Journal of Universal Computer Science, vol. 30, no. 13 (2024), 1807-1828
submitted: 6/11/2023, accepted: 26/6/2024, appeared: 28/12/2024 CC BY-ND 4.0

The Bart-based Model for Scientific Articles Summarization

Mehtap Ülker
(Computer Engineering Department, Firat University, Elazig, Turkey
https://orcid.org/0000-0001-8680-8518, m.ulker@firat.edu.tr)

A. Bedri Özer
(Computer Engineering Department, Firat University, Elazig, Turkey
https://orcid.org/0000-0002-8005-7386, bedriozer@firat.edu.tr)

Abstract: With the development of deep learning techniques, many models have been proposed for abstractive text summarization. However, the problem of summarizing source documents while preserving their integrity persists due to token restrictions and the inability to adequately extract semantic word relations between different sentences. To overcome this problem, a fine-tuning BART-based model was proposed, which generates a scientific summary by selecting important words contained in the text in the input document. The input text consists of terminology and keywords from the source document. The proposed model is based on the working principle of graph-based methods. Thus, the proposed model can summarize the source document with as few words as possible that are relevant to the content. The proposed model was compared with baseline models and the results of human evaluation. The experimental results demonstrate that the proposed model outperforms the baseline methods with a 37.60 ROUGE-L score.

Keywords: Text summarization, Abstractive method, SciIE, TF-IDF, TextRank
Categories: H.3.1, H.3.2, H.3.3, H.3.7, H.5.1
DOI: 10.3897/jucs.115121

1 Introduction

Researchers publish novel scientific articles to contribute to science. They want to access the salient information in long documents as quickly as possible without reading the entire document. However, manually extracting salient information from these documents is very challenging in terms of time and cost. Therefore, text summarization has recently become widespread. Text summarization presents the content of the document as quickly and concisely as possible while preserving the text's integrity [Alomari et al., 22, Al-Zahrani et al., 15, Li et al., 22]. It is applied in two basic ways: extractive and abstractive. In extractive methods, the source document is summarized by extracting the most important sentences, whereas in abstractive methods, it is summarized according to the rewriting of sentences with novel words. In this respect, abstractive approaches resemble human-generated summaries and are more difficult than extraction approaches.

With the development of deep learning techniques, the success of abstractive text summarization methods has improved considerably. In particular, fine-tuning-based methods such as BART [Yadav, H., et al., 2023], BERT [Aksenov, D., et al., 20], and

T5 [Etemad, A.G., et al., 21] are widely preferred for abstractive text summarization. The success of fine-tuning models is based on (i) latent representations of words in the source document [Pang et al., 22], (ii) the length of the source document [Aksenov, D., et al. 20, Bajaj, A., et al. 21, Borah, M., et al. 22], and (iii) domain knowledge [Laskar, M. T. R., et al. 22, Moro, G., and Ragazzi, L. 22]. Latent representations are extracted from the bottom-up using the transformer encoder. Self-attention-based models on transformer encoders face the problem of quadratic complexity with respect to sequence length [Pang et al., 22]. In addition, the length of the source document is restricted because of the token limitation of transformer models [Feijo, D. D. V., and Moreira, V. P. 2023]. Therefore, recent studies have focused on models for summarizing longer documents [Cao, S., and Wang, L. 23; Pu, D., et al. 2022]. However, it is very challenging to summarize long documents due to factors such as preserving semantic information in the content [Grail, Q., et al. 21], extended structured input context [Zhang, H., Liu, X., and Zhang, J. 2022], and memory [Moro, G., and Ragazzi, L. 22].

Scientific articles are longer than other text types, such as news and tweets. Summarization of scientific articles requires expertise in domain-related terminology. The salient features are not adequately researched and discussed in each section of scientific articles. Some subsections contain general information about the subject of the document. Therefore, it is not the right approach to use each section for generating a summary. The aim of scientific article summarization is to extract important information and transform it into a semantic and coherent sentence [Cai et al., 22, Altmami et al., 20]. Therefore, recent studies have focused on summarizing models and extracting salient content-related information [Wang et al., 20, Wang et al., 20]. In particular, graph-based methods (GNN- Graph Neural Network [Xu, K., et al. 2018], GAN-Graph Attention Network [Veličković, P., et al., 2017], and GAT- Graph Attention Transformer [Yun, S., et al., 2019]) generate summary sentences on salient words or word groups related to the content [Gupta, S., and Gupta, S. K. 2019, Zhao, L., Xu, W., and Guo, J. 2020]. In graph-based methods, methods such as named entity extraction and automatic key extraction are used for word selection [Koncel-Kedziorski, R., et al. 2019]. Candidate sentences are generated by learning the relationships between nodes. The graph-based method is widely preferred for long document summarization because it preserves the integrity of the document.

