IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 December 2022, accepted 19 February 2023, date of publication 27 February 2023, date of current version 13 March 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3249783*

RESEARCH ARTICLE

# Neural Attention Model for Abstractive Text Summarization Using Linguistic Feature Space

**ANIQA DILAWARI[1], MUHAMMAD USMAN GHANI KHAN[2], SUMMRA SALEEM[3], ZAHOOR-UR-REHMAN[4], AND FATEMA SABEEN SHAIKH[5]**

[1]Department of Computer Science and Information Technology, University of Home Economics, Lahore 54792, Pakistan
[2]Department of Computer Science, University of Engineering and Technology, Lahore (UET Lahore), Lahore 39161, Pakistan
[3]Rheinland-Pfälzische Technische Universität, Kaiserslautern, 67663 Kaiserslautern, Germany
[4]Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock 45550, Pakistan
[5]Computer Information Systems Department, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 34212, Saudi Arabia

Corresponding author: Zahoor-Ur-Rehman (xahoor@cuiatk.edu.pk)

**ABSTRACT** Summarization generates a brief and concise summary which portrays the main idea of the source text. There are two forms of summarization: abstractive and extractive. Extractive summarization chooses important sentences from the text to form a summary whereas abstractive summarization paraphrase using advanced and nearer-to human explanation by adding novel words or phrases. For a human annotator, producing summary of a document is time consuming and expensive because it requires going through the long document and composing a short summary. An automatic feature-rich model for text summarization is proposed that can reduce the amount of labor and produce a quick summary by using both extractive and abstractive approach. A feature-rich extractor highlights the important sentences in the text and linguistic characteristics are used to enhance results. The extracted summary is then fed to an abstracter to further provide information using features such as named entity tags, part of speech tags and term weights. Furthermore, a loss function is introduced to normalize the inconsistency between word-level and sentence-level attentions. The proposed two-staged network achieved a ROUGE score of 37.76% on the benchmark CNN/DailyMail dataset, outperforming the earlier work. Human evaluation is also conducted to measure the comprehensiveness, conciseness and informativeness of the generated summary.

**INDEX TERMS** Abstractive summarization, encoder-decoder, extractive summarization, feature rich model, linguistic features, summarization evaluation.

## I. INTRODUCTION

In summarization, a compact version of textual information is generated, which typically contains the important information of the original document. There are two types of summarizations: extractive and abstractive summarization. In extractive summarization summaries are assembled exclusively from passages; it is a simpler approach because copying data from a source document ensure grammatical accuracy. On the other hand, abstraction not only signifies a summary of the mere selection of a few sentences or passages but also rephrases the main contents of a document. The task may transfer a long text sequence of words into a shorter sequence encompassing informative content. Most of the earlier work on summarization focused on extractive summarization [3], [21], [23]. In abstractive summarization [21], [26], sophisticated mechanisms have been employed to paraphrase and generate expressions unseen in the original document.

There is a plethora of real-world applications for automatic text summarization. It can assist in education, research, media monitoring, search engines, question-answering systems, social media analysis, and video scripting. For education and media monitoring, automatic summarization can

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

support us to grasp the core idea of the document. A personal assistant for question-answering systems can be improved by collecting documents that are relevant to the extractive summarized answers. Video scripting can help individuals to select desired videos based on the summarized caption of a video instead of watching the complete length. Recently a sequence-to-sequence model was used to map an input sequence into a corresponding output sequence; the approach has been successful in speech recognition [5], machine translation [1], and video captioning [27]. Similarly, an attentional encoder-decoder based neural network [1] was used for abstractive summarization. However, these models could reproduce inaccurate factual details at best and had no ability to manage OOV (out-of-vocabulary) words. To date, there have not been many studies in the existing literature that fed detailed information to a DNN (deep neural network) for abstractive summarization, hence resulting in not so high accuracies. To improve the compactness of summarized text additional information can be passed alongside actual word embeddings. In the proposed work we are incorporating multiple text features such as POS (part of speech) tags, term weights, and named entities. Although machine translation and abstractive summarization have many things in common, they are not the same task. A translation is lossless, and a strong one-to-one word-level alignment exists in machine translation between the source and the target. In abstractive summarization, the target does not depend on the length of a document and is generally short. Moreover, the original document is compressed in a lossy manner to preserve the most important contents from the original.

The major contribution of the proposed study is the use of a rich feature set for document summarization. The proposed feature set contains a sentence position, term weights, named entity tags, POS tags, and the total numbers of numerals and of proper nouns. The increasing number of features improves the comprehensiveness of a resultant summary. To this end, we propose a unified model of extractive and abstractive summarization. Firstly, we handle sentence-level attention by using extractive summarization. Secondly, by using abstractive summarization, we modulate the output at word-level attention. The approach allows extractive summarization to help abstractive summarization that mitigates forged word-level attention. The experiment uses the CNN/DailyMail dataset [21] having more than 300k news articles. We show that the approach has resulted in a ROUGE score of 37.76%, outperforming the earlier work. Human evaluation is also conducted to ensure the significance of the two-stage summarized network.

## II. LITERATURE SURVEY

Most of the recent work in text summarization relied on extractive techniques in which sentences and phrases were identified in a source document and were reproduced as a summary [6], [9], [10], [20], [30]. Several surveys exists on automatic text summarization systems using attention models, datasets, and evaluation methods to assess the quality of the summaries.

