ROUGE-SS: A New ROUGE Variant for Evaluation of Text Summarization

Sandeep Kumar¹ and Arun Solanki¹

¹Gautam Buddha University

July 20, 2023

Abstract

Evaluation is the systematic collection and analysis of data to make judgments about a software system’s value, worth or quality. The authenticity or accuracy of any software system is based on evaluation techniques. If the evaluation technique covers more features for evaluating a software system, then it will be beneficial for determining the proposed system’s accuracy, validity, and reliability. The ROUGE score metrics and their variants are utilized for evaluating text summarization models. While ROUGE metrics are suitable for extractive approaches, they are inadequate for abstractive approaches as it rely on exact word matching. Therefore, we have proposed a new variant of the ROUGE metric called ROUGE-SS, which also considers word’s synonyms in addition to exact matches. Our experiments have shown that ROUGE-SS is more effective than other variants of ROUGE scores. The f1-score of proposed ROUGE-SS metric is increase by an average of 8.8%.

INTRODUCTION

Natural Language Processing (NLP) is a subfield of computer science and Artificial Intelligence (AI) that focuses on developing algorithms and models that enable machines to process and analyze natural language data [1]–[3]. One of the fundamental tasks in NLP is text summarization, which involves compressing a large piece of text into a shorter version while retaining the essential information. Text summarization can be approached using either extractive or abstractive methods. Extractive methods involve selecting the most important sentences or phrases from the original text and combining them to create a summary. This could expressed as an optimization problem, the objective is to maximize relevance of the selected sentences while minimizing redundancy. On the other hand, Abstractive methods involve generating new sentences that convey the original text’s meaning but may not necessarily contain the same words. This can be formulated as a sequence-to-sequence generation problem, where the goal is to generate a summary that maximizes the likelihood of the summary given the original text [4], [5].

The evaluation of text summarization can be done using various metrics, such as “ROUGE (Recall-Oriented Understudy for Gisting Evaluation)” [6] and “BLEU (BiLingual Evaluation Understudy)” [7]. ROUGE is a
recall-oriented metric that measures the overlap between the generated and reference summaries in terms of n-gram sequences. BLEU is a bilingual evaluation metric that measures the quality of machine-generated translations by comparing them to human translations. Evaluation is a crucial step in the development cycle of a machine learning model. It allows us to understand how well the model performs on the task it was designed for and provides insight into its strengths and weaknesses. Evaluating a machine learning model involves testing it on a set of data that the model has not seen before. This testing data is often referred to as the validation or test set. By evaluating the model on a testing dataset, we can determine how well the model is able to generalize to new and unseen data. Without effective evaluations, comparing different models or determining which model is the best for a particular task is difficult. Evaluations also help researchers to identify areas where improvements can be made and to guide the development of new machine learning algorithms [8], [9]. Overall, evaluations are essential for ensuring that machine learning models are effective and reliable, and it plays a crucial role in advancing the field of machine learning.

Text summarization is a critical work in natural language processing that has received more attention in recent years as digital data has grown exponentially. Considering the emergence of big data, text summarization has become crucial in information retrieval, content management, and knowledge representation [10], [11]. Evaluating the quality of text summarization is a significant challenge as there are no well-defined objective measures for it. This paper discussed the evaluation of text summarization in detail, including the points and techniques used.

Figure 1 shows the Evaluation Techniques as, the Human, Automatic, and Hybrid evaluation are used to evaluate the quality of text summarization [12]. Figure 2 shows the type of automatic evaluation techniques. It is automatically use algorithms to evaluate the summary quality. Some popular automated evaluation techniques are ROUGE, BLEU, and METEOR. These techniques use statistical measures to compare the summary with the original document and determine its quality and result in terms of accuracy.

