Abstractive Document Summarization Using Machine Learning Approach

Anuja Parve [1], Purushottam Nimje [2], Mayur Kardile[3], Pratik Bachche[4], Kanchan Pradhan[5]

U.G. Students, Department of Computer Engineering, JSPM's Bhivrabai Sawant Institue of Technology and Research, Wagholi, Pune, Maharashtra, India[1,2,3,4]

Associate Professor, Department of Computer Engineering, JSPM's Bhivrabai Sawant Institue of Technology and Research, Wagholi, Pune, Maharashtra, India[5]

ABSTRACT: We present a novel divide-and-conquer strategy for the neuralsummary of large documents. Our solution makes use of thedocument's discourse structure and sentence similarity. Divide thetask into smaller summarization challenges. We divide a lengthy document and its summary into numerous source-target pairs,which are then used to train a model. learns to sum up each section of the document individually After The partial summaries are then combined to form a final comprehensive summary. In contrast to the conventional approach, we may break down the challenge of long document summaries into smaller pieces and simpler problems, reducing computational complexity and increasing the number of training examples while also reducing There is noise in the target summaries. There are two publicly available datasets.

KEYWORDS: Complexity, Contrast, Summarization, Machine Learning, Classification, NLP(Natural Language Processing).

I. Introduction

In this present era, where huge quantity of information is generating on the internet day by day. So, it is necessary to provide a better mechanism to extract useful information fast and effectively. Text summarization is one of the methods of identifying the important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meaning. It reduces the time required for reading whole documents and also it is a space problem that is needed for storing a large amount of data. The automatic text summarization problem has two sub-problems that is a single document and multiple documents. In the single document, the single document is taken as the input and summarized information is extracted from that particular single document. In Multiple documents, the multiple documents of the single topic are taken as input and the output which is generated should be related to that topic.

Automatic Text summarization has two approaches 1) Abstractive text summarization and 2) Extractive text summarization. Anextractive text summarization means important information or sentence are extracted from the given text file or original document. An extractive text summarization approach uses linguistic or statistical features for selecting useful informative sentences. An Abstractive text summarization will try to understand the input file or original file and re-generate the output in a few words by identifying the main concept of the input file. In many research papers, they have mentioned that extractive text summarization is sentence ranking. Extractive text summarization isdivided into two phases: 1) Pre-processing and 2) Processing. In this paper, weare explaining extractive text summarization on a single document.

This thesis aims to replicate and extend knowledge of automatic text summarizing. The success of sequence-to-sequence (seq2seq) modelling in NLP has pushed rapid and significant growthin this field. However, the focus of the study has been on developing frameworks for short, single-document summarizing, rather than extended and multi-document summarization. The ultimate goal of this topic is to provide a framework that can produce high-quality summaries regardless of the length of the source documents, whether it's a single or multi-document assignment and irrespective of domainor language.

Given the recent impressive progress in a short document summarizing, the next challenge will be to reproduce the results usinglonger documents. The goal of this thesis is to contribute to the knowledge of expertise in the field of long document summaries. The evaluation of summaries and systems for Abstractive document summarizing are two particularly relevant areas that we identify as core issues in this thesis.

II. RELATED WORK

1.Paper name: Text Summarization using Sentiment Analysis for DUCData.
Author: Nidhika Yadav, Niladri Chatterjee
Abstract: The basic purpose of this study is to see if sentiments can be used to summarize material effectively. We describe a text summarization technique based on the emotions associated with key phrases that is computationally efficient. Sentiment analysis is already in use for large-scale text data interpretationand opinion mining in a variety of fields. The current research shows thatsentiment analysis can be utilized to summarize text as well. Our findings were compared to those of the conventional DUC2002 datasets.A text summary is a technique for determining a text's main points. In the literature, various strategies have been presented, and many ofthese are currently in use in commercial systems
The work in this paper is divided into two stages. 1) Text- Detection 2) Inpainting. Text detection is done by applying morphological open-close and close-open filters and combining the images. Thereafter, the gradient is applied to detect the edges followed by thresholding and morphological dilation, erosion operation. Then, connected component labelling is performed to label each object separately. Finally, the set of selection criteria is applied to filter out non-text regions. After text detection, text inpainting is accomplished by using exemplar based Inpainting algorithm.