The Neural networks were used by Jean et al. [14] and Yin et al. [34] where sentences were mapped into vectors for processing. Nallapati et al. [21] and Cheng and Lapata [3] employed RNNs (recurrent neural networks) to create representations for a document. Narayan et al. [23] adopted a sentence classifier to choose sentences by utilizing additional information such as titles and image captions. Yasunaga et al. [33] combined graph convolutional networks and RNNs to compute the importance of each sentence. Although some extractive summarization models have achieved good ROUGE scores, they typically had readability problems.

Abstractive document summarization has not received enough attention prior to recent neural models. For the first time, Jing [15] created summaries by removing unimportant parts of sentences. The abstractive summarization task was standardized in DUC-2003 and 2004 competitions. TOPIARY [35] was an accomplishment on the DUC-2004 task; it used various linguistically motivated compression techniques and detection algorithms in which keywords extracted from a document were appended onto the output. Cheung and Penn [4] created sentence fusion by using dependency trees.

A modern neural network applied to abstractive text summarization was proposed by Rush et al. [25], where convolutional models were used to encode input text. To generate a summary an attentional feed-forward neural network was employed. Vinyals et al. [28] introduced a pointer network, which was a sequence-to-sequence model based on the soft attention distribution method of Bahdanau et al. [2]. The pointer network has also created hybrid approaches to language modeling, neural machine translation [11], and summarization [16], [21]. Rush et al. [25] was an extension of this work, which used the same convolutional method for the encoder, but the decoder was replaced with RNN to achieve improved performance. Hu et al. [13] used text summarization to show the auspicious performance of the Chinese dataset by employing RNN.

For extractive text summarization of the source, an RNN-based encoder-decoder was used by Cheng and Lapata [3]. A sequence-to-sequence model was used by Nallapati et al. [21] who evaluated the work using the CNN/DailyMail dataset. The traditional training matrix was replaced with an evaluation matrix (e.g., ROUGE and BLEU) by Ranzato et al. [24]. To manage OOV words See et al. [26] and Jin et al. [16] adopted pointer networks in their desired models. To mitigate repeated phrases in a summary, a different model was proposed by See et al. [26]. Yadav et al. [7] used reinforcement learning with an attention layer as the base model. Generative adversarial networks were used by Li et al. [17] to achieve a high score with human evaluation. An attention mechanism was proposed by Bahdanau et al. [1]. For document classification, a hierarchical attention mechanism was proposed by Yang et al. [32].

Nallapati et al. [21] combined word and sentence-level attention where their sentence attention was dynamic.

There are a lot of advances in auto feature engineering for developing feature selection models which include meta-learning [36], [37], [38] aka learning to learn. It focuses on how to learn and adapt even if the data is sparse.

In this research study, we propose an end-to-end model for extractive summarization followed by abstractive summarization. The encoded words are features rich preserving the linguistic information of each word. These linguistic characteristics of words are fed to the extractor and abstractor. Furthermore, the model incorporates sentence-level summarization from an extractive model and word-level summarization from an abstractive model. The concepts of different attentions have been employed by previous researchers, but attention to characteristic linguistics have not been merged for the unified model. The advantage of using words and sentence-level attention in a sequential model with feature-rich word encoding is an approach toward comprehensive summarization.

## III. APPROACH

We explore an approach that associates the strength of a state-of-the-art extractor [22] and the feature-rich abstracter [21]. This paper adopts the following notation throughout the discussion. Firstly, both an extractor and an abstracter take a sequence of $n$ words $w = \{w_1, w_2, \ldots, w_i, \ldots, w_n\}$ as input, where $i$ represents an index of a word. A sequence of words jointly forms sentences $s = \{s_1, s_2, \ldots, s_j, \ldots, s_n\}$, where $j$ is an index of a sentence. The $i^{th}$ word is associated with the $j\,(i)^{th}$ sentence, where $j$ is a mapping function. Extractive and abstractive summarization are the selection of significant sentences and words in a document. Hence sentence and word-level attention is employed in the model to generate a concise output summary. The extractor assigns sentence-level attention $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_m, \ldots\}$, where $\alpha_m$ shows the probability of $m^{th}$ sentence being extracted into a decoded summary.

In contrast, the abstracter dynamically casts word-level attention $\beta = \left\{\beta_1^t, \beta_2^t, \ldots, \beta_n^t, \ldots\right\}$ while producing $n^{th}$ word at a time step $t$ in a summary.

### A. PRE-PROCESSING

Text pre-processing involves the reduction of ambiguities caused by, e.g., the use of several forms of a certain verb, or the singular/plural form of a word. Further, stop words, such as *a, the, of, is*, do not carry much information toward our goal of summarization. Described below are multiple operations employed for preprocessing of documents.

#### 1) DOCUMENT SEGMENTATION

A text is divided into several paragraphs to find where each sentence is placed in its respective paragraph.

#### 2) STEMMING

We apply stemming to bring a word to its root or base form. The examples include the use of a singular form rather than using plural or the removal of -ing from a verb. To this end, StanfordNLP stemmer[1] is employed in this paper.

#### 3) PARAGRAPH SEGMENTATION

Paragraph segmentation divides a paragraph into sentences using sentence tags <s>.