Motivation for the proposed work

For abstractive text summarization, the ROUGE evaluation metric are used by the past researchers. The existing variants of ROUGE evaluation metric does not take into account semantic meaning. It only considers the exact words or phrases used in the text, not the underlying meaning. This means that ROUGE will not recognize the similarity, if two sentences have the same purpose but use different words to convey the same idea. In this paper, we have delved into the details of ROUGE score metrics and their variants, which are widely used for evaluating text summarization models. ROUGE metrics, which, rely on precision, recall, and F1 scores [13], are typically well-suited for extractive text summarization approaches, which involve selecting sentences or phrases from the original text. However, ROUGE may not be as effective for abstractive approaches, which generate novel summaries by paraphrasing and rephrasing the original content. This is because ROUGE metrics are based on exact word matching, which can be challenging for abstractive models that produce summaries that may not contain the same words as the source text. To overcome this constraint, we have proposed a new variant of the ROUGE metric called ROUGE-SS, which considers exact word matching as well as synonyms of words. Specifically, ROUGE-SS uses WordNet online dictionary [14], a lexical database [15] that groups words into synonyms called synsets, to identify matches between the source text and the summary beyond exact word overlap. Our experiments demonstrate that ROUGE-SS outperforms other ROUGE variants. It achieves higher F1-score and better performance for abstractive summarization models [16]. By incorporating synonyms and expanding the scope of matching beyond exact words, ROUGE-SS provides a more nuanced, comprehensive evaluation metric, making it a more effective evaluation metric for abstractive models.
The paper’s organization is represented in Figure 3. The section 1 presents the Introduction. Section 2 presents the Literature review. Section 3 presents existing evaluation metrics, and challenges. Section 4 describes the proposed approach for the ROUGE-SS evaluation metric, with a diagram illustrating the process. Section 5 includes the Implementation, Results and Analysis of the ROUGE-SS metric and compares it with existing evaluation metrics. Section 6 compares the proposed evaluation technique with the existing ROUGE variant as ROUGE-1/2/L on various text summarizer models. Finally, Section 7 concludes the article with future directions in this evaluation technique.

LITERATURE REVIEW

Abstractive text summarization is a challenging task that involves generating a summary that captures the main ideas of the original text concisely and coherently. Evaluating abstractive text summarization models is crucial to assess their effectiveness and compare different approaches. However, evaluating the quality of abstractive summaries is difficult since there is no clear objective criterion for measuring their quality. One of the most widely used evaluation metrics for text summarization is the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score, which measures the overlap between the text and its summary regarding n-gram co-occurrence and word order [6]. However, the ROUGE metrics are based on exact word matching. The semantic similarity between the original text and its summary could not be captured, which is a major limitation for evaluating abstractive summarization models. Several studies have proposed variants of the ROUGE metric that incorporate semantic information to address this limitation. For example, using WordNet, ROUGE-W [17] measures the word-level similarity between the text and its summary. This lexical database contains information about synonymy, hyponymy, and hypernymy relationships between words. ROUGE-S [18] measures the similarity between the text and its summary based on their sentence-level semantic content, as captured by “latent semantic analysis (LSA) or latent Dirichlet allocation (LDA)” models. Other evaluation metrics that have been proposed for abstractive text summarization include the Content-Units Metric (CU-Metric) [4], which measures the number of content units (i.e., concepts or propositions) that are covered by the summary. The Pyramid method [5] evaluates the informativeness and coherence of the summary using a pyramid structure that represents the hierarchy.

Nallapati et al. (2016) [19] proposed a sequence-to-sequence RNN [20] and beyond model with an attention mechanism for abstractive text summarization. They evaluated their model using the ROUGE score and showed that it outperformed a number of other state-of-art models. Chopra et al. (2016) [21] proposed a neural attention-based model for abstractive text summarization. They evaluated their model using the ROUGE score and showed that it outperformed several other baselines, including the extractive summarization model. Paulus et al. (2018) [22] A deep reinforced model for abstractive text summarization was presented. Which incorporated the global context of the document. They evaluated their model using the ROUGE score and showed that it outperformed several other state-of-art models. Liu et al. (2018) [23] proposed a hierarchical encoder-decoder model for abstractive text summarization that considers the document’s hierarchical structure. They evaluated their model using the ROUGE score and showed that it outperformed several other state-of-art models. Zhang et al. (2019) [24] proposed a semantic-aware model for abstractive text summarization that takes into account the semantic relationship between words. They evaluated their model using the ROUGE score and showed that it outperformed several other state-of-art models. Chen and Bansal (2018) [25] proposed a transformer-based model for abstractive text summarization that utilizes the self-attention mechanism. They evaluated their model using the ROUGE score and showed that it outperformed several other state-of-art models. Cao et al. (2020) [26] proposed a model for abstractive text summarization that used a pre-trained language model to generate summaries. They evaluated their model using the ROUGE score and showed that it outperformed several other advanced methods.