2. Paper name: Multi-document Summarization by using Text Rank andMaximal Marginal Relevance for Text in Bahasa Indonesia.
Author: Multi-document Summarization by using Text Rank andMaximalMarginal Relevance for Text in Bahasa Indonesia
Abstract: The text summarizer eliminates unnecessary information by selecting the most important sentences. There's a chance that twoor more crucial statements in a multi-document summary will containthe same information. If those sentences are included in the summaryresult, the information will be redundant. The goal of this research isto condense similar lines from a variety of sources containing similarinformation into a more concise text summary. This study divides avariety of internet news articles into six groups to achieve its goal. The articles are concatenated and pre-processed to create clean text.Following the acquisition of clean text, the Text Rank algorithm is used to extract the important sentences based on similarity measurement. This procedure produced the summary text.

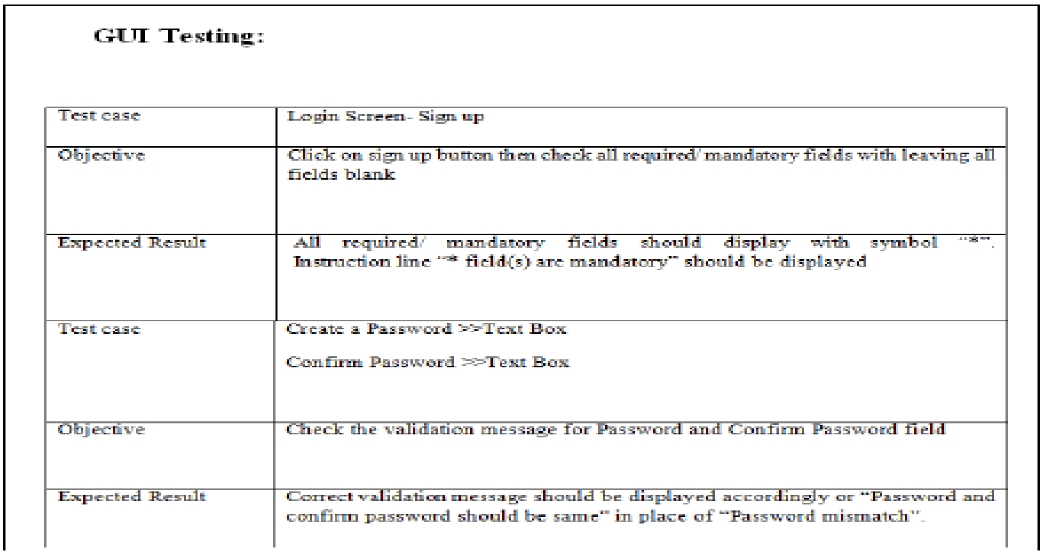
3. Paper name: Automatic Text Summarization by Local Scoringand Ranking for Improving Coherence
Author: P. Krishnaveni, Dr.S. R. Balasundaram.
Abstract: Due to a large amount of textual material available on theinternet, machine-generated summarization has received a lot of attention. Manually summarizing these online text documents is difficult for most people. As a result, we will require an automatic textsummarizer. Automatic Text Summarization (ATS) is defined as "shortening the original text while preserving its information value and overall meaning." Despite the fact that automatic text summarization has been studied since the 1950s, more cohesive and meaningful summaries still exist. The proposed technique generates anautomatic feature-based extractive heading-wise text summarizer that im-proves the summary text's coherence and, as a result, its understandability. It just summarizes the input document using local scoring and ranking, and that's it.

4. Paper name: A Divide-and-Conquer Approach to the Summarization of Long Documents.
Author: Alexios Gidiotis and Grigorios Tsoumakas
Abstract: We propose a novel divide-and-conquer strategy for the neural summarization of large documents. Our solution takes advantage of the document's discourse structure and divides the challenge into smaller summarization problems using sentence similarity. We divide a lengthy text and its summary into numerous source-target pairs, which are then used to train a model that learns to summarize each section of the document separately. The partial summaries are then combined to create a final full summary.