#### 4) WORD NORMALIZATION

Each sentence consists of multiple normalized words. Through normalization and lemmatization, individual words become one common form, stemming down to their roots. Ambiguities are removed by Porter's algorithm [29].

#### 5) STOP WORD FILTERING

Stop words can be filtered out after performing other pre-processing steps. There is no uniform rule for selecting a stop word because it depends on individual tasks. In this work, words such as *a, is, in, the, of* are selected as stop words and are filtered out from the document. In text mining, applications stop word filtering is considered a standard step.

### B. FEATURE RICH EXTRACTOR MODEL FOR EXTRACTIVE SUMMARIZATION

After ambiguity removal and complexity reduction, a document is arranged into a sentence-feature matrix. Each sentence is processed to extract features and all these feature vectors are used to make a matrix. Upon trial and error of various features, we have chosen the combination of the following sentence features for an extractor model.

*Sentence Position:*

The position of a sentence in a document has been proven very effective for document compression. Most of the existing methodologies take advantage of a sentence position because they are more content representative. Typically, significant information is described at the beginning of a document. It has been observed that the first sentence is the most significant for text summarization and the effectiveness decreases as the distance from the start of a document increases.

*Number of Numerals:*

Numerals often incorporate great importance to present facts. For each sentence, we calculate the number of numerals to the total number of words given in Equation (1).

$$sen_{num} = \frac{\#numerals}{total\#words} \tag{1}$$

#### 1) MODEL ARCHITECTURE

The extractor builds on the work by Nallapati et al. [22], however, the architecture is different in that it generates highly-ranked sentences by computing informativeness based on recall scores. Recall scores are calculated using the ground truth, which indicates whether each significant sentences should be part of a summary. Using hierarchical bidirectional

[1] https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

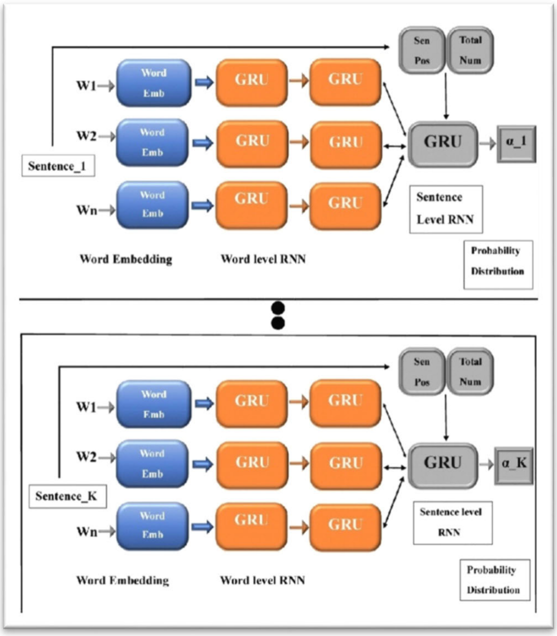


**FIGURE 1. Feature-rich extractor model for n sentences: the model architecture comprises hierarchical bidirectional LSTM (long short-term memory) which extracts sentence-level representation. It is followed by a classification layer for computing sentence-level attention $\alpha_n$.**

GRUs (gated recurrent units) [31] the network extracts sentences, assigning sentence attention $\alpha_{\text{n}}$ at the classification layer (see Figure 1). A sigmoid cross entropy function is used for computing the extractor loss $\text{Loss}_{\text{ext}}$ given by equation 2:

$$Loss_{ext} = -\frac{1}{K}\sum_{k=1}^{K}((g_k log\alpha_k) + (1 - g_k)log(1 - \alpha_k)) \quad (2)$$

where $K$ is the total number of sentences, and $g_k$ is either 0 or 1, depicting the ground truth for the $k^{th}$ sentence. Note that $g_k = 1$ narrates the $k^{th}$ sentence of a document, which is given attention to support abstractive summarization. The extractor selects highly informative sentences. It means that selected sentences should comprise relevant information required to produce abstractive summaries. To attain ground truth labels $g = g_k k$ for sentences an informativeness parameter is derived for every sentence $s_k$ of an input text by calculating the ROUGE-L score [18] of the reference summary and the sentence $\hat{z} = \hat{z}^t t$. After that, sentences are sorted in decreasing order of informativeness, and highly ranked sentences are extracted. If any upcoming sentence raises information from previously selected sentences it is added to the list of chosen sentences. The extractor is trained using the ground truth labels of sentences to minimize the loss as calculated by Equation (3). The model focuses on the recall scores of ROUGE rather than the F-1 scores to target highly informative sentences in reference to the ground truth.