Huang et al. (2021) [27] proposed a hierarchical model for summarizing abstractive text using a graph neural network to capture the semantic relationship between sentences. They evaluated their model using the ROUGE score and showed that it outperformed several other state-of-art models. Xiong et al. (2021) [28] planned a transformer-based summarization model that utilizes a new cross-task learning method. They evaluated their model using the ROUGE score and showed that it outperformed several other models. The
authors suggested that their proposed cross-task learning methods can improve the model’s generalization ability. Liu et al. (2021) [29] proposed a conditional BERT-based model for abstractive text summarization that utilizes a conditional encoder-decoder architecture. They evaluated their model using the ROUGE score and showed that it gave outperform results. The authors recommended that their proposed conditional encoder-decoder architecture can more effectively capture the semantic relationship between the input text and its summary. Jha et al. (2021) [30] proposed a transformer-based model that utilizes a new attention mechanism for abstractive text summarization. They evaluated their model using the ROUGE score and showed that it outperformed several other models. The authors suggested that their proposed attention mechanism can improve the model’s ability to attend to important parts of the input text. Tao et al. (2021) [31] proposed a transformer-based model for abstractive text summarization that utilizes a new self-distillation method. They evaluated their model using the ROUGE score. The authors suggested that their proposed self-distillation method can improve the model’s ability to capture the important features of the input text.

Ding et al. (2021) [32] proposed a hybrid attention-based abstractive text summarization model that combines the strengths of both extractive and abstractive approaches. They also evaluated their model using the ROUGE score. According to the authors, their proposed hybrid technique can increase the model’s overall performance. Zheng, C. et al. (2021) [33] The proposed topic-aware abstractive summarization (TAAS) framework addresses a gap in the current state-of-art models by incorporating global semantics and latent topics into the summarization process. It builds upon the Transformer-based encoder-decoder architecture commonly used in abstractive summarization. The authors highlight that while existing models capture contextual information well, they overlook the importance of incorporating global semantics. Experiments on real-world datasets show that TAAS beats BART, a recognized cutting-edge model, regarding F1 scores of ROUGE-1, ROUGE-2, and ROUGE-L. Additionally, TAAS achieves comparable performance to PEGASUS and ProphetNet, despite being trained with less computational resources. The results validate the effectiveness of incorporating topic modeling and global semantics in abstractive summarization. Ng, J.P. et al. (2021) [34] The paper addresses the limitations of the widely used ROUGE evaluation measure for text summarization, which is biased towards surface lexical similarities and lacks effectiveness in assessing abstractive summarization and paraphrased summaries. The project investigates the use of word embedding for calculating the semantic equivalents of terms in summary, with the goal of overcoming ROUGE’s inadequacies. The experimental findings show that the suggested method obtains better correlations with human judgements, as evaluated by “the Spearman and Kendall rank coefficients”. This research contributes to enhancing the evaluation of text summarization techniques by leveraging word embedding for a more comprehensive assessment of semantic similarities.

Wang, G. et al. (2022) [35] The paper addresses the limitations of existing text summarization models, including unfaithfulness and factual errors, by proposing a novel approach called Knowledge-aware Abstractive Text Summarization (KATSum). The model leverages the advantages offered by Knowledge Graphs to enhance the standard Seq2Seq model and extract relevant keywords with relational information. Experiment findings on data from the real world show that the suggested structure works well. utilizes Knowledge Graph information and significantly reduces factual errors in the summaries. The integration of Knowledge Graphs allows for more coherent and accurate summarization. This research highlights the importance of incorporating external knowledge sources in abstractive text summarization to improve the quality of generated summaries. Kauris, P. et al. (2022) [36] The paper presents a novel framework for text summarization that combines semantic graph representations and deep learning predictions. The framework leverages abstract meaning representation models and explores various graph construction and transformation methods. Multiple deep learning architectures are examined, and the problem is formulated as a graph-to-summary learning task. The paper introduces an extended metric for evaluating factual consistency in generated summaries. The experimental evaluation showcases the effectiveness of the proposed approach. This research contributes to the field by providing a comprehensive and innovative approach to abstractive text summarization using semantic graphs and deep learning techniques. Marcello, B. et al. (2022) [37] The paper focuses on evaluating the quality of text summarization techniques using the widely used ROUGE metric. It discusses the
distinction between extractive and abstractive summarization approaches and the importance of assessing the similarity between summaries and the original text. The study aims to accurately estimate the behavior of the ROUGE metric by conducting experiments comparing its effectiveness for evaluating abstractive and extractive summarization algorithms. Additionally, the paper compares the performance of a single summarization algorithm execution with multiple executions of different algorithms on the same text. The results indicate that the ROUGE metric performs similarly for both abstractive and extractive algorithms and multiple executions generally yield better results than a single execution. This research contributes to the understanding of the evaluation process for text summarization techniques and the utilization of the ROUGE metric.

