2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT) | 979-8-3503-0546-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/CSNT60213.2024.10546173

Fusion of Word Embedding and Encoder Decoder model for Text Summarization

Sachin Solanki, Suresh jain, Kailash Chandra Bandhu
Department of Computer Science & Engineering
Medi-Caps University, Indore
Indore, India
sol.sachin@gmail.com, suresh.jain@rediffmail.com, kailashchandra.bandhu@gmail.com

***Abstract*—Extractive text summarization is a core activity in the field of natural language processing, aiming to condense the most important information from a given text while preserving its core meaning. This study investigates a novel methodology that integrates word embeddings using GLOVE and LSTM based encoder decoder model, two widely recognized methodologies, to enhance the effectiveness of extractive summarization. This novel approach incorporates the advantages of both word embeddings and LSTM to enhance the summarization process. Word embeddings can capture semantic links between words and phrases, so facilitating a more profound comprehension of textual context. Encoder decoder-based LSTM model identifies predicted summary based on the original summary. BLEU and cosine similarity are metrics that evaluates the importance of terms in a collection of documents, so guaranteeing that crucial phrases are given suitable weighting. An algorithm is provided for the task of extracting text summarizing, wherein sentence embedding is achieved by following word embedding, and the score of the summary is obtained. The model is tested on a dataset obtained from Kaggle and consists of news summaries. The model that was suggested obtained a BLEU score of 59.4% and a cosine similarity of 50.2 %; when these findings were compared with the state of the art work, it was found that the proposed model produced superior results.**

Keywords- Word Embedding, Extractive text summarization, GLOVE, Cosine similarity, BLEU Score

I. INTRODUCTION

The proliferation of electronic media and the increasing popularity of the Internet have led to a significant expansion in the quantity and lack of organization of web resources, including many formats such as text, audio, images, video, and more. Extracting relevant information from sources such as news items, books, legal files, and e-learning materials may present challenges because of the large volume of data and the absence of proper structure within the corpus. Consequently, consumers invest a significant amount of time in the process of collecting crucial information due to their inability to peruse all search results, consequently diminishing overall performance. Consequently, endeavors are undertaken to employ text filtering and text summarization techniques with the aim of minimizing content volume. The process of manual summarization is sometimes accompanied by challenges and time constraints experienced by individuals [1]. When conducting extensive text searches, the utilization of artificial intelligence (AI), machine learning (ML), and deep learning (DL) is imperative in the summarization and processing of the text [2, 3]. The issue can be addressed by the utilization of automatic text summarization (ATS). The categorization of summarization methods can be delineated into three distinct types: abstractive, extractive, and hybrid [4]. Abstractive processes involve the construction of phrases by drawing upon the core notion of a given document, followed by the semantic organization of these sentences using Natural Language Processing (NLP) techniques. Consequently, the condensed version does not incorporate the original sentences in their present structure. condensedontrary, extractive text summarization (ETS) systems aim to condense articles by finding a selection of sentences that encapsulate the primary themes [5]. This approach uses multiple phrases to provide a comprehensive outline. In the realm of text summarization, numerous scholars have introduced a range of machine learning and deep learning models that are based on artificial intelligence (AI) [6,7,8]. Further research is required to reduce lexical repetition and propose a more exact model than has been previously established. To attain these improvements, it is possible to employ many techniques such as intra-attention on decoding yield, a training approach based on reinforcement learning, pointer generation networks, and bidirectional long short-term memory (Bi-LSTM).

The project aims to replicate the use of voice data as accurately as possible by employing the BBC news dataset acquired from Kaggle [9]. Authors adopt the Bi-LSTM machine learning model to tackle the task of generating summaries. Moreover, they place emphasis on distinguishing between summarizing content at the sentence and paragraph levels. The performance of our model in the domain of sentence-level text summarization was shown to be significantly superior when tested using both techniques, as determined by researchers.