### 2) WEIGHTED ATTENTIONS

The attention mechanism plays an important role in natural language processing. To deal with this challenge, a simple approach to combining the word-level $\beta_n^T$ attention with the sentence-level attention $\alpha_m$ through re-normalization and scalar multiplication has been proposed [12]. Multiplication operation is performed where calculated attention at a word level and at a sentence level is high. As the sentence-level generation has already achieved a high ROUGE score by the extractor, the word-level attention can be used to remove the contrived words from the less attended sentences. Our focus is to update the word-level attention to enhance abstractive summarization. Instead of the conclusive nature between attention at a word and a sentence level, we focus on how these two types of attention are consistent with each other during the process of training. The aim of our work is to achieve high word-level attention when sentence-level attention is high. Hence, we propose the inconsistent loss $Loss_{ics}$ defined in Equation (3).

$$Loss_{ics} = -\frac{1}{N}\sum_{T=1}^{N} log\left(\frac{1}{|W|}\sum_{n\in W}^{max} \beta_n^T \alpha_{m(n)}\right) \quad (3)$$

where $W$ represents the most attended words whereas $N$ is the total number of words present in a particular sentence. This fortifies word-level attention at a time when sentence-level attention is high. We have utilized the different loss functions for both the extractor and the abstracter to avoid a degenerated solution for the distribution of words where word-level and sentence-level attention are both high. This inconsistency loss function facilitates both the extractor and the abstracter in our proposed two-stage unified model.

## C. FEATURE RICH ABSTRACTER MODEL FOR ABSTRACTIVE SUMMARIZATION

After pre-processing, significant linguistic features are calculated. It has been observed that the following word features are most relevant for summarization.

### 1) NE (NAMED ENTITY) TAGS

NE tagging algorithms identify proper nouns in a string of text (e.g., sentence, paragraph). Sentences, having reference to named entities, such as a personal name and a company name, are of great importance to make a factual description. NE tags consist of seven classes: location, person, date, organization, money, percent, and time. In this work, we use the Stanford NE tagger.[2]

### 2) POS TAGS

Words in a text are marked and classified with their POS (Parts-of-Speech) categories such as nouns, verbs, and adverbs. There are several algorithms that are used to perform POS tagging, including statistical approaches such as a hidden Markov model. In this work, we use the Stanford POS tagger.[3]

[2] https://nlp.stanford.edu/software/CRF-NER.shtml
[3] https://stanfordnlp.github.io/CoreNLP/pos.html

### 3) NUMBER OF PROPER NOUNS

In corpus linguistics, a word is marked in a text that depends on a certain part of speech. It is not only based on its definition but also on the context, for example, the relationship of a word with its adjacent words in the paragraph. POS tagging identifies words as a noun, pronouns, verbs, adverbs, adjectives, etc. POS tagging is performed based on hidden parts of speech and discrete terms. There are rule-based and statistical approaches. Brill tagger is a rule-based algorithm and is a widely used English POS tagger. This feature is used to count the words in the sentences, which have a considerable number of proper nouns. To compute a few proper nouns, the Stanford POS tagger is used.

### 4) TERM WEIGHTS

Term weights are another important feature to deal with text summarization. The term frequency TF shows the importance of a word in the respective document. It measures how many times a word is repeated in the document, which is then divided by the length of the document for normalization which is defined in Equation (4).

$$TF(n) = \frac{\#times\ term\ n\ appear\ in\ the\ document}{total\ \#\ of\ \ terms\ in\ the\ document} \tag{4}$$

The inverse document frequency *IDF* measures the importance of a term in the document. Terms that occur rarely are often more important (hence scaled up) than frequent terms (which are weighed down) as shown in equation (5):

$$IDF\ (n) = log_e \frac{total\ \#\ documents}{\#documents\ having\ term\ n} \tag{5}$$

Finally, the term weight is calculated as TF × IDF.

### 5) MODEL ARCHITECTURE

Identification of fundamental concepts and main entities around which a story develops is a challenging task for text summarization. To this end, we incorporate linguistic features such as POS tags, NE tags, and term weights of an input document. We generate appended look-up-based vector embedding for retaining the lexical characteristics of words. POS tags replace textual descriptions with grammatical tags to encompass linguistic information. TF and IDF are continuous features that are transformed into explicit values by discretization. One hot encoding of their values represents the bin value they belong to. Consequently, in the look-up dictionary, each word is associated with word embedding and four linguistic features, i.e., POS tags, NE tags, TF and IDF. Word embedding is appended with tags and hot encodings to a single long vector. On the decoder side, only word embedding is fed as input. Figure 2 illustrates the complete architecture of a feature-rich abstracter model, which generates a comprehensive summary based on robust linguistic features.

Repetition is the major problem when generating a summary consisting of multiple sentences using sequence-to-sequence models. To address this problem, we have adopted the coverage model, in which a coverage vector $v^t$ in

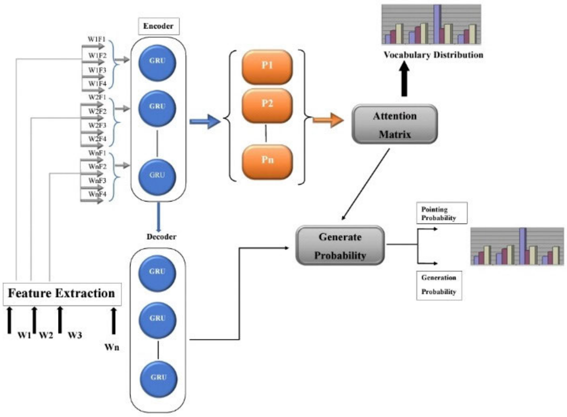


FIGURE 2. Feature rich abstracter model for n words: F1, F2, F3 and F4 represent four linguistic features, term weight, POS tag, NE tag and the number of proper nouns, that are computed for each word. They are concatenated with word embedding, then fed to the abstracter network.