In fine-tuning models, the input text is limited to token numbers [Feijo, D. D. V., and Moreira, V. P. 2023]. Most studies summarize scientific articles with token lengths within certain ranges, which is a problem that needs to be addressed. To address these problems, a fine-tuning Bart-based model that summarizes scientific articles in as few words as possible was proposed. In this study, a novel model was developed with inspiration from the working principle of graph-based models, which is to select the most important words or word groups in the document and generate sentences based on the relations between the words. In the proposed model, semantic relationships between words are extracted using a transformer encoder. Because the number of input tokens is constrained and selected according to certain criteria, it is easier to extract the relations at the transformer encoder than in a longer document. Thus, it is important to summarize the source document in as few words as possible while preserving its integrity. The main contribution of this study is to summarize scientific articles in as few words as possible, based on the working principle of graph-based methods. More specifically,

- A novel fine-tuning BART-based model that generates a summary from entity names, relations (terminology related to content), and keywords is proposed to summarize scientific articles.
- Two datasets are proposed for abstractive scientific summarization, which consist of entity names, relations, keywords, and abstracts.

The rest of the study is organized as follows. In Section 2, related work is explained. In Section 3, the proposed model is explained. In Section 4, the comparison results of the proposed model with state-of-the-art methods and a discussion are given. The last section provides a general evaluation of this study.

2 Related Work

In this section, existing studies on abstractive text summarization that summarize long documents are reviewed. [Xiao et al., 21] proposed the PRIMERA model for multi-document representation, eliminating the need for large amounts of fine-tuning labeled data. Encoder– decoder transformers were used to facilitate the processing of combined input documents. Experiments were performed on six multi-document summary datasets from three different fields and compared with baseline methods. The experimental results showed that the proposed model outperformed the baseline methods. [La Quatra and Cagliero, 22] emphasized the main challenges of implementing recurrent neural networks as the need for ad hoc linguistic features and inefficiency in learning long-range text dependencies. To improve the effectiveness of the approach, they proposed a text summarization model based on highlight extraction, which used the attention mechanism in the transformer model. Experiments were conducted using three different benchmark datasets. The results of the proposed model were notably superior to those of the baseline models.

[Oh et al., 23] mentioned that the automatic summary of scientific articles is quite difficult because of their structured format and longer length than other text types. To extract a balanced summary from a section of a scientific article, they proposed a novel model that summarizes the article in the IMRaD format. [Saini et al., 23] presented a multi-view clustering (MVC) framework for scientific text summarization. The MVC framework evaluates scientific documents from two perspectives: semantic and syntactic. Sentence-level features such as sentence length and position were analyzed to extract high-scoring sentences for the summary. This study utilizes two approaches for representing document sentences in a semantic space: two embedding spaces and citation contexts. The proposed model was compared with a single-view clustering framework using the CL-SciSumm 2016 dataset and CL-SciSumm 2017. According to the evaluation results, the MVC framework outperformed single-view clustering.

[Pang et al., 22] mentioned that the latent representations of words or tokens in the source document are considered crucial for summarization. A bottom-up transformer encoder model is used to extract the latent representation. However, the self-attention model faces the problem of sequence length. To improve summarization models, they proposed a novel framework that focused on uncovering long-text dependency, assuming that the token's sub-level preserves details.

[Mishra et al., 22] introduced a novel approach based on a multi-objective differential evolution technique for summarizing scientific documents. Salient sentences were identified using citation contextualization. These sentences were

clustered using multi objective clustering. The multi-objective differential evolution algorithm was used to measure the compactness and separation of sentence clusters using two objective functions, namely the PBM index and the XB index. The experiments were conducted on CL-SciSummNet datasets and compared with the baseline model. The experimental results indicated that the proposed model outperformed the baseline models.

[Cachola et al., 20] noted that the summarization of scientific articles requires expert background knowledge and an understanding of complex domain-specific language. To overcome this problem, they proposed a BART-based model (CATTS) that generates a summary using the titles. As a result of both automatic measurements and human evaluations, CATTS has demonstrated exceptional performance.

[Huang et al., 21] mentioned that large transformers have quadratic computation and memory complexities, which can limit their scalability for long document summarizations such as scientific articles. To address this problem, they proposed a novel, efficient encoder-decoder attention with head-wise positional strides (HEPOS). Therefore, they extracted the salient content from the source documents.

[Krishna et al., 23] pointed out that existing summary datasets lack sufficient additional explanations to support research on all aspects of summarization. To overcome this problem, they presented a USB benchmark consisting of eight tasks that challenge the multi-dimensional understanding of summarization. The benchmark was trained and evaluated on transformer-based fine-tuned models, including RoBERTa-Large, T5-Large, Flan-T5-Large, Flan-T5-XL, and Flan-T5-XL. It has been shown that scientific articles can be summarized effectively with these models.

[Kondadadi et al., 21] proposed a BART-based model for radiology report summarization. The proposed method was trained with MIMIC-CXR, consisting of the radiology reports. The proposed model was compared with the T5 and PEGASUS models, and it was stated that the proposed model was superior in summarizing these data.

[Shen et al., 23] emphasized that pre-trained language models (PLMs) are not as successful at multi-document abstraction (MDS), where interactions between documents are more complex. To perform multi-document interactions, they enabled hierarchical coding in an existing PLM by modifying the attention masking scheme. Therefore, it ensures that the PLM does not introduce new parameters and can be fine-tuned directly on an MDS task dataset. They trained the Bart-based model and made it possible to summarize it in their scientific articles.