Liu, Y. et al. (2022) [38] The paper addresses the limitations of reference-based evaluation for text summarization and the authors provide a new reference-free assessment measure. Using pre-trained language models, the measure compares the semantic distribution of the original document with the summary and integrates the summary compression ratio. In terms of coherence, consistency, relevance, and fluency, experimental results show that the suggested metric is more closely aligned with human judgement. This research contributes to the field by providing a flexible and effective approach to evaluate text summarization without the need for reference summaries. The combination of semantic correlation analysis and compression ratio assessment offers a comprehensive evaluation framework for summarization techniques. Manik Bhandari et al. (2020) [39] attempted to re-evaluate the text summarization assessment technique. The authors used extractive and abstractive text summarization models for re-evaluation with the help of top-scoring system output. They made a public dataset of human judgements gathered from 25, top-scoring neural summarizing systems (fourteen abstractive & Eleven extractives).

Dima et al. (2020) [40] conducted a study on deep learning models [41] for abstractive text summarization. They analyzed the datasets, evaluation metrics, and challenges associated with summarization. The researchers discovered that recurrent neural networks contain attention mechanisms. The LSTM were the most commonly used techniques [42], and pre-trained encoder models achieved the highest ROUGE scores. However, ongoing issues include OOV terms, repeated summaries, incorrect phrases, and false facts. The study by J. Ding et al. (2020) [43] highlights the importance of factual accuracy in abstract summarization, which has been overlooked in previous studies that relied on human evaluation. The authors examined the factors that influence factual correctness by studying humans and found that confirmed error rate and assessment methods depend on the summarization models and datasets used. Moving forward, the authors plan to retrain FactCCX on their generated datasets and evaluate top pre-trained summarization models when they become available.

EXISTING ROUGE METRICS

Past researchers have used the ROUGE score as an evaluation metric for text summarization [6]. This determines the degree of similarity between a candidate document and a set of reference documents. The ROUGE score assesses the effectiveness of document translation and summarization methods. There are mainly five variants of the ROUGE evaluation metrics as ROUGE_N, ROUGE_L, ROUGE_W, ROUGE_S, and ROUGE_SU.

ROUGE’s variants are:

- **ROUGE_N**: This evaluates n-gram overlaps between the system and reference summaries.
- **ROUGE_L**: This evaluates the longest common subsequence between the system and reference summaries.
- **ROUGE_W**: This evaluates weighted n-gram overlaps between the system and reference summaries.
- **ROUGE_S**: This measures skip-bigram co-occurrence statistics between the system and reference summaries.
- **ROUGE_SU**: This modified version of ROUGE-S uses a stemmer before measuring skip-bigram co-occurrence statistics. Other well-liked measures for assessing the summary intrinsically include precession, recall, and f1-measure [13].

3.1. Mathematical models

The mathematical formulas behind the ROUGE variant are defined here with equation numbers (1), (2),
Here, $S_{\text{Ref}}$ is the reference summary and $S_{\text{Gen}}$ is the generated summary, $\text{gram}_n$ is the $n$-gram which can be uni-gram (1-gram) and bi-gram (2-gram), LCS is the Longest Common Subsequence, $N_r$ is the number of words in the reference summary, $N_g$ is the number of words in the generated summary, WLCS is the Weighted-LCS, SKIP2 is the number of skip-bigram matches between $S_{\text{Ref}}$ and $S_{\text{Gen}}$, SKIP is the number of skip-unigram matches between $S_{\text{Ref}}$ and $S_{\text{Gen}}$. 

3.2. Limitations in ROUGE metrics

Due to the growing amount of digital information available recently, text summarization has been a focus of ongoing research. The evaluation of summarization models is crucial to assess their effectiveness and to compare different approaches. One of the most widely used evaluation metrics is the ROUGE score, which measures the overlap between the text and its summary regarding $n$-gram co-occurrence and word order. However, the ROUGE metrics have limitations regarding abstractive text summarization approaches, which generate summaries by paraphrasing or rephrasing the original text. Abstractive summarization models often produce summaries that differ significantly from the original text, making exact word matching impossible.

In contrast, ROUGE measures are based on exact word matching and cannot detect semantic similarities between the original text and its summary. Several studies have proposed variants of the ROUGE metric that incorporate semantic information to address this limitation. For example, using WordNet, ROUGE-W [17] measures the word-level similarity between the text and its summary. This lexical database contains information about synonymy, hyponymy, and hypernymy relationships between words. ROUGE-W is limited to WordNet and may not capture all synonyms of words. At the same time, ROUGE-S requires complex topic models and may not accurately capture the semantic similarity between the text and its summary.