Equation (6) represents the integrated effect of attention distributions computed from previous time steps.

$$v^T = \sum_{t=0}^{T-1} B^t \tag{6}$$

$v^T$ represents the division of the words from the source document and exhibits that these words have been selected using the attention mechanism. Here $v_0$ represents the vector having the zero degrees because at the first time step no source documents have been concealed.

We introduced the coverage vector into the attention mechanism presented in Equation (7).

$$E_j^T = V^t tanh(w_H H_j + w_k K_T + W_{cv} C_j^T) \tag{7}$$

where $W_{cv}$ is a trainable parameter having the same length as the vector $v$. It indicates that the attention mechanism for the current decision is endowed by the precedent decisions. This helps the attention mechanism by avoiding repeated words from the same location. Coverage loss has been utilized to penalize the process of attending the same location of the source document. The coverage loss is bounded and different from the machine translation loss. A translation ratio exists in machine translation, and the final convergence vector is penalized depending on the resultant. The employed loss function is adjustable, as constant coverage is not required in summarization. The purpose of the coverage loss function is to penalize the overlapped attention distribution and coverage to avoid reciprocal attention.

## IV. EXPERIMETAL RESULTS

### 1) DATASET

To evaluate the approach, we have used CNN/DailyMail dataset [21] that consists of online news articles (on average 781 tokens are used). These articles are also used for multi-sentence summaries (56 tokens on average or 3.75 sentences). The data set is split into the following sets: 13,368 validation pairs, 287,226 training pairs, and 11,490 test pairs.

### 2) PROCESS FOR EXPERIMENTS

Both the extractor and the abstracter are trained with 128-dimensional word embeddings. Following See et al. [26] and Nallapati et al. [22], 200 and 256 hidden states are used for the extractor and the abstracter, respectively. The vocabulary size is 50,000 words by Nallapati et al. [21], as the network is capable of handling OOV words. The pointer-generator and the coverage structure introduce a very small number of trainable parameters (1153 and 512 spare parameters) to the network. Rather than employing pre-trained embedding for word representation [21], we use embeddings learned from scratch during the network training. The learning rate for the network is 0.15 for both the extractor and the abstracter. The accumulator value is set to 0.1 using the Adagrad optimizer [8]. We apply the early stopping scheme which terminates the network training as soon as the validation data appears overfit.

During training and testing of the network, the source text length is limited to tokens of words. The maximum length for a reference summary is 100 tokens. As a comparison 120 tokens are decoded during testing. To further reduce the training time of the network, we minimize the encoding and decoding steps to 100 and 50 tokens in the early stage, thus accelerating the network training speed. Truncation of articles increases the training speed; we start training with largely truncated articles, then gradually increase the length until convergence. Moreover, we trained the model having a batch size of 4 for 48k iterations.

The linguistic features of words are concatenated with word embedding while training the vector representation of words. Thus, the features do not affect the training time of the main network. In the end-to-end model, the reduction of encoding and decoding steps further minimizes the time of training. The extractor is trained so that the length of the text is reduced, while the abstracter training aims to minimize the loss functions with $\lambda_1 = 5$ and $\lambda_2, \lambda_3, \lambda_4 = 1$. The abstracter uses extracted sentences with $g_n = 1$ as input. The combination of the extractor and the abstracter makes a two-stage network. This setting means that we use the sentence-level attention $\alpha$ as hard attention selected by the pre-trained extractor. The extractor is used as a classifier to select sentences with high attention where $\alpha_n$ is greater than a threshold.

### 3) RESULTS AND DISCUSSION

The network was trained on an 11GB GPU using a batch size of four. Training of the extractor and the abstracter required 6 days and 18 hours. It has been observed that during the early stages of training the accuracy increases exponentially. For the first 118 thousand iterations, we did not use any coverage mechanism for the abstracter.[4] The training was continued for approximately two thousand more iterations (4 hours) introducing coverage with the weighted coverage loss value of $\lambda = 1$, with the coverage loss starting from 0.9 and dropping to 0.27.[5] Finally, a beam size of four was used during testing. The network generated a non-anonymized sequence of words in the summary. The performance was measured by calculating ROUGE scores [19] between the system-generated summaries and the reference summaries. Table 1 compares F-measures for unigrams, bigrams, and the longest common subsequences between the work by Nallapati et al. [21] and the two-stage network.

**TABLE 1. Comparison of the model by Nallapati et al. [21], pointer generator [26], and the two-stage network: F-measures for unigrams (ROUGE-1), bigrams (ROUGE-2) and the longest common subsequences (ROUGE-L) are shown. We used pyrouge library (https://pypi.org/project/pyrouge/0.1.3/) when calculating these scores.**

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| Seq-seq + attention baseline (150k vocab) | 30.49 | 11.17 | 28.08 |
| Seq-seq + attention baseline (50k vocab) | 31.33 | 11.81 | 28.83 |
| Pointer Generator | 36.44 | 15.66 | 33.42 |
| Two-stage network (proposed) | **37.76** | **17.81** | **33.83** |

The table illustrates that the proposed two-stage approach with the extractor and the abstracter clearly outperformed the existing work by a clear margin. Further experiments showed that even a huge vocabulary size of 150 thousand words did not appear beneficial for efficiency raise. The performance efficient model failed to capture some significant information. For most generated summaries detailed descriptions were found redundant, and infrequent words were often replaced with words occurring more frequently. Despite a large vocabulary trick (LVT) being employed for handling OOV words, we nevertheless found a redundancy of words in the generated summaries. However, it was observed that the number of repeated words was reduced by introducing the coverage function. The redundancy problem was alleviated at the cost of 1.6% of extra training time for the coverage mechanism.