Transformer-based fine-tuned models are one of the most commonly used methods for summarizing long documents. Content-based latent semantic information is easily obtained using the attention mechanism in the encoder. Therefore, it is necessary to analyze the entire article as thoroughly as possible. However, this is not possible because of the token constraint. To address the difficulty of reading and the comprehensibility of scientific article summaries, they proposed a Flan-T5-based model [1] that rewrites the summaries and converts them into a format that can be understood by everyone. Flan-T5 is a model trained with a dataset of 1.8 K NLP tasks initiated with a T5 checkpoint. FLAN-T5 has achieved impressive results on various NLP tasks [Shen et al., 23].

[1] https://huggingface.co/haining/scientific_abstract_simplification#Usage

The input text of BART-based models is limited to token number. Therefore, there are problems with summarizing long documents. Within the scope of the study, the BART model was trained to summarize the document in as few but important words as possible. The proposed model was carried out by adopting the approach of graph-based models in the selection of input text.

3 Proposed Model: Fine Tuning Bart-Based Scientific Summarization Model

The fine-tuning Bart-based summarization model is shown in Figure 1. The basic processes of the proposed model are as follows: during the processing stage, the stop words and numbers are removed from the article. In the keyword extraction stage, the common keywords are obtained using the TextRank and TF-IDF algorithms. In the SciIE system, the entity names and relations are extracted from the source document.

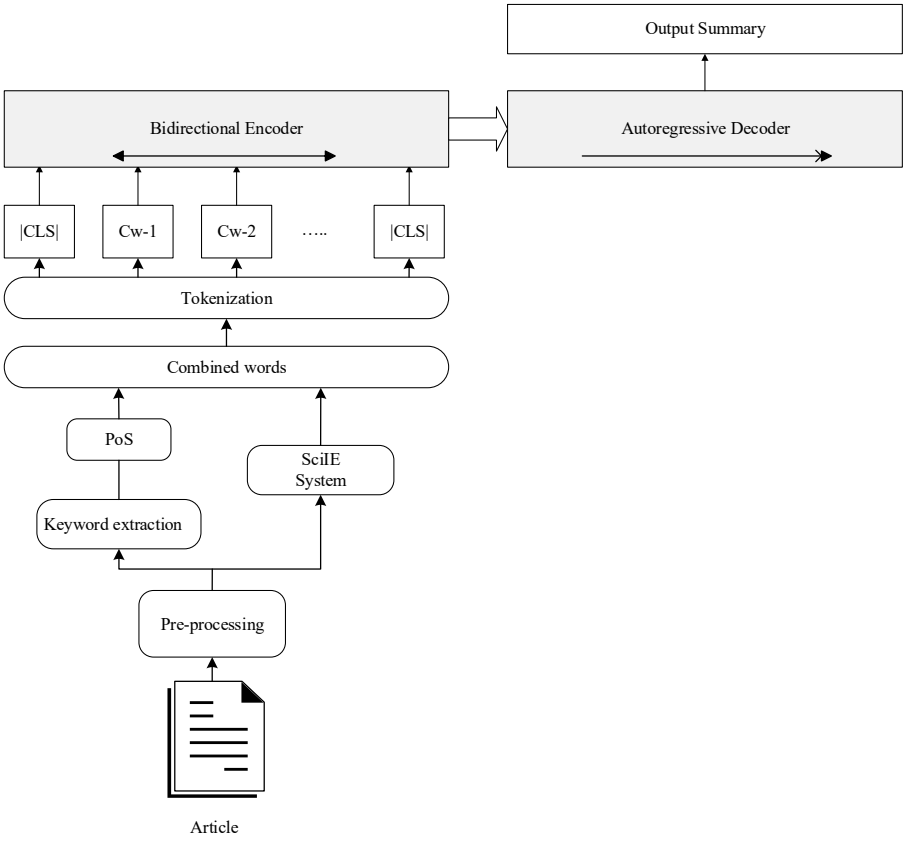

Figure 1: BART-based summarization model

3.1 Keyword Extraction: TextRank & TD-IDF Algorithm

Statistical-based methods are one of the most preferred methods for keyword extraction because of their language independence. Using this method, the importance of the sentence is determined according to the frequency of the words in the sentence [Gupta et al., 10]. One of the most commonly used methods is term frequency. The term frequency refers to the number of times a word is repeated in a document. It is calculated using Equation (1).

$$f(w) = \frac{n(w)}{N} \tag{1}$$

where $n(w)$ is the frequency of the word w in the document, N is the total number of words in the document. S_j represents the weight of each sentence (S).

With term frequency, conjunctions are determined as important words along with keywords. To handle this problem, the term frequency-inverse document frequency (TF-IDF) is proposed. TF-IDF determines words that best describe the content according to the frequency of the words in the text. To calculate the TF-IDF values, TF (term frequency) and IDF (inverse document frequency) are calculated in Equation (2) [Salton et al., 88].

$$TF - IDF_i = TF_{ij} * \log\frac{|n|}{|Df_i|} \tag{2}$$

where TF_{ij} indicates the number of occurrences of word i in the j document. |n| represents the number of documents. Df_i indicates the number of documents containing the word i.