To address these limitations, this paper proposed a new variant of the ROUGE metric called ROUGE-SS ($\text{SynonymS}$), which considers both exact word matching and synonyms of words in the text and its summary. Our experiments on several benchmark datasets have shown that ROUGE-SS outperforms other ROUGE variants in capturing the semantic similarity between the text and its summary.

PROPOSED MODEL

This section describes the proposed model for evaluating text summaries, ROUGE-SS. This evaluation technique extends the ROUGE-1 metric, focusing on unigrams or single words. The procedure for our proposed model is illustrated in Figure 4.
Fig. 4: ROUGE-SS: The proposed metric for text summary evaluation

The proposed model for the ROUGE-SS technique utilizes the following steps.

**Step 1: Dataset Collection**
Collect a labelled dataset consisting of text and corresponding summary pairs.

*Let D be a labelled dataset of pair (x, y), where x is a text, and y is its corresponding summary.*

**Step 2: Text Summarization**
Pass the labelled dataset to a text summarizer model to generate summaries for each text.

*Let M be a text summarization model, that takes a text x as input and generates a summary y’ = M(x). For each pair (x, y) in D, apply M to generate a summary y’ for x.*

**Step 3: Text Preprocessing**
In the text preprocessing stage, pass the reference summary through the Syn_Dictionary tool that uses synsets from Wordnet Dictionary to find synonyms for each word/token in the reference summary.

Store the words in a set data structure to remove duplicates.
Apply text preprocessing techniques, such as lowercasing, stemming, and lemmatization, to both the generated and reference summaries.

Tokenize the text to create a list of unigrams for both the generated and reference summaries.

Let \( P \) be a text preprocessing function that inputs a text \( x \) and outputs a preprocessed text \( P(x) \). Let \( \text{Syn\_Dictionary} \) be a function that takes a word \( w \) as input and returns a set of synonyms \( S(w) \) obtained from WordNet Dictionary. For each summary \( y \) in \( D \) and its corresponding generated summary \( y' \), do the following:
- Pass \( y \) through \( \text{Syn\_Dictionary} \) to obtain a set of synonyms \( SY(y) \). Store the words in \( SY(y) \) in a set data structure to remove duplicates.
- Apply text preprocessing techniques, such as lowercase, stemming, and lemmatization, to both \( y \) and \( y' \).
- Tokenize \( y \) and \( y' \) to create a list of unigrams \( Y \) and \( Y' \), respectively.

**Step 4: ROUGE-SS Metric Calculation**

Pass the preprocessed generated summary and merged synonyms reference summary, as input parameters to the ROUGE-SS metric function.

Measure the Precision, Recall, and F1-measure of ROUGE-SS for the input parameters.

Let \( \text{ROUGE-SS} \) be a metric that calculates the similarity between two sets of unigrams. For each pair \((y, y')\) in \( D \) and their corresponding preprocessed unigram sets \( Y \) and \( Y' \), do the following:
- Calculate the ROUGE-SS scores between \( Y \) and \( Y' \), considering the synonyms in \( SY(y') \).
- Measure the Precision, Recall, and F1-score for the generated summary \( y' \) against the reference summary \( y \) using the ROUGE-SS scores.

**IMPLEMENTATION, RESULTS AND ANALYSIS**

This section describes the implementations of the ROUGE-SS evaluation metric, which utilizes synonym features of the reference summary. The pseudo procedure for this technique is also explained, along with the steps to evaluate a generated summary.

**Implementation**

The proposed model is implemented in Python using the NLTK, ROUGE, and WordNet libraries. The pseudocode for the proposed model is as follows:

Step 1: Import the necessary Python libraries, including nltk, wordnet, and omw-1.4.
Step 2: Generate a summary in string format that you want to evaluate against one or more reference summaries.
Step 3: Import the rouge library and calculate the ROUGE scores of the generated summary using this metric.
Step 4: For each word in the reference summary, create a set data structure and populate it with synonyms using WordNet.
Step 5: Update the reference summary by replacing each word with its corresponding set of synonyms.
Step 6: Create a function for ROUGE-SS that takes two parameters: the updated reference summary with synonyms and the generated summary.
Step 7: Tokenize the text, perform stemming to reduce words to their base form, and calculate the ROUGE-SS score using the updated reference summary.
Step 8: Call the ROUGE-SS function with the updated reference summary and the generated summary as parameters.
Step 9: Print the ROUGE-SS score in precision, recall, and F1 format.