Figure 3 shows the comparison of extractive and abstractive summaries on a news article generated by the feature-rich model using the two-stage network.

Additional Information in the feature rich extractor summary includes the following:

First Read (Without reading original article):
- Saili's signature information
- Head coach Anthony statement
- Saili's team mates

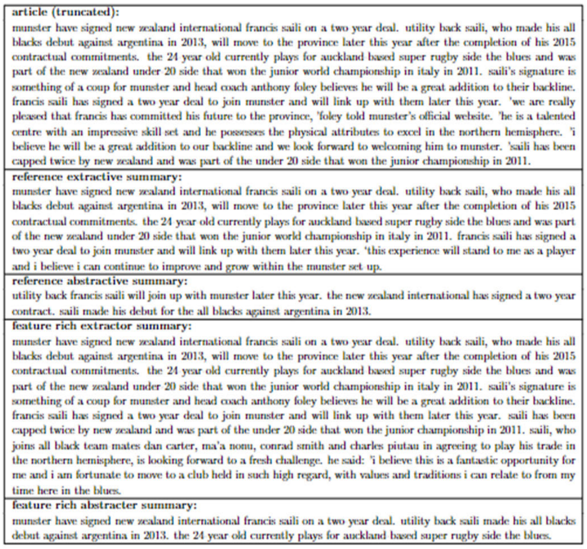
Second Read:
- Article: 2.5 min approx. (complete understanding)
- Reference Extractive Summary: 30 sec
- Feature rich extractive summary: 50 sec – 1 min approx.

First time read
- Reference abstractive summary:

Understanding of Context with respect to article: Average
- Feature Rich abstractive summary:

[4] We also tested the coverage mechanism from the first iteration of training to preserve contexts for all words. This trial, however, adversely affected the performance without sufficient reduction of redundancy.

[5] When trained the network with $\lambda = 1.5$ and $\lambda = 2$, we observed reduction in the coverage loss but increase in the basic training loss, hence we did not employ these values.



**article (truncated):**
munster have signed new zealand international francis saili on a two year deal. utility back saili, who made his all blacks debut against argentina in 2013, will move to the province later this year after the completion of his 2015 contractual commitments. the 24 year old currently plays for auckland based super rugby side the blues and was part of the new zealand under 20 side that won the junior world championship in italy in 2011. saili's signature is something of a coup for munster and head coach anthony foley believes he will be a great addition to their backline. francis saili has signed a two year deal to join munster and will link up with them later this year. 'we are really pleased that francis has committed his future to the province, 'foley told munster's official website. 'he is a talented centre with an impressive skill set and he possesses the physical attributes to excel in the northern hemisphere. 'i believe he will be a great addition to our backline and we look forward to welcoming him to munster. 'saili has been capped twice by new zealand and was part of the under 20 side that won the junior championship in 2011.

**reference extractive summary:**
munster have signed new zealand international francis saili on a two year deal. utility back saili, who made his all blacks debut against argentina in 2013, will move to the province later this year after the completion of his 2015 contractual commitments. the 24 year old currently plays for auckland based super rugby side the blues and was part of the new zealand under 20 side that won the junior world championship in italy in 2011. francis saili has signed a two year deal to join munster and will link up with them later this year. 'this experience will stand to me as a player and i believe i can continue to improve and grow within the munster set up.

**reference abstractive summary:**
utility back francis saili will join up with munster later this year. the new zealand international has signed a two year contract. saili made his debut for the all blacks against argentina in 2013.

**feature rich extractor summary:**
munster have signed new zealand international francis saili on a two year deal. utility back saili, who made his all blacks debut against argentina in 2013, will move to the province later this year after the completion of his 2015 contractual commitments. the 24 year old currently plays for auckland based super rugby side the blues and was part of the new zealand under 20 side that won the junior world championship in italy in 2011. saili's signature is something of a coup for munster and head coach anthony foley believes he will be a great addition to their backline. francis saili has signed a two year deal to join munster and will link up with them later this year. saili has been capped twice by new zealand and was part of the under 20 side that won the junior championship in 2011. saili, who joins all black team mates dan carter, ma'a nonu, conrad smith and charles piutau in agreeing to play his trade in the northern hemisphere, is looking forward to a fresh challenge. he said: 'i believe this is a fantastic opportunity for me and i am fortunate to move to a club held in such high regard, with values and traditions i can relate to from my time here in the blues.

**feature rich abstracter summary:**
munster have signed new zealand international francis saili on a two year deal. utility back saili made his all blacks debut against argentina in 2013. the 24 year old currently plays for auckland based super rugby side the blues.