Despite TF-IDF's strong algorithm for extracting keywords, the semantic relationship between tokens is completely ignored in this method. In this study, this problem was overcome using the TextRank algorithm with TF-IDF. The TextRank algorithm can be used to extract keywords based on semantic similarity without requiring linguistic or domain knowledge. The TextRank algorithm based on graphs [Mihalcea et al., 04] was developed to extract salient or important words from the source document. In graph G, which is expressed as G = (V, E), V denotes nodes composed of words, and E (edge) represents word scores. This algorithm is also widely used for text summarization and keyword extraction. The mathematical expression is given by Equation (4).

$$W_{ij} = \frac{V_i \cdot V_j}{\|V_i\|\|V_j\|} = \frac{\sum_{k=1}^{n} V_{k,i} V_{k,j}}{\sqrt{\sum_{k=1}^{n} {V_{k,i}}^2}\sqrt{\sum_{k=1}^{n} {V_{k,j}}^2}} \tag{3}$$

$$S(V_i) = (1-d) + d * \sum_{j \in V_j} \frac{S(V_j)}{|Out(V_j)|} \tag{4}$$

where, V_i refers to the words in the document, $Out(V_i)$ the set of connections coming out of the Vi node and d. d is a parameter that can be adjusted between 0 and 1.

In this study, statistically based keywords were extracted using Equation (2). Each token was represented in the node after the sentences were tokenized. To determine candidate keywords, the weight (W_{ij}) according to the V_i and V_j vectors was calculated as in Equation (3). Common keywords from both algorithms were combined. Keywords

with the part of speech) tags "Noun" and "Propn" were extracted. The terms with the highest weight were expressed as keywords.

3.2 Entity Name and Relation Extraction with the SciIE System

In human-generated summaries, the summaries are generated after entity names and relations are well understood. Keywords are not sufficient to be used alone in a descriptive summary, as it is not guaranteed that the keywords in the article contain entity names or relations. It is possible to generate more descriptive summaries by combining these words with keywords. The SciIE system is one of the NER systems that extract entities, relations, and co-references developed for scientific articles [Luan et al., 18]. As entity names, six different types are extracted: task, method, metric, material, other scientific term, and generic. As relations, seven different types are extracted: compare, part-of, conjunction, evaluate-for, feature-of, used-for, and hyponym-of. The mathematical expression of these operations is given in Equation (5).

$$P(E,R,C\,|D) = P(E,R,C,S\,|D) = \prod_{i=1}^{N} P(e_i|D)P(c_i|D)\prod_{j=1}^{N} P(r_{ij}|D),$$
$$P(e_i = e|D) = \frac{\exp\left(\Phi_E(e,s_i)\right)}{\sum e' \in L_E\, exp(\Phi_E(e',s_i))}$$
$$P\left(r_{ij} = r\middle|D\right) = \frac{\exp\left(\Phi_R\left(r,s_i,s_j\right)\right)}{\sum r' \in L_R\, exp(\Phi_R\left(r',s_i,s_j\right))} \tag{5}$$
$$P(c_i = j|D) = \frac{\exp\left(\Phi_C\left(s_i,s_j\right)\right)}{\sum j' \in \{1,\ldots,i-1,\mathcal{E}\}\, exp(\Phi_c\left(s_i,s_{j'}\right))'}$$

where N denotes sentence numbers in the document. D represents a sequence of words, and S represents the set of all possible within-sentence word sequence spans in the source document. SciIE generates three outputs: entity types (E), relations (R), and co-reference links (C). Φ_E represents the unnormalized model score for an entity type (e) and a span (s_i). Φ_R represents the score for relation types (r) and span pairs. Φ_c represents the score for a binary reference link between span pairs. L_E refer to all possible entity types including the null-type $\mathcal{E}$. L_R refer to all possible relation types including the null-type $\mathcal{E}$.

In the preprocessing stages, punctuation marks, special characters, and co-references were removed from the entity names and relations extracted by the SciIE system. Then, common keywords, which were extracted using TextRank and the TF-IDF algorithm, were combined with entity names and relations. Therefore, keywords were selected from the semantic, statistical, and entity names in the document. To generate semantic sentences, the relations were combined with keywords. The combined words (cw) are tokenized as in Equation (6).

$$\text{Combined words} = \{Cw_1, Cw_2, \ldots, Cw_n\} \tag{6}$$

Algorithm 1. Scientific summarization algorithm

Inputs:
d: document
d ←preprocessing(d)
Common keywords (TF-IDF(d), TextRank(d))
common_keywords ← []
tfidf_words← TF-IDF(d)
textrank_words← TextRank(d)
for i←0 to tfidfwords.length **do**
for j←0 to textrank_words. length **do**
if(tf-idf_words[i]== textrank_words[j]) **then**
common_keywords← tfidf_words[i]
end if
end for
end for
return common_keywords
common_keywords← PoS_Tag(common_keywords)
Entity names, Relations ← SciIE(d)
Combined words (common_keywords, Entity names, Relations)
Entity names← preprocessing_ner(Entity names)
Relations ← preprocessing_ner(Relations)
combined_words ← common_keywords ∪ Entity names ∪ Relations
return combined_words
Encode(text_input←combined_words or common_words):
tokenizer ← tokenizer (model_name)
model ← model(model_name)
inputs ←tokenizer(text_input, return_tensors="pt")
labels←torch.tensor([1]).unsqueeze(0)
h ← model(**inputs)
return(h)
$\boldsymbol{y_t = SOS, t = 0}$
$\boldsymbol{y_t}$**: Prediction output**
Decode(h)
while $\boldsymbol{y_t \neq}$ **|CLS| do**
$\boldsymbol{c_t}$← attention layer(h)
$\boldsymbol{y_t}$← autoregressive_decoder($\boldsymbol{c_t}$)
$\boldsymbol{t}$ +←1
end while
return $\boldsymbol{y_t}$

