. S. El-Kassas, C. R. Salama, A. A. Rafea and H. K. Mohamed, “Automatic text summarization: A comprehensive survey,” Expert Systems with Applications, vol. 165, p. 113679, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420305030

[6] N. Moratanch and C. Gopalan, “A survey on extractive text summarization,” 01 2017, pp. 1–6. [Online]. Available: https://www.researchgate.net/publication/317420253_A_survey_on_extractive_text_summarization

[7] M. Allahyari, S. A. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. J. Kochut, “Text summarization techniques: A brief survey,” CoRR, vol. abs/1707.02268, 2017. [Online]. Available: http://arxiv.org/abs/1707.02268

[8] M. Gambhir and V. Gupta, “Recent automatic text summarization techniques: a survey,” Artificial Intelligence Review, vol. 47, 01 2017. [Online]. Available: https://www.researchgate.net/publication/299499824_Recent_automatic_text_summarization_techniques_a_survey

[9] N. Moratanch and S. Chitrakala, “A survey on abstractive text summarization,” in 2016 International Conference on Circuit, power and computing technologies (ICCPCT). IEEE, 2016, pp. 1–7.

[10] K. S. Jones et al., “Automatic summarizing: factors and directions,” Advances in automatic text summarization, pp. 1–12, 1999.

[11] E. Hovy, C.-Y. Lin et al., “Automated text summarization in summarist,” Advances in automatic text summarization, vol. 14, pp. 81–94, 1999.

[12] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, “A discourse-aware attention model for abstractive summarization of long documents,” arXiv preprint arXiv:1804.05685, 2018.

Thank You