**FIGURE 3.** Comparison of extractive and abstractive summaries on a news article generated by the proposed model.

**TABLE 2.** Comparison of Nallapati et al. [21], pointer generator [26], the two-stage network (proposed), and the reference summary.

| Model | Comprehensiveness | Conciseness | Informativeness |
|---|---|---|---|
| Nallapati et. al | 3.49 | 2.64 | 3.17 |
| Pointer Generator | 3.53 | 2.84 | 3.34 |
| Two-stage network (proposed) | **3.89** | **2.97** | **3.62** |

Understanding of Context with respect to the article: Good

Conclusively, the above points depict the significance of the feature-rich model as important proper nouns, numbers, and phrases are retained in generated summary. The generated summary remarkably reduced the reading time encompassing the information.

### 4) HUMAN EVALUATION

Human evaluation was conducted using Amazon Mechanical Turk.[6] We selected 50 test samples at random, with each sample consisting of an original article, the baseline, the two-stage network, and the reference summaries. Three summaries were anonymized, and their order was randomized when presented to human evaluators. Summaries were evaluated for three aspects, i.e., comprehensiveness, conciseness and informativeness.

- Comprehensiveness: a well-reported summary that is fluent and grammatically correct.
- Conciseness: presentation of summary with clear understanding without repetition.
- Informativeness: a summary that encompasses significant aspects of an article.

For the above three parameters of each summary, eight human subjects were assigned a score between 1 and 5, with 5 being the highest score.

Table 2 presents a comparison of the work by Nallapati et al. [21], the pointer generator [26], the two-stage network (proposed), and the reference summaries. The two-stage model achieved well for comprehensiveness. Most of the recent summarization techniques created a summary from the main article, while the two-stage network picked up information based on its linguistic characteristics, thus resulting in a higher comprehensiveness score. For the conciseness parameter, reference summaries scored the highest among all.

[6] https://www.mturk.com/

## V. CONCLUSION

In this paper, we presented the approach by combining the strengths of an extractor and an abstracter model to generate a comprehensive summary. The word embeddings encompassing linguistic information of words are fed to the neural network of the extractor and abstracter model. The incorporated word features include sentence position, number of numerals, POS tags, NE tags, term weights and number of proper nouns. In addition to this, attention layers highlight the most significant information for the extractor and abstracter models by using sentence and word attention parameters, respectively. The proposed approach combines attention weights for sentences and words in order to compute a loss function efficiently. The two-stage model enabled extractive and abstractive summarization in the single network. The proposed network was trained and tested using the CNN/DailyMail dataset. It was evaluated by calculating the ROUGE scores as well as by human subjects. The outcomes indicated that the approach outperformed the existing techniques with ROUGE score of 37.76%, with high comprehensiveness and informativeness.

## REFERENCES

[1] D. Bahdanau, K. Cho, and Y. Bengio, ''Neural machine translation by jointly learning to align and translate,'' 2014, *arXiv:1409.0473*.

[2] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, ''End-to-end attention-based large vocabulary speech recognition,'' in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2016, pp. 4945–4949.

[3] J. Cheng and M. Lapata, ''Neural summarization by extracting sentences and words,'' in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2016, pp. 484–494.

[4] J. C. K. Cheung and G. Penn, ''Unsupervised sentence enhancement for automatic summarization,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Oct. 2014, pp. 775–786.

[5] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, ''Attention-based models for speech recognition,'' in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, vol. 1, Jun. 2015, pp. 577–585.

[6] C. A. Colmenares, M. Litvak, A. Mantrach, and F. Silvestri, ''HEADS: Headline generation as sequence prediction using an abstract feature-rich space,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2015, pp. 133–142.

[7] A. K. Yadav, A. Singh, M. Dhiman, R. Kaundal, A. Verma, and D. Yadav, ''Extractive text summarization using deep learning approach,'' *Int. J. Inf. Technol.*, vol. 14, no. 5, pp. 2407–2415, 2022.

[8] J. Duchi, H. Elad, and Y. Singer, ''Adaptive subgradient methods for online learning and stochastic optimization,'' *J. Mach. Learn. Res.*, vol. 12, no. 7, pp. 1–39, Jul. 2011.

[9] G. Erkan and D. R. Radev, ''LexRank: Graph-based lexical centrality as salience in text summarization,'' *J. Artif. Intell. Res.*, vol. 22, pp. 457–479, Dec. 2004.

[10] K. Filippova and Y. Altun, ''Overcoming the lack of parallel data in sentence compression,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Oct. 2013, pp. 1–11.

[11] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, ''On using monolingual corpora in neural machine translation,'' 2015, *arXiv:1503.03535*.

[12] W.-T. Hsu, C.-K. Lin, M.-Y. Lee, K. Min, J. Tang, and M. Sun, "A unified model for extractive and abstractive summarization using inconsistency loss," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2018, pp. 132–141.

[13] B. Hu, Q. Chen, and F. Zhu, "LCSTS: A large scale Chinese short text summarization dataset," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1967–1972.

[14] S. Jean, K. Cho, R. Memisevic, and Y. Bengio, "On using very large target vocabulary for neural machine translation," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, vol. 1, 2015, pp. 1–10.

[15] H. Jing, "Sentence reduction for automatic text summarization," in *Proc. 6th Conf. Appl. natural Lang. Process.*, 2000, pp. 310–315.

[16] J. Jin, P. Ji, and R. Gu, "Identifying comparative customer requirements from product online reviews for competitor analysis," *Eng. Appl. Artif. Intell.*, vol. 49, pp. 61–73, Mar. 2016.