3.3 Fine-tuning the BART-based model

BART is a transformer-based model constructed by combining a bidirectional encoder and an auto-regressive [Lewis et al., 19]. The performance of the BART model is increased by transformation functions such as token masking, token deletion, sentence permutation, document rotation, and text infilling. In token masking, random tokens

are replaced with [MASK]. In token deletion, random tokens are deleted and expected that the model would find out at which positions the entries are missing. In sentence permutation, sentences are separated according to full stops. Document rotation is started by rotating a random token from the document. This task trains the model to identify the start of the document. Text padding samples different span lengths in the distribution and replaces each span with a sequence of [MASK] tokens of the same length. As a result, the proposed model reduces the transformations that cause noise in the source document before pre-training, thanks to having as few words as possible. According to this, the proposed model is trained using the reconstructed source document.

The input sequence is given to the encoder, and the output is generated autoregressively at the decoder input. When the decoder receives the word- or sentence-level vector from the encoder, it must interpret these vectors to map the target output. The number of layers is N=12. Each layer consists of a multi-headed attention network and a feed-forward network. While designing the decoder, they were inspired by the GPT (Generative Pre-Training) model. In the GPT model [Radford et al., 18], words are left context-conditioned. Therefore, they cannot learn bidirectional interactions. In this study, the BART-based model was fine-tuned according to the hyperparameters given as fallows. We set the number of beams to 4 and the batch size to 32. The fine-tuning step is performed using the Adam optimizer with a maximum learning rate of 1e−5, hidden size of 768, dropout rate is .0.1.

4 Research Findings and Discussion

The proposed model performed the experiment on NVIDIA 1650 and compared it with the baseline models on SciSummNet and the proposed datasets, which contain scientific articles. To demonstrate the importance of entities and relations in text summarization, the proposed model was trained on two different scenarios.

- **Scenario-1 ($BARTSum_{CommonWords}$):** The proposed model was trained with the common keywords extracted by the TF-IDF and TextRank algorithms and the abstract.
- **Scenario-2 ($BARTSum_{CombinedWords}$):** The proposed model was trained by combining the keywords and entities, i.e., the relations extracted from the SciIE system, with the common keywords.

4.1 Dataset

To evaluate the performance of the proposed model, experiments on the SciSummNet dataset and the proposed dataset were conducted. The proposed dataset (ArxivComp) comprises 16 K scientific articles. The datasets were crawled from the arXiv website. The summaries written by the author **(gen)** were used to evaluate the summaries produced **(gold)** by the proposed model. The dataset was prepared from two different sections.

- **Dataset-section-1** consists of abstracts written by the author (gen) and common keywords obtained from the TF-IDF and TextRank algorithms.
- **Dataset-section-2** consists of summaries written by the author (gen) and combined words (common words, entities, and relations).

The SciSummNet dataset was prepared using the same approach (Scenario-1 and Scenario-2). However, some articles were not evaluated because entity names could not be extracted from all these articles using SciIE systems. Average word length is calculated based on the ratio of the total number of words in the document to its size. The dataset properties are given in Table 1.

Dataset	Document Size	Source Avg. Word Length
SciSummNet [Yasunaga et al., 19]	676 - 813	17.55 (Scenario-1)
		93.10 (Scenario-2)
Proposed dataset (ArxivComp)	16.040–16.807	12.27 (Scenario-1)
		92.50 (Scenario-2)

Table 1: Dataset properties

4.2 Fine-tuning Phase

To ensure fair comparison of the proposed model performance, BART-based fine-tuning was performed using the same set of hyperparameters for the two datasets. First, the input training data of the proposed dataset, which consists of 16.040–16.807 scientific articles (Scenario-1 and Scenario-2) and abstracts written by the author, are passed to the auto-regressive encoder and trained with training arguments. Then, the input training data of the SciSummNet dataset, which consists of 676–813 scientific articles (Scenario-1 and Scenario-2) and abstracts written by the author, is also passed to the auto-regressive encoder and trained with training arguments. Specifically, we set the number of beams to 4 and the batch size to 32. The fine-tuning step is performed using the Adam optimizer with a maximum learning rate of 1e−5.

4.3 Evaluation Metric

The Rouge metric (Recall Oriented Understudy of Gisting Evaluation) has been widely used to evaluate the similarity between the author-created summary **(gen)** and the model-created summary **(gold)**. [Lin, 04]. ROUGE calculates n-gram-based recall to determine whether it overlaps gen and gold. Although the three commonly used metrics are ROUGE-1, ROUGE-2, and ROUGE-L, it is preferred over others in demonstrating the performance of a Rouge-L abstractive summarization model that deals with the overlap of the longest common words (LCS). The calculation of ROUGE-N is given in Equation (7).

$$ROUGE - N = \frac{\sum_{S \in \{ReferanceSummaries\}} \sum_{gram_n \in S} Count_{match}\,(gram_n)}{\sum_{S \in \{ReferanceSummaries\}} \sum_{gram_n \in S} Count\,(gram_n)} \tag{7}$$

where N, $Count\ (gram_n)$, $Count_{match}(gram_n)$ are represented as the length of the N-gram, the count of n-grams in the reference summaries, and the maximum number of matching words in the candidate summary, respectively.