The algorithm of ROUGE-SS evaluates a generated summary based on its similarity with the reference summary. The process includes using synonyms to match terms in the produced summary to those in the reference summary. The Algorithm for the synonyms features (ROUGE-SS) is as follows:

```plaintext
// Ref = reference summary, //Gen = Generated summary
1. Tokenize Ref & Gen
2. SRef = r1, r2, r3, .........rn ⋯ Ref
3. SGen = g1, g2, g3, .........gn ⋯ Gen
4. Count number of words in SRef & SGen
5. Nr = math.count(SRef)
6. Ng = math.count(SGen)
```
7. Initialize Set $L = \text{NULL}$
8. For each $r$ in $S_{Ref}$:
9. \hspace{1cm} $sys = \text{find}_\text{synonyms}(r)$
10. If $sys$ is not empty:
11. \hspace{1.5cm} $L = L + \text{syn}$
12. \hspace{1.5cm} $L = L + S_{ref}$ \hspace{0.2cm} //Merge $S_{Ref}$ $\&$ $L$
13. \hspace{1cm} Initialize count $= 0$
14. \hspace{1.5cm} $count = \text{length}(S_{gen} \text{ or } L)$
15. \hspace{1.5cm} Find recall, precision, F1-score
16. \hspace{1.5cm} $\text{Recall}_{SS} = \frac{count}{N_r}$
17. \hspace{1.5cm} $\text{Precision}_{SS} = \frac{count}{N_g}$
18. $f1\text{-measure} = 2 \times \frac{\text{Recall}_{SS} \times \text{Precision}_{SS}}{\text{Recall}_{SS} + \text{Precision}_{SS}}$

This algorithm is a method for computing the recall, precision, and F1-score of a text summarization system using a synonym set. The algorithm starts by tokenizing the reference and generated summary, then counting the number of words in each summary. It initializes an empty set and finds synonyms for each word in the reference summary. It merges the reference summary with the synonym set to create a new set $L$. The algorithm then counts the number of words in the generated summary that appear in $L$ and computes the recall, precision, and F1-score based on this count.

This method is useful for evaluating abstractive text summarization models that may use different words to express the same idea. This process evaluates the similarity between a generated and a reference summary using the ROUGE-SS metric with synonym features.

Results & Analysis

In this subsection, we presented the results and analysis of the ROUGE-SS metric, comparing the performance of different text summarizer models using reference and generated summaries. We also compare these results to those of existing ROUGE scores.

\[
\begin{align*}
('\text{rouge-1}': \{'r': 0.4, 'p': 0.6666666666666666, 'f': 0.49999995312500066, 's': 0.49999995312500066},
('\text{rouge-2}': \{'r': 0.2222222222222222, 'p': 0.4, 'f': 0.2857142857142857, 's': 0.2857142857142857})),
('\text{rouge-1}': \{'r': 0.4, 'p': 0.6666666666666666, 'f': 0.49999995312500066})
\end{align*}
\]

Fig. 5: Testing example for different variant of ROUGE
Fig. 6: The evaluation of the ROUGE metrics
Fig. 7: The evaluation of the ROUGE-SS metric

In Figure 5, we obtained a reference summary and generated a summary according to the ROUGE-SS pseudo-code. Figure 6, presented the evaluation of the existing ROUGE score’s Precision, Recall, and F1 of “the ROUGE-1, ROUGE-2, and ROUGE-L”. The Figure 7 presented ROUGE-SS metrics on these parameters, including both the reference and generated summary. The comparison of these metrics are shown in Table 1 and a graphical representation of these results are presented in Graph 1.

Table 1: The comparison of ROUGE-SS with existing ROUGE on T2SAM summarizer model with obtained reference and generated summary [46]
Table 1 and Graph 1 show the performance comparison of evaluation metrics. The ROUGE-SS metric outperforms the other variants of ROUGE metrics. The F1-measures of the existing ROUGE-1, ROUGE-2, ROUGE-L and ROUGE-SS is 49%, 28%, 49% and 62% respectively. The F1-score of ROUGE-SS is 62%, which is increased by 13% of top baseline model.

Different Case study of ROUGE-SS metric evaluation

Table 2, 3, 4, 5, 6, 7, 8, 9, and 10 show that the ROUGE-SS scores outperform the results of the existing variants of ROUGE scores. These tables compared the performance of different summarizer models based on their reference and generated summaries.

### Table 2: The comparison of ROUGE-SS with existing ROUGE on Sy-Bi+multi-head atten summarizer model [44]

<table>
<thead>
<tr>
<th>Summarizer Model</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sy-Bi+multi-head atten</td>
<td>35.48</td>
<td>40</td>
<td>49</td>
<td>62</td>
<td>66</td>
<td>49</td>
<td>66</td>
<td>49</td>
</tr>
</tbody>
</table>

### Table 3: The comparison of ROUGE-SS with existing ROUGE on Multi-head attention summarizer model [44]

<table>
<thead>
<tr>
<th>Summarizer Model</th>
<th>ROUGE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sy-Bi+multi-head atten</td>
<td>35.48</td>
</tr>
</tbody>
</table>

Reference summary: There are many ways to improve memory, such as developing good eating and sleeping habits, and enough rest in the brain can improve memory.