III. METHODOLOGY

- **TEST CASES**

1. GUI TESTING

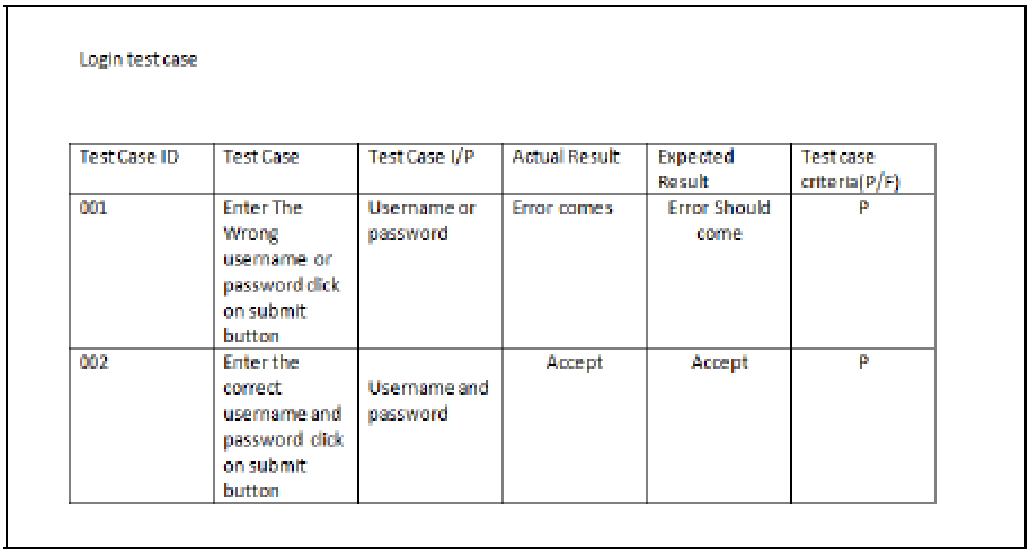


GUI Testing:

Test case	Login Screen- Sign up
Objective	Click on sign up button then check all required/ mandatory fields with leaving all fields blank
Expected Result	All required/ mandatory fields should display with symbol "*". Instruction line "* field(s) are mandatory" should be displayed
Test case	Create a Password >>Text Box Confirm Password >>Text Box
Objective	Check the validation message for Password and Confirm Password field
Expected Result	Correct validation message should be displayed accordingly or "Password and confirm password should be same" in place of "Password mismatch".

Figure 1: GUI Test Case

2. LOGIN TEST CASE

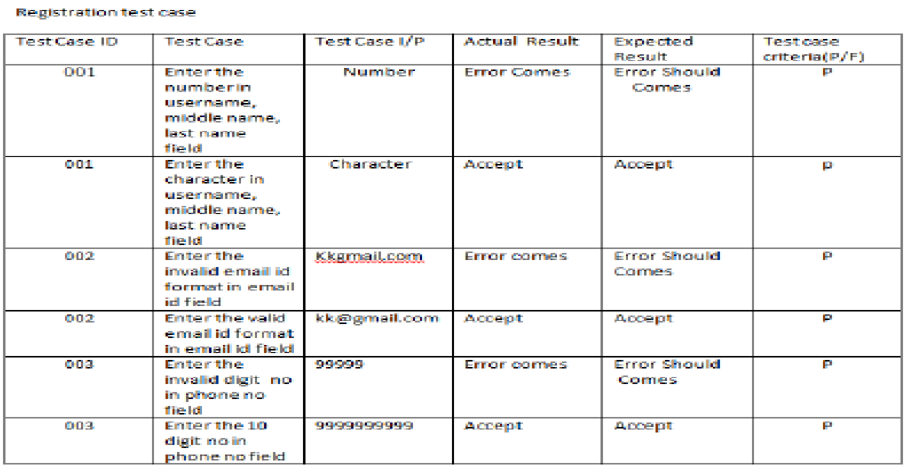


Login test case

Test Case ID	Test Case	Test Case I/P	Actual Result	Expected Result	Test case criteria(P/F)
001	Enter The Wrong username or password click on submit button	Username or password	Error comes	Error Should come	P
002	Enter the correct username and password click on submit button	Username and password	Accept	Accept	P