In addition, the researchers of the study have assessed the dataset by employing a range of advanced deep learning models and subsequently juxtaposed their respective findings. Scholars have utilized sophisticated methodologies, such as the comparison of ROUGE and BLEU values, to determine the comparative efficacy of various summaries. The essay concludes by analyzing potential paths for enhancing the existing procedures. This includes exploring

979-8-3503-0546-3/24/$31.00 ©2024 IEEE
DOI: 10.1109/CSNT.2024.186

the use of data indexing methods, such as the implementation of complex data structures like a wavelet tree. This research consists of six main sections. Section 2 of this paper offers a comprehensive explanation of the text summaries, essential points, and discoveries identified within the literature. Section 3 introduces the proposed methodology, which includes the presentation of an algorithm. Section 4 of this study presents a comprehensive exposition and elucidation of the data and analysis obtained from the upcoming investigation. Additionally, section 4 offers an exhaustive examination of the proposed work in accordance with the existing state-of-the-art. The inclusion of pertinent citations is presented after the discussion in Section 5, which covers the final remarks of the study and delineates the prospective directions for future investigation.

II. Literature Review

The act of generating a concise overview of a compilation of written materials is commonly referred to as text summarization. The manual summarization of content is a time-consuming and arduous task for human experts. In contemporary society, the utilization of automated text summarizing (ATS) approaches has become prevalent as a substitute for conventional methods. This shift aims to streamline and speed up summary, particularly when dealing with substantial volumes of data. Previous studies have proposed and implemented several applications in text processing by employing machine learning and deep learning techniques. These applications include tasks such as part-of-speech tagging, sentence similarity determination using deep learning, and text categorization mining [10, 11]. The researchers not only focused on text processing but also explored the utilization of machine learning and deep learning in various domains such as handwritten text analysis [12], cyberbullying detection [13], and product review analysis [14].

A comprehensive review on the topic of Text Summarization Extractive Methods was recently released by a group of experts [3]. Extractive summarizing is a method that entails the identification and consolidation of relevant phrases, paragraphs, and other elements from the source material with the aim of producing a concise version. The determination of sentence importance is accomplished by the inspection and analysis of statistical and linguistic aspects. Text summarizing techniques can be categorized as text mining tasks since they aim to generate a succinct summary or abstract from one or more input text sources [15]. A wide range of heuristic and semi-supervised learning techniques have undergone thorough investigation. The study aims to assess the effectiveness of frequently employed summarization techniques in producing extracts of varying lengths from a single source. The claim that the subject matter of the text influences the efficacy of integrating sentence scoring systems in producing a summary is supported by the results outlined in the scholarly article entitled "A Context Based Text Summarization System" [8]. This concept can be analyzed within three discrete frameworks, specifically headlines, news articles, and online content. The results support the stated hypothesis and illustrate the varying degrees of efficacy associated with different strategies in each of the examined settings.

In [12] authors have directed their attention towards the development of a method that is specific to style for the purpose of summarizing spoken news articles. The researchers evaluated the impact of various features on the summarization of news articles and broadcast data, focusing on both stylistic and content-based contributions. Their findings indicate that, in the case of news articles, the position feature plays a dominant role, while features related to content information have relatively less importance. However, for broadcast data, both stylistic and content features were found to be significant. In summary, the researchers reached the conclusion that the inclusion of the position characteristic holds significant importance in the process of summarizing news articles and broadcast data. Song et al. [6] devised a two-phased methodology for abstract text summarization utilizing deep learning techniques. The initial stage of the procedure entails identifying and extracting pertinent sentences from the primary input materials. The subsequent stage entails the generation of a concise overview with the aid of deep learning techniques. It is released a review study that focused on a notable writing audit conducted in the field of text summarization (TS), specifically in relation to human language technologies (HLT). This review study examines the latest methods for generic text summarization, as well as specialized strategies for summarizing specific purposes such as sentiment analysis, blogs, and similar platforms. The authors thoroughly analyzed these methodologies to provide a comprehensive overview. In subsequent portions of the research, the authors elaborate on the integration of text summarization with other intelligent systems rooted in Human Language Technology (HLT) [14]. Intelligent systems encompass several functionalities such as information retrieval, question answering, text classification, among others.

III. Proposed Methodology

In this paper, the extractive text summarization is performed using fusion of word embedding and LSTM. Fusion of word embeddings and LSTM based encoder decoder model for extractive text summarization involves combining the strengths of both techniques to create a more informative and context-aware summary. This fusion approach aims to enhance the summarization process by considering semantic relationships (word embeddings). The proposed model contains seven phases' names as: Data Preparation, data splitting, model building, training, inference, evaluation and hyperparameter tuning, shown in figure The performance of the model can be affected by the GloVe embeddings that are selected (for example, 50-dimensional, 100-dimensional, or 300-dimensional). It may be useful to make use of pretrained embeddings that correspond to the context and domain of the summarization task. LSTM, or long-term short-term memory, will be utilized in the construction of an encoder-decoder model. The model that is being presented is intended to take textual input to generate a summary as its final output. The

architecture of the LSTM-based model is extremely important, and this includes the number of layers, hidden units, and attention processes. Models that are more complicated have a better chance of capturing the intricate linkages that exist within the text and producing summaries of a higher quality.