The ROUGE score compares **gold** based on the overlapping of the n-grams with the **gen**. However, it is not sufficient to prove the quality of the generated summaries. To handle this problem, gold was also evaluated using human judgment. The Fleiss' kappa coefficient was used to measure the reliability of the comparative agreement between more than two fixed numbers of volunteers. The Fleiss' kappa coefficient is calculated using Equation (8) [Fleiss, J. L. 71].

$$K = \frac{\bar{P} - \bar{P_e}}{1 - \bar{P_e}} \tag{8}$$

$$\bar{P} = \frac{1}{N} \sum_i^N P_i \tag{9}$$

$$\bar{P_e} = \sum_{j=1}^k \rho_j^2 \tag{10}$$

$$\rho_j = \frac{1}{N * n} \sum_{i=1}^{N} n_{ij} \tag{11}$$

$$P_i = \frac{1}{n(n-1)} \sum_{i=1}^{N} n_{ij}\ (n_{ij} - 1) \tag{12}$$

where N is the number of samples to be evaluated, n is the total number of evaluators, k represents the number of categories to be used in the evaluation, $\bar{P}$ is observation agreement, $\bar{P_e}$ is sum of criterion-based evaluation probability, and p_j is possibility of criterion-based evaluation. As a result, the K value obtains a value between 0 and 1. The closer this value is to 1, the higher is the consistency among raters.

4.4 Baseline Methods

To ensure fairness in comparison, the proposed model performance was compared with the baseline models, which are fine-tuned models summarizing scientific articles. These models were trained using our dataset and fine-tuning parameters.

- **T5-small-finetuned-summarize-scientific-articles model** [2] is a model fine-tuned to summarize scientific models, which on huggingface is a T5-based model that provides a comprehensive framework for addressing all text-based language problems [Raffel et al., 20].
- **Bart-base-finetuned-summarize-scientific-articles**[3] is a model fine-tuned to summarize scientific models.
- **Easumm** is a model that summarizes scientific articles in the biomedical field, based on graph-based methods [Frisoni, Giacomo, et al. 23]

[2] https://huggingface.co/mrm8488/t5-base-finetuned-summarize-news
[3] https://huggingface.co/sana-ngu/bart-base-finetuned-summarize-scientific-articles

4.5 Evaluation Results

To evaluate the performance of the proposed model, two different scenarios (Scenario-1 / $\boldsymbol{BARTSum_{CommonWords}}$ and Scenario-2 / $\boldsymbol{BARTSum_{CombinedWords}}$) were handled, and experiments were conducted on SciSummNet and the proposed dataset. The evaluation results were compared with the baseline methods.

4.5.1 Evaluation results of Scenario 1

As shown in Table 1, the average word length for Scenario-1 of the proposed dataset is 12.27 and the average word length for the SciSummNet dataset is 17.55. When the results in Table 2 are examined, the proposed model and baseline models show higher performance on the SciSummNet dataset than the proposed dataset. These results show that more keyword extraction improves the success of document summarization.

The results in Table 2 show that the summaries produced are not effective enough when the keywords are given as the input text to the baseline models instead of the entire document. With the proposed model, it can be seen that the value of ROUGE-1/2/L improves significantly and generates a summary with as few keywords as possible. When common keywords are given instead of the entire text, Rouge L 15.81 and 17.31 are better than baseline models.

The proposed model showed superior performance compared with the baseline model on both datasets. However, the proposed model achieved performance comparable to that of the model Bart-base-finetuned-summarize-scientific-article. Both models are Bart-based. However, the main difference between these models is that we trained the proposed model with our own hyperparameter using a pre-trained model with CNN Daily Mail. In the other model, a language model trained with scientific articles was trained with our hyperparameters. It can be clearly seen in Table 2 that the proposed models and Bart-base-finetuned-summarize-scientific-articles outperformed the T5 model thanks to the auto-regressive decoder and bidirectional encoder.

EAsumm extracts scientific content according to domain knowledge and creates a graph-based model. In the proposed model, scientific contents were extracted with the SciIE system. The proposed data set consists of either common words or scientific content. It is very difficult to produce a summary by re-extracting scientific content by the semantic integrity of the sentence from the proposed data set with the proposed data set. When Table-2 and 3 results are examined, it shows that the easumm model is not suitable for summary generation according to Scenario-1 and Scenario-2.

	Model	Rouge-1	Rouge-2	Rouge-L
Proposed dataset	T5-small-finetuned-summarize-scientific-articles model (T5-based)	12.80	0.05	11.44
	Bart-base-finetuned-summarize-scientific-articles	28.52	5.55	**16.78**
	EASumm	11.74	0.02	10.05
	Proposed Model	**29.18**	**4.79**	15.81
SciSummNet dataset	T5-small-finetuned-summarize-scientific-articles model (T5-based)	16.52	1.12	14.01
	Bart-base-finetuned-summarize-scientific-articles	27.02	4.76	16.68
	EASumm	12.25	2.17	12.18
	Proposed Model	**29.11**	**5.55**	**17.31**

Table 2: Evaluation results of Scenario-1

4.5.2 Evaluation results of Scenario 2

Entity names and relations (combined words) are critical words for revealing the main topic in scientific articles. These can be used in generating a summary as they contain information such as subject, method, and task. Within the scope of Scenario 2, the fine-tuned Bart-based model was trained using common keywords and combined words.