Figure 2: Login Test Case

3. Registration Test Case

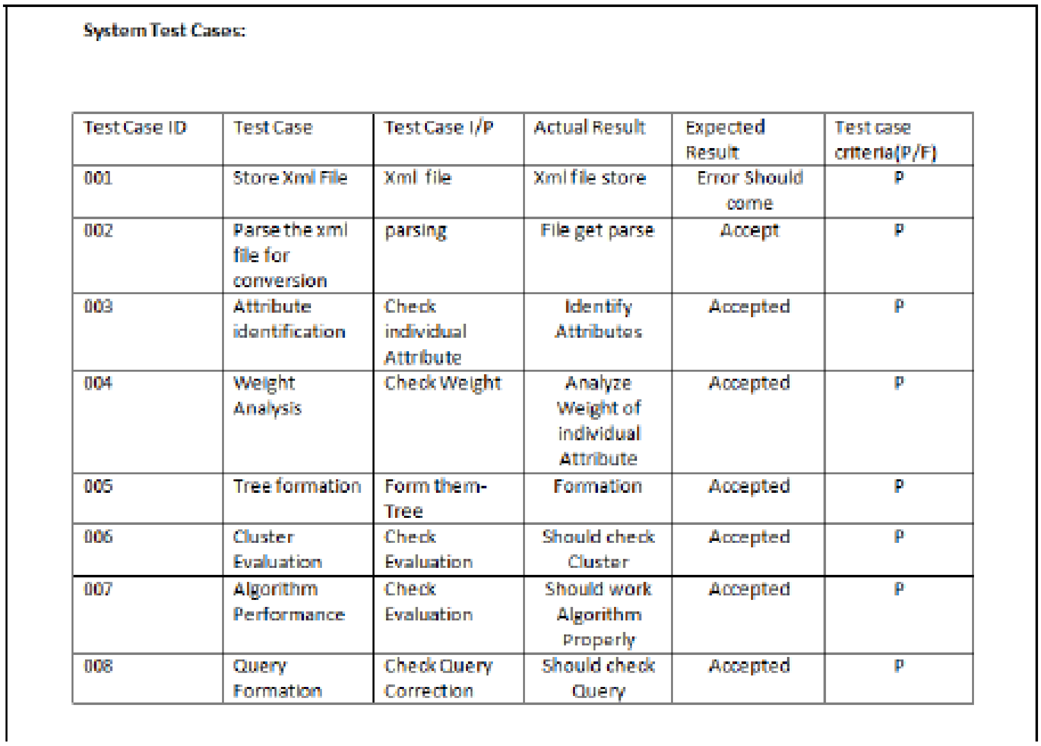


Registration test case

Test Case ID	Test Case	Test Case I/P	Actual Result	Expected Result	Test case criteria(P/F)
001	Enter the number in username, middle name, last name field	Number	Error Comes	Error Should Comes	P
001	Enter the character in username, middle name, last name field	Character	Accept	Accept	p
002	Enter the invalid email id format in email id field	Kkgmail.com	Error comes	Error Should Comes	P
002	Enter the valid email id format in email id field	kk@gmail.com	Accept	Accept	P
003	Enter the invalid digit no in phone no field	99999	Error comes	Error Should Comes	P
003	Enter the 10 digit no in phone no field	9999999999	Accept	Accept	P

Figure 3: Registration Test Case

4. System Test Cases



System Test Cases:

Test Case ID	Test Case	Test Case I/P	Actual Result	Expected Result	Test case criteria(P/F)
001	Store Xml File	Xml file	Xml file store	Error Should come	P
002	Parse the xml file for conversion	parsing	File get parse	Accept	P
003	Attribute identification	Check individual Attribute	Identify Attributes	Accepted	P
004	Weight Analysis	Check Weight	Analyze Weight of individual Attribute	Accepted	P
005	Tree formation	Form them-Tree	Formation	Accepted	P
006	Cluster Evaluation	Check Evaluation	Should check Cluster	Accepted	P
007	Algorithm Performance	Check Evaluation	Should work Algorithm Properly	Accepted	P
008	Query Formation	Check Query Correction	Should check Query	Accepted	P