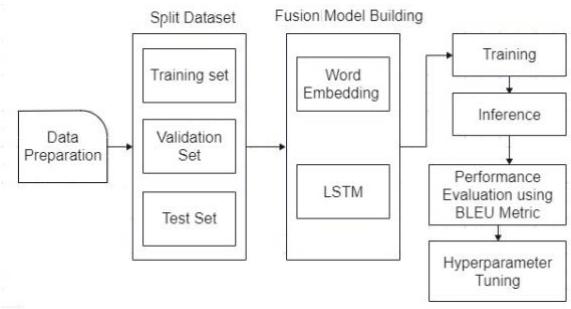

Figure 1. Architecture of Proposed Model (Fusion of GLOVE and Encoder decoder model)

IV. ALGORITHM

Text summarization using GloVe embeddings and an encoder-decoder LSTM model is a natural language processing task where you aim to generate a concise summary of a longer text. Here is a high-level algorithm for text summarization using these components.

A. *Data Preperation*

- Compile a dataset of paired text documents containing both the original text and its summary.
- Tokenize the text into words or subword tokens (using a tokenizer such as NLTK,).
- Transform words or tokens into embedded GloVe words. GloVe embeddings provide vector representations for words, and you can obtain sentence embeddings by averaging the embeddings for all the words in a sentence.

B. *Split Dataset*

- The dataset should be divided into three distinct sets, namely the training set, validation set, and test set.

C. *Build Model*

- An Encoder-Decoder model employing LSTM (Long Short-Term Memory) layers will be constructed. The proposed model is designed to accept textual input and produce a summary as its output.
- The user's text does not contain any content to rewrite. The encoder component of the model is responsible for processing the input text.
- The input sequences, whether in the form of words or embeddings, are transformed into a fixed-length representation by employing an LSTM layer. Stacking numerous LSTM layers can potentially improve performance.
- The ultimate latent state of the encoder Long Short-Term Memory (LSTM) will be employed as the beginning state of the decoder LSTM.
- The decoder component of the model is responsible for generating the summary:
- The decoder LSTM is initialized with the concealed state obtained from the encoder.
- An LSTM layer is utilized for the purpose of decoding the summary in a sequential manner, generating one word at a time. The LSTM layer's output at each time step is utilized for the purpose of predicting the subsequent word in the summary.
- To forecast the probability distribution over the vocabulary for the following word, a substantial layer with SoftMax activation is implemented.
- During the training process, it is recommended to employ a technique called trainer forcing. This involves providing the true summary phrases as inputs to the model, which facilitates its learning process.

D. *Training*

- The model should be trained using the provided training data and optimized to minimize the disparity between the predicted summary and the actual summary.
- It is recommended to employ an appropriate loss function such as categorical cross-entropy.
- It is advisable to monitor the performance of the model on the validation set in order to mitigate the risk of overfitting

E. *Inference*

To generate summaries through inference, the decoder component of the model that has been trained should be utilized.

- The process commences by initiating an initial input, typically represented by a start token. Subsequently, words are generated in a recursive manner until either an end token or a predetermined maximum length is reached.
- The utilization of beam search or greedy decoding techniques has the potential to enhance the overall quality of generated summaries.
- The word indices that have been generated can be transformed back into text by utilizing the corresponding vocabulary.

F. *Evaluation*

- The quality of the generated summaries produced by the model can be assessed by employing the BLEU metric

G. *Hyperparameter Tuning*

- Conduct experiments by varying hyperparameters such as the number of LSTM layers, LSTM units,

learning rate, batch size, and embedding dimensions to enhance the performance of the model.