[17] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generation," 2017, *arXiv:1701.06547*.

[18] C.-Y. Lin, G. Cao, J. Gao, and J.-Y. Nie, "An information-theoretic approach to automatic evaluation of summaries," in *Proc. Main Conf. Hum. Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2006, pp. 463–470.

[19] C. Y. Lin, "Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough?" in *Proc. NTCIR*, Jun. 2004, pp. 1–10.

[20] S. Jadooki, D. Mohamad, T. Saba, A. S. Almazyad, and A. Rehman, "Fused features mining for depth-based hand gesture recognition to classify blind human communication," *Neural Comput. Appl.*, vol. 28, no. 11, pp. 3285–3294, 2017.

[21] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *Proc. 20th SIGNLL Conf. Comput. Natural Lang. Learn.*, 2016, pp. 280–290.

[22] R. Nallapati, F. Zhai, and B. Zhou, "SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 3075–3081.

[23] S. Narayan, N. Papasarantopoulos, S. B. Cohen, and M. Lapata, "Neural extractive summarization with side information," 2017, *arXiv:1704.04530*.

[24] M. A. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–16.

[25] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1–11.

[26] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 1073–1083.

[27] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence—Video to text," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4534–4542.

[28] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, vol. 2, 2015, pp. 2692–2700.

[29] P. Willett, "The Porter stemming algorithm: Then and now," *Program, Electron. Library Inf. Syst.*, vol. 40, no. 3, pp. 219–223, 2006.

[30] K.-F. Wong, M. Wu, and W. Li, "Extractive summarization using supervised and semi-supervised learning," in *Proc. 22nd Int. Conf. Comput. Linguistics COLING*, 2008, pp. 985–992.

[31] J. Amin, M. Sharif, M. Raza, T. Saba, R. Sial, and S. A. Shad, "Brain tumor detection: A long short-term memory (LSTM)-based learning model," *Neural Comput. Appl.*, vol. 32, no. 20, pp. 15965–15973, Oct. 2020.

[32] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1480–1489.

[33] M. Yasunaga, R. Zhang, K. Meelu, A. Pareek, K. Srinivasan, and D. Radev, "Graph-based neural multi-document summarization," in *Proc. 21st Conf. Comput. Natural Lang. Learn. (CoNLL)*, 2017, pp. 452–462.

[34] X. C. Yin, W. Y. Pei, J. Zhang, and H. W. Hao, "Multi-orientation scene text detection with adaptive clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1930–1937, Sep. 2015.

[35] D. Zajic, B. Dorr, and R. Schwartz, "Topiary," in *Proc. HLT-NAACL Document Understand. Workshop*, Boston, MA, USA, 2004, pp. 112–119.

[36] J. Li, B. Chiu, S. Feng, and H. Wang, "Few-shot named entity recognition via meta-learning," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 9, pp. 4245–4256, Sep. 2022.

[37] J. Li, S. Shang, and L. Chen, "Domain generalization for named entity boundary detection via metalearning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 3819–3830, Sep. 2021.

[38] J. Li, P. Han, X. Ren, J. Hu, L. Chen, and S. Shang, "Sequence labeling with meta-learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 3072–3086, Mar. 2023.

**ANIQA DILAWARI** is currently with the Department of Computer Science and Information Technology, University of Home Economics, Lahore, Pakistan. She has also worked on multiple artificial intelligence research projects. Her research interests include image processing, natural language processing, pattern recognition, and deep learning in image/video analysis.

**MUHAMMAD USMAN GHANI KHAN** was the Director of the Intelligent Criminology Laboratory, Center of Artificial Intelligence. He is currently the Head of the Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan. He is also the Director and the Founder of five research laboratories, including the Computer Vision and Machine Learning Laboratory, the Bioinformatics Laboratory, the Virtual Reality and Gaming Laboratory, the Data Science Laboratory, and the Software Systems Research Laboratory. He has more than 18 years of research experience in top journals and conference papers, especially in the areas of image processing, computer vision, bioinformatics, medical imaging, computational linguistics, and machine learning.

**SUMMRA SALEEM** is currently pursuing the Ph.D. degree with the Rheinland-Pfälzische Technische Universität, Kaiserslautern, Germany. She has six years of research experience. Her research interests include image processing, natural language processing, and pattern recognition.

**ZAHOOR-UR-REHMAN** has completed his educational and academic training at the University of Peshawar, Foundation University Islamabad, and University of Engineering and Technology, Lahore (UET Lahore), Pakistan. He joined COMSATS University Islamabad, in 2015. He has experience in both academia and research. Along with teaching responsibilities, he is an active researcher and a reviewer of various conferences and reputed journals.

**FATEMA SABEEN SHAIKH** received the Ph.D. degree in heterogeneous wireless networks from Middlesex University, London. She is currently an Assistant Professor with Imam Abdulrahman Bin Faisal University. Her specialization includes the development of intelligent, context-aware solutions to facilitate the seamless functioning of multi-homed, multi-interfaced mobile clients in converged and heterogeneous networks. She is also involved in the development of pedagogical solutions to facilitate curriculum internationalization. Her research interests include big data solutions for the development of software architectures for self-learning health systems to support personalized treatments and clinical trials. She is a fellow of the Higher Education Academy, U.K.