Figure 4: System Test Cases

IV. RESULT

1. GUI MAIN PAGE

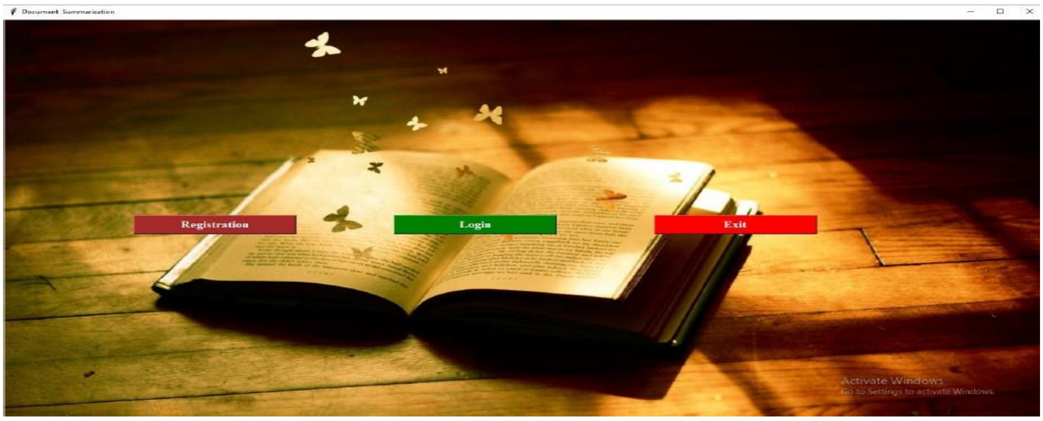
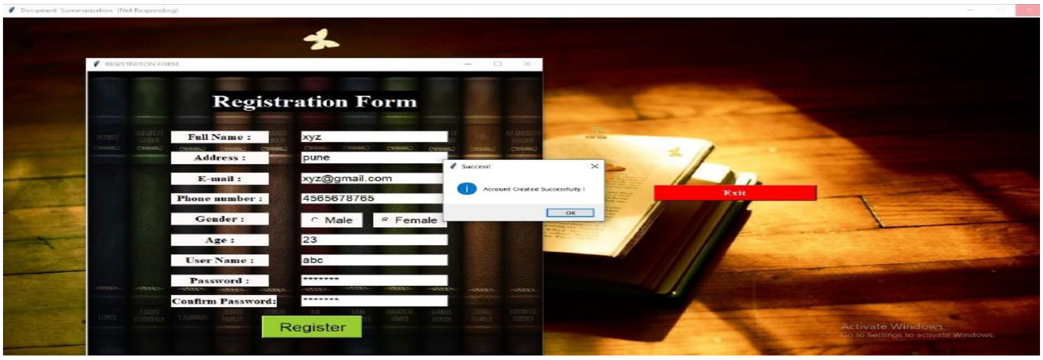


Figure 5: GUI Main Page

2. REGISTRATION


Document Summarization (Not Responding)
REGISTRATION FORM
Exit
Activate Windows
Go to Settings to activate Windows

Registration Form

Full Name : xyz

Address : pune

E-mail : xyz@gmail.com

Phone number : 4565678765

Gender : ○ Male ● Female

Age : 23

User Name : abc

Password : *******

Confirm Password: *******

Register

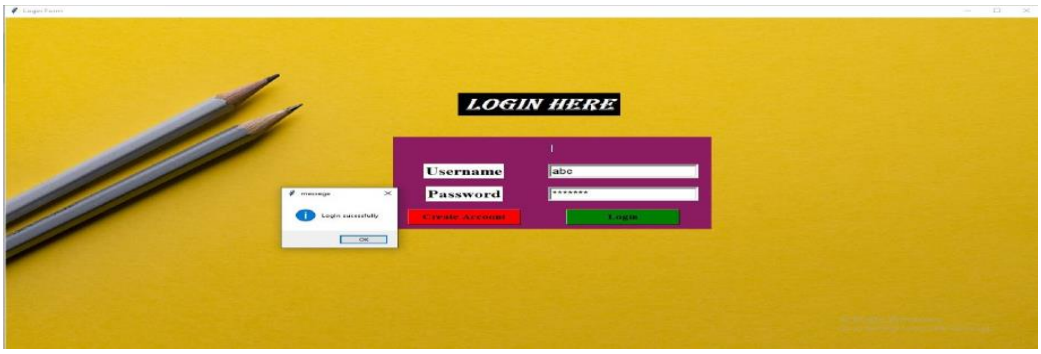
Success!

Account Created Successfully !