As shown in Table 1, the average word length of the proposed dataset for Scenario 2 is 92.50, and for the SciSummNet dataset, it is 93.10. The average document lengths for both datasets are approximately the same. The results in Table 3 show that the best result is obtained in the proposed dataset when the document length is the same. The main reason is that the proposed dataset is trained with a larger document size than the SciSummNet dataset. The proposed model shows superior performance compared to Scenario-1 because of training in terminology related to the source documents, which are extracted with the SciIE system.

The proposed model was prepared based on the graph-based approach. Words or word groups related to terminology were extracted with the SciIE system from the document. Therefore, Scenario 2 in both datasets was prepared with this approach. The results in Table 3 show that the success of summaries based on graph-based approach (Scenario 2) is higher than the success of summaries extracted with common words.

As shown in Table 3, the proposed model showed superior performance compared with the baseline methods in Scenario 1. It results from uncovering and summarizing terminology that covers the entire document. For the SciSummNet dataset, the results were very similar. The Rouge-2 result of the T5-based method was lower than that of the other results, with 6.76. Therefore, it appears to be lower in summarization compared with both the proposed model and the Bart-based method in Tables 2 and 3. The highest result was obtained using the proposed method for the proposed dataset.

The proposed model is affected document size, average word length (document length), and input word types (common words/combined words). The document size during training can affect summarization positively or negatively. As can be seen in Table-1, the document size of the SciSummNet dataset is less than the proposed dataset. The average word numbers in the two data sets are less in Scenario-1 than in Scenario-2. Scenario-1 consists of common words, and Scenario-2 consists of combined words (keywords and terminology-related words). As seen in Table 2, the summarization performance is lower even on the proposed dataset, which is larger in size but generates summaries with only common words. When the results in Table 3 are examined, increasing document length, increasing average word length, and using terminology-related words along with keywords significantly increase the success rate of the model. However, generating a summary with common keywords is a disadvantage of the proposed model in that it cannot achieve the integrity of the document.

	Model	**Rouge-1**	**Rouge-2**	**Rouge-L**
Proposed dataset	T5-small-finetuned-summarize-scientific-articles model(T5-based)	24.28	7.18	21.56
	Bart-base-finetuned-summarize-scientific-articles (BART-model)	49.94	23.85	35.37
	EASumm	22.98	6.15	20.15
	Proposed Model	**52.92**	**27.32**	**37.60**
SciSummNet dataset	T5-small-finetuned-summarize-scientific-articles model (T5-based)	24.70	6.76	21.51
	Bart-base-finetuned-summarize-scientific-articles (BART-model)	40.30	16.30	**24.27**
	EASumm	23.18	8.25	20.22
	Proposed Model	**43.44**	**18.19**	23.75

Table 3: Evaluation results of Scenario 2

The input parameters of these fine-tuning models are limited to a certain range of words. For example, the maximum input text length is 512 for the BERT-based baseline [Devlin, J., et al. 2018, Chalkidis, I., et al. 20]. The maximum input text length is 512 for the T5-based baseline [Raffel et al., 20]. For the Bart-based model, the maximum input text length is 512 [Lewis et al., 19]. In general, the main reason for limiting the word lengths of models is memory [Rohde, T., Wu, X., and Liu, Y. 2021]. In this study, we propose a method that summarizes the text by considering the integrity of the source document without the need for token restrictions. The proposed model summarizes the text and integrity of the source document without token restriction. Thus, long documents can be summarized in as few words as possible. As shown in Table 1, salient words and terminology related to the source document are extracted. While the average word count for the proposed dataset is 92.50, the average word count for SciSummnet is 93.10. The result in Table 3 shows that it is less than the maximum word limit of the BART-based model, and good success is achieved with Rouge L 37.60. Thus, long documents can be summarized in as few words as possible.

The proposed method for Scenarios 1 and 2 is superior to the baseline methods, with an F1 score Rouge-L value of 15.81 and an F1 score Rouge-L value of 37.60. With the aim of the proposed model, it has been observed that long document summarization can be easily performed with combined words. Thus, it has been concluded that scientific articles can be summarized in as few, however salient, words as possible.

4.5.3 Human evaluation results

In Table 4 and Table 5, the human evaluation results are given according to Scenarios 1 and 2. As evaluation criteria, the five criteria were considered [Cai et al., 22]. The evaluation criteria are as follows:

- **Conciseness (Cn)** is whether to avoid redundant information.
- **Informativeness (I)** is whether it contains important information.
- **Coherence (Ch)** is whether the content of the gen is appropriate for gold.
- **Readability (R)** means that the generated summary is easy to understand and fluent.
- **Grammatically (G)** indicates whether the sentences are appropriate to grammar rules.

The five different undergraduate students in computer engineering were asked whether the summaries produced by the proposed model were suitable for the summaries produced by the author according to the above criteria. Each person wanted to rate it on a scale of 0 (worst) to 5 (best) by randomly selecting 50 articles. Tables 6 and 7 include summaries generated by both scenarios.