- The categorical cross-entropy loss is employed to evaluate and compare the predicted summary with the ground truth summary.
- The loss received at each time step t can be expressed in eq1.

$$L(y(t),ytrue(t))=-v\sum ytrue(t)(v)\log(y(t)(v)) \quad \text{eq.1}$$

- The variable "v" iterates over the vocabulary.
- The cumulative loss for the entire sequence is obtained by summing the individual losses across all time steps, shown in eq2.

$$L(model)=t\sum L(y(t), ytrue(t)) \quad \text{eq.2}$$

Where:
y(t) - Predicted summary at time t
ytrue – Actual summary at a time t
V - Vocabulary size

V. RESULT AND DISCUSSION

The proposed approach was tested on a dataset of news summaries obtained from Kaggle and evaluated with a BLUE score, which compared the projected summary to the original summary. The complexity of the words used in news headlines as compared to the data found in news articles is illustrated in Figure 2. The number of words in headlines is typically lower and they are dispersed among several sentences, whereas the number of words in text data is typically higher and they are more cohesively tied to other content. Words in a document are shown to have cohesive correlations between themselves when they share semantic similarities with other words. The idea of similarity is what drives word embedding, which is then used for the generation of word vectors and sentence vectors, both of which can be utilized as input for an encoder decoder model. Word embedding with a dimension of 300 was accomplished with the use of GLOVE in the task that is being proposed. The total number of words in the corpus is 3665888, and there are 90013 words that are unique to the corpus. The length of the "word 2 vec" is 70834, and the number of words that are included in both the dataset corpus and the glove vectors is 70834, which is about 79.0% of the total amount of words. The dataset is then divided into an 80:20 ratio, with 20 being the size of the test dataset. The model is compiled with Adam optimizer and Sparse categorical cross-entropy' loss, both of which are presented in eq1 and eq2 and epoch 10 respectively. The loss comparison of the train data and the test data is shown in Figure 3. It can be evident that there is not a significant difference between the train loss and the test loss, which indicates that the proposed approach performed the best. The results of LSTM-based seq to seq model and the fusion of GLOVE and encoder decoder model are shown in Table 2. These results are shown in terms of BLEU score and cosine similarity score. It can be obtained from this table that GLOVE-based model gives 59.5% BLUE score and 50.2% cosine similarity score. The examples of reviews, their original summaries, and the predicted summaries that were created by the GLOVE-based encoder decoder model are presented in Table 3. It is possible to establish that the model has provided a summary that is approximately accurate.

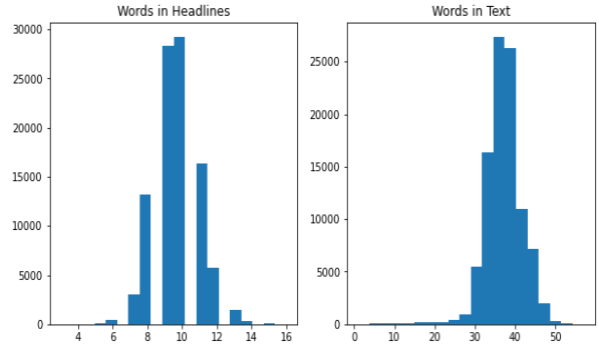

Figure 2. Cohesive similarity of Words in news headlines and text column of news summary dataset.

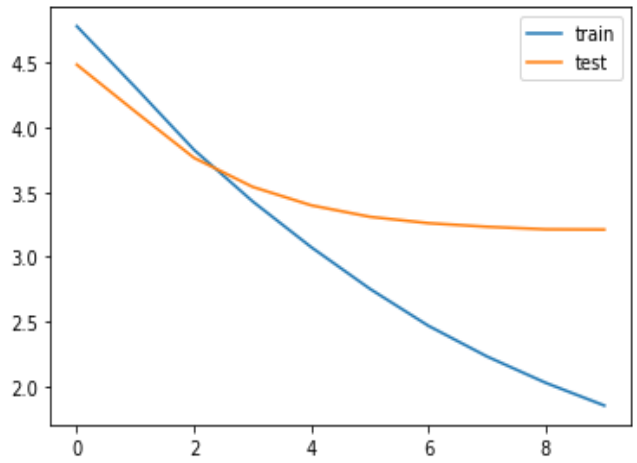

Figure 3. Loss comparison between train and test dataset according to eq1 and eq2.

TABLE 2. PERFORMANCE OF PROPOSED MODEL AND SEQ-TO-SEQ MODEL WITH BLEU AND COSINE SIMILARITY SCORE.