Generated summary: To improve memory, you must develop good habits, pay attention to diet and sleep, and then use some methods to strengthen brain memory.
Table 4: The comparison of ROUGE-SS with existing ROUGE on En-semantic-model+pos-w2ePro+dimen summarizer model [44]

<table>
<thead>
<tr>
<th>Reference summary:</th>
<th>Generated summary:</th>
<th>Summarizer Model</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>There are many ways to improve memory, such as developing good eating and sleeping habits, and</td>
<td>There are some method to improve memory by developing good eating and sleeping habits, and</td>
<td>En-semantic-model+pos-w2ePro+dimen</td>
<td>62.50</td>
</tr>
</tbody>
</table>

Table 5: The comparison of ROUGE-SS with existing ROUGE on Seq2ASeq + Attention summarizer model [45]

<table>
<thead>
<tr>
<th>Reference summary:</th>
<th>Generated summary:</th>
<th>Summarizer Model</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>british no 1 defeated dominic thiem in miami open quarter finals. andy murray celebrated his</td>
<td>andy murray beat UNK bedene 6-3, 6-4, 6-1 in an hour and three quarters. british no 1 believes</td>
<td>Seq2ASeq + Attention model</td>
<td>27.02</td>
</tr>
</tbody>
</table>

Table 6: The comparison of ROUGE-SS with existing ROUGE on Pointer-Generator without Coverage summarizer model [45]

<table>
<thead>
<tr>
<th>Reference summary:</th>
<th>Generated summary:</th>
<th>Summarizer Model</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>british no 1 defeated dominic thiem in miami open quarter finals. andy murray celebrated his</td>
<td>andy murray came close to giving himself some extra preparation for his wedding next week</td>
<td>Pointer-Generator without Coverage</td>
<td>23.68</td>
</tr>
</tbody>
</table>

Table 7: The comparison of ROUGE-SS with existing ROUGE on Pointer-Generator, with Coverage summarizer model [45]

<table>
<thead>
<tr>
<th>Reference summary:</th>
<th>Generated summary:</th>
<th>Summarizer Model</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>british no 1 defeated dominic thiem in miami open quarter finals. andy murray celebrated his</td>
<td>murray was awaiting the winner from the last eight match between tomas berdych and a</td>
<td>Pointer-Generator, with Coverage</td>
<td>25.71</td>
</tr>
</tbody>
</table>

Table 8: The comparison of ROUGE-SS with existing ROUGE on ABS summarizer model [19]
Table 9: The comparison of ROUGE-SS with existing ROUGE on ABS+ summarizer model [19]

<table>
<thead>
<tr>
<th>Reference summary:</th>
<th>Reference summary:</th>
<th>Generated summary:</th>
<th>Generated summary:</th>
<th>Summarizer Model</th>
<th>Precision</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>tennis : davydenko pulls out of sydney with injury</td>
<td>tennis : davydenko pulls out of sydney with injury</td>
<td>davydenko pulls out of sydney international with foot injury</td>
<td>davydenko pulls out of sydney international with foot injury</td>
<td>ABS</td>
<td>77.77</td>
<td>77.77</td>
<td>77.77</td>
<td>77.77</td>
</tr>
</tbody>
</table>

Table 10: The comparison of ROUGE-SS with existing ROUGE on words-lvt5k-1sent summarizer model [19]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>norway protests russia barring fisheries research ships</td>
<td>norway grants diplomatic protest to Russia</td>
<td>norway grants diplomatic protest to Russia</td>
<td>norway grants diplomatic protest to Russia</td>
<td>words-lvt5k-1sent</td>
<td>16.66</td>
</tr>
</tbody>
</table>

The case study of the above mentioned tables shows that the proposed text summarizer evaluation technique ROUGE-SS performs better than other variants of ROUGE metrics, such as ROUGE-1, ROUGE-2, and ROUGE-L. These tables show the Precision, Recall, and F1 of different summarizer models based on the text’s reference and generated summary. Tables 6 and 7 indicate that when the reference summary does not contain synonyms, the F1 score of ROUGE-SS is similar to that of ROUGE-1.