OK

Figure 6: Registration

3. LOGIN PAGE



LOGIN HERE

Username abc

Password *******

Create Account

Login

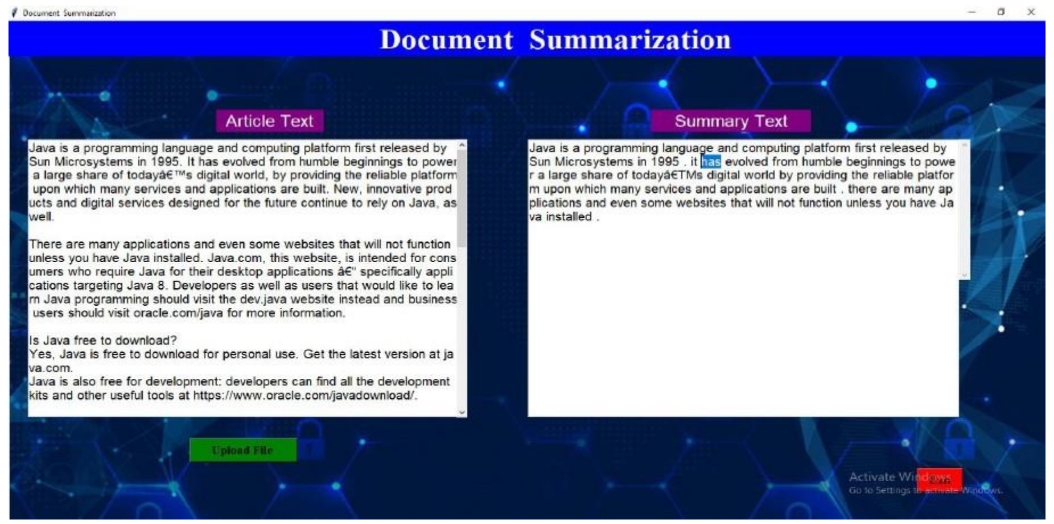
message

Login successfully

OK

Figure 7: Login Page

4. FINAL OUTPUT



Java is a programming language and computing platform first released by Sun Microsystems in 1995. It has evolved from humble beginnings to power a large share of todayâ€™s digital world, by providing the reliable platform upon which many services and applications are built. New, innovative products and digital services designed for the future continue to rely on Java, as well.

There are many applications and even some websites that will not function unless you have Java installed. Java.com, this website, is intended for consumers who require Java for their desktop applications â€" specifically applications targeting Java 8. Developers as well as users that would like to learn Java programming should visit the dev.java website instead and business users should visit oracle.com/java for more information.

Is Java free to download?
Yes, Java is free to download for personal use. Get the latest version at java.com.
Java is also free for development: developers can find all the development kits and other useful tools at https://www.oracle.com/javadownload/.

Java is a programming language and computing platform first released by Sun Microsystems in 1995 . it has evolved from humble beginnings to power a large share of todayâ€TMs digital world by providing the reliable platform upon which many services and applications are built . there are many applications and even some websites that will not function unless you have Java installed .

Figure 8: Result

V. CONCLUSION

Autonomy abstracted Document summarizing is a multi-step process with several sub-tasks. Every subtask has the ability to generate high-quality summaries. Abstractive document summarization requires identifying necessary paragraphs from a given document. We suggested an abstractive based text summary inthis research, based on a statistical new approach based on sentence ranking, in which the summarizer selects phrases based on their rating. Abstract sentences are created as a summarized text that isthen turned into audio. The proposed model outperforms the old technique in terms of accuracy.

REFERENCES

[1] R. Socher, "Boiling the information ocean," 2020. [Online].Available: http://tiny.cc/45ohlz
[2] S. Chopra, M. Auli, and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist.: Human Lang.Technol., 2016, pp. 93–98.
[3] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarizationwith pointer-generator networks," in Proc. Annu. Meet. Assoc. Comput.Linguist., 2017, pp. 1073–1083.
[4] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model forabstractive summarization," in Proc. Int. Conf. Learn. Representations, 2018.
[5] Y. Liu and M. Lapata, "Text summarization with pretrained encoders,"in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process., 2019, pp. 3721–3731.
[6] K. Song, X. Tan, T. Qin, J. Lu, and T. Y. Liu, "MASS: Masked sequenceto sequence pre-training for language generation," in Proc. Int. Conf. Mach.Learn., 2019, pp. 5926–5936.
[7] L. Dong et al., "Unified language model pre-training for natural language understanding and generation," in Adv. Neural Inform. Process.Syst., 2019, pp. 13 042–13 054.
[8] Y. Yan et al., "ProphetNet: Predicting future N-gram for sequence- to sequence pre-training," 2020, arXiv:2001.04063.
[9] K. M. Hermann et al., "Teaching machines to read and comprehend,"inAdv. Neural Inform. Process. Syst., 2015, pp. 1693–1701.
[10] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs andbeyond," in Proc. SIGNLL Conf. Comput. Natural Lang. Learn. Stroudsburg, PA,USA, 2016, pp. 280–290.