METHOD	BLEU Score	Cosine Similarity
Seq-to Seq Model based on LSTM	***40.78***	***39.54***
Word Embedding GLOVE + Encoder Decoder Model	***59.5***	***50.2***

TABLE 3. EXAMPLE OF REVIEW WITH ITS ORIGINAL AND PREDICTED

1.Review: streaker peace love written across chest crashed Olympic skating rink wearing pink tutu winter games Pyeongchang incident occurred right men 1 000-meter speed skating event Friday streaker used monkey pouch cover genitals later fell face forward ice

Original summary: man crashes Olympic skating rink wearing tutu

Predicted summary: runner makes Olympic skating rink cage qualifies.

2.Review: Delhi high court upheld trial court order granting anticipatory bail congress leader Sajjan Kumar alleged murder three Sikhs 1984 riots pm Indira Gandhi assassination court noted special investigation team challenging trial court order make ground bail cancellation

Original summary: HC upholds bail order cong leader accused 1984 riots.

Predicted summary: HC upholds plea seeking death tandoor 1984 riots case

VI. CONCLUSION

The purpose of the approach that has been suggested is to improve the performance of text summarization by combining word embedding with an encoder decoder model. This will be accomplished. Utilizing preprocessed data from the CNN/Daily Mail dataset is the first step in the process of creating a summary context vector. This data is obtained from the dataset. The BLEU score was utilized in order to perform the analysis that was necessary to determine how effective the summary was. It was discovered, following the completion of the pre-processing on the Sequence-to-Sequence model, that the attention mechanism accurately summarizes seventy percent of the test data. It is envisaged that in subsequent study, hyperparameter tuning will be carried out to carry out additional tests and determine the strategy that is most suitable for improving performance in text summarization. In addition, other performance measures like ROUGE, METEOR, and ATEC will be utilized to facilitate comparisons.

REFERENCES

[1] Yadav, Arun Kumar, et al. "Extractive text summarization using deep learning approach." International Journal of Information Technology, 14.5 (2022): 2407-2415.

[2] Mohsen F, Wang J, Al-Sabahi K (2020) “A hierarchical self-attentive neural extractive summarizer via reinforcement learning (hsasrl),” Applied Intelligence, pp. 1-14

[3] Joshi, Akanksha, et al. "SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders." Expert Systems with Applications 129 (2019): 200-215.

[4] Belwal, Ramesh Chandra, Sawan Rai, and Atul Gupta. "Text summarization using topic-based vector space model and semantic measure." Information Processing & Management 58.3 (2021): 102536.

[5] Nenkova, Ani, and Kathleen McKeown. "A survey of text summarization techniques." Mining text data (2012): 43-76.

[6] Belwal, Ramesh Chandra, Sawan Rai, and Atul Gupta. "A new graph-based extractive text summarization using keywords or topic modeling." Journal of Ambient Intelligence and Humanized Computing 12.10 (2021): 8975-8990.

[7] Dalal, Vipul, and Latesh Malik. "A survey of extractive and abstractive text summarization techniques." 2013 6th international conference on emerging trends in engineering and technology. IEEE, 2013.

[8] Madhuri, J. N., and R. Ganesh Kumar. "Extractive text summarization using sentence ranking." 2019 international conference on data science and communication (IconDSC). IEEE, 2019.

[9] Yadav, Arun Kumar, et al. "Extractive text summarization using deep learning approach." International Journal of Information Technology 14.5 (2022): 2407-2415.

[10] Yadav, A. K., Singh, A., Dhiman, M., Vineet, Kaundal, R., Verma, A., & Yadav, D. (2022). Extractive text summarization using deep learning approach. International Journal of Information Technology, 14(5), 2407-2415.

[11] Taner Uçkan , Ali Karcı. Extractive multi-document text summarization based on graph independent sets. Egytian informatics journal. 2020.

[12] Begum Mutlua , Ebru A. Sezer∗,b , M. Ali Akcayola. Candidate sentence selection for extractive text summarization. 2020.

[13] Abdel-Salam, S., & Rafea, A. (2022). Performance study on extractive text summarization using BERT models. Information, 13(2), 67.

[14] Cengiz Harka,∗ , Ali Karcıb. Karcı summarization: A simple and effective approach for automatic text summarization using Karcı entropy. 2020.

[15] ZHIXIN LI 1 , ZHI PENG1 , SUQIN TANG1 , CANLONG ZHANG1 , AND HUIFANG MA. Text Summarization Method Based on Double Attention Pointer Network. 2020. IEEE Access.