COMPARISON

In this section, we equate the F1 score of ROUGE-SS score to the existing variants of ROUGE metrics with different datasets and text summarizer models.

Table 11: The comparison of evaluation techniques on different text summarizer models

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Model</th>
<th>Dataset</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>Sy-Bi+multi-head atten [44]</td>
<td>SOGOU News</td>
<td>39.13</td>
<td>11.32</td>
<td>32.72</td>
</tr>
<tr>
<td>3.</td>
<td>Multi-head attention [44]</td>
<td>SOGOU News</td>
<td>38.09</td>
<td>04.44</td>
<td>32.55</td>
</tr>
<tr>
<td>4.</td>
<td>En-semantic-model+pos-w2ePro+dimen [44]</td>
<td>SOGOU News</td>
<td>65.21</td>
<td>32.65</td>
<td>56.52</td>
</tr>
<tr>
<td>7.</td>
<td>Pointer-Generator, with Coverage [45]</td>
<td>CNN/Daily Mail</td>
<td>32.55</td>
<td>05.19</td>
<td>25.71</td>
</tr>
<tr>
<td>8.</td>
<td>ABS [19]</td>
<td>DUC Corpus</td>
<td>77.77</td>
<td>50.00</td>
<td>77.77</td>
</tr>
</tbody>
</table>
Table 11 displays the F1 measures of the ROUGE-SS metric. The table compares this metric with other variants of ROUGE metrics. The comparison is made on different text summarizer models.

The table shows that the proposed metric, ROUGE-SS, performs better than other ROUGE metrics on different datasets. On average, the F1 score increases by 8.8% when using ROUGE-SS compared to the base evaluation metric, ROUGE-1. The F1 score increases, which means the performance of the summarizer model will also improve.

CONCLUSION AND FUTURE DIRECTION

The proposed ROUGE-SS technique is an evaluation metric of the text summarization model. There are many evaluation metrics of text summarization, like METEOR, BLEU, and ROUGE. ROUGE are available in different variants. These are discussed in the Introduction section. We studied various research papers on this domain and found the research gaps in the text summary evaluation metric. We focused on the most popular metric, ROUGE-1, which gave better results than other ROUGE variants like ROUGE-2, ROUGE-L (in terms of F1 measures).

But the problem/gap is: It worked on the exact matching of the word from the reference summary and generated summary. So this evaluation technique is correct for the extractive approach but not for the abstractive approach. So we have to propose a new variant of the ROUGE-1 evaluation metric called ROUGE-SS that works on both exact matches of words and synonyms of particular words.

The implementation of this proposed technique (ROUGE-SS) is done in Python language. The proposed ROUGE-SS evaluation technique is evaluated in different datasets like CNN/Daily Mail, DUC-2004, Gigawords, and Inshorts News datasets. The ROUGE-SS gives a better result than other ROUGE variant
metrics. The demonstration is also shown in Tables 1 and 2. The comparison table of the proposed metric and the existing metrics are in Table 3.

The pictorial representation of the comparison table is also shown in Graph 2. The proposed technique is helpful in the measurement of the text summarization model, increasing the accuracy of the evaluation metric.

These metrics have some limitations, such as being based on word overlap or not accounting for the semantic content of the summary. Another limitation is human subjectivity. Ultimately, the quality of a summary is subjective and may vary depending on the reader’s preferences and expectations. Therefore, even human evaluators may not always agree on the quality of a summary.

In the future, text summarization tools might be evaluated using more advanced methods that look at things like how well the summary flows, how easy it is to read, and how informative it is. It will make the evaluation process more accurate and valuable for people who want to summarize the text. Another possible direction is to use more advanced NLP techniques, such as semantic parsing and semantic role labelling, to evaluate the semantic content of a summary.

References


and Information Sciences, 34 (4), 1029-1046.


[34] Ng, J. P., & Abrecht, V. (2015). Better summarization evaluation with word embeddings for


An abstract meaning representation graph-to-text deep learning approach.


[38] Liu, Y., Jia, Q., & Zhu, K. (2022, July). Reference-free Summarization Evaluation via Semantic Cor-
relation and Compression Ratio. In *Proceedings of the 2022 Conference of the North American Chapter of
the Association for Computational Linguistics: Human Language Technologies* (pp. 2109-2115).


[40] D. Suleiman and A. Awajan, “Deep Learning Based Abstractive Text Summarization: Approaches,
10.1155/2020/9365340.

Caption Translation Using Transformers and Object Detectors,” in *Applications of Artificial Intelligence


in abstractive summarization.


[46] Kumar, S., Solanki, A. An abstractive text summarization technique using transformer model with