Fleiss's kappa coefficient [Fleiss, J. L. 71] was used to measure the reliability of comparative agreement between volunteers. The Fleiss kappa value is between 0 and 1. Because of Fleiss's kappa, the closer this value is to 1, the higher the consistency among volunteers.

	Cn	I	Ch	R	G
Mean/Variance	2.62 (0.26)	2.58(0.29)	2.43(0.29)	4.52(0.25)	4.59(0.28)
Fleiss kappa	0.34	0.38	0.44	0.42	0.39

Table 4. Human evaluation results for Scenario 1

	Cn	I	Ch	R	G
Mean/Variance	3.42 (0.24)	3.56(0.25)	4.54(0.26)	4.61(0.24)	4.76(0.14)
Fleiss kappa	0.36	0.40	0.66	0.57	0.54

Table 5: Human evaluation results for Scenario 2

Fleiss kappa results were evaluated according to 5 criteria: Conciseness, Informativeness, Coherence, Readability and Grammatically. Evaluation criteria are based on consistency between evaluators. Between 0 and 1, a value closer to 1 is considered closer to reality [Fleiss, J. L. 71] According to Tables 4 and 5, Scenario 2 produced more realistic summaries because the result of each criterion was higher than that of Scenario 1. It has been concluded that it is not sufficient to use only common keywords to generate a scientific article. It is seen that entity names and relations extracted from scientific articles can be summarized using common keywords without using long documents as input to summarize scientific articles.

Tables 6 and 7 include summaries produced according to Scenario-1 and Scenario-2. The generated summaries clearly show that the choice of keywords affects the production of summaries. Producing summaries with words related to terminology performs better in terms of being shorter, fluent, and concise, like summaries written by the author. As seen in Tables 4 and 5, human evaluation results show that the proposed model seems successful compared to method in Scenario-1 according to the "Conciseness, Informativeness, Coherence, Readability, and Grammatically " criteria.

Prediction-1 with $BARTSum_{CommonWords}$
In this paper, we introduce the concept of inertia in machine code generation. In order to explain the concept, we first define it as a programming language implementing an arbitrary stack-based procedure. We then introduce the notion of stack-to-stackvirtual machine (SCVMs). This machine is implemented by an abstract machine called an Inertia-Inertia Stack-To-Machine (ITAS). InITAS, the stack register is increasively incremented until only the register with the most recent instruction remains untouched. Finally, we present an implementation of our concept on a small number of benchmark programs.
Prediction-2 with $BARTSum_{CommmonWords}$
Autonomous vehicles have become more and more popular due to their large storage capacity and ability to store data. This has led to a never-increasing need for monitoring and maintenance of these machines. Logistic regression is a machine learning method that learns the relationship between the amount of available data and the machine state. In this paper, we propose an automated machine learning (AutoML) framework to automatically train Autonomous Vehicle (AutoVMs) models to detect machine failures, and apply a machine learning algorithm to automatically correct these failures. The AutoVMs are trained on a set of datasets, and an AutoML algorithm is trained on top of top-down machine learning techniques.
Prediction-3 with $T5-based\ model$
Machine learning (ML) is an emerging technology in the automotive industry. we propose a new machine learning methodologies based on machine learning. we propose a new approach to machine learning safety.
Prediction-4 with $BART-model$
Machine Learning (ML) is an open source framework for the development of machine learning. In this paper, we present a framework for the development of machine learning models.

Table 6: Summaries generated with Scenario-1

Prediction-1 with $BARTSum_{CombinedWords}$
Communication tools make the world like a small village and as a consequence people can contact with others who are from different societies or who speak different languages. This communication cannot happen effectively without Machine Translation because they can be found anytime and everywhere. There are numerous studies that have developed Machine Translation for the English language with so many other languages except the Arabic it has not been considered yet. Therefore, we aim to highlight a roadmap for our proposed translation machine to provide an enhanced Arabic English translation based on Semantic.
Prediction-2 with $BARTSum_{CombinedWords}$
Interpretable machine learning models are based on the mind's construction of concepts and meaning. In this paper, we present a method for classifying concepts and their meanings from the perspective of machine learning, which is used in daily life to make decisions and automatically create interpretable models.
Prediction-3 with $T5-based\ model$
Communication tools make the world like a small village and as a consequence people can contact with. In this paper, we present a proposed translation machine for arabic english translation. In this paper, we present a proposed translation machine for the arabic English translation.
Prediction-4 with $BART-model$
The last decade has seen huge progress in the development of advanced machinelearning models. Machine learning is a method for classifying concepts in daily life interpretable machine learning models. In this paper, we propose a novel method of classifying concepts and meaning.

Table 7: Summaries generated with Scenario 2

5 Conclusions

The summarization of scientific articles is still a challenging problem because it includes information such as token limitations, domain knowledge, and memory. In this study, a novel Bart-based model is proposed to address these problems. In addition, two different datasets have been proposed for abstractive scientific articles with keywords, entity names, and relations. It has been observed that the selection of these words affects summarization performance. When the input text is selected among keywords that cover the entire document, summarization performance decreases. However, it has been observed that when the input text is prepared with terminology and common keywords, the summarization results outperform. According to the evaluation results, the source document can be summarized in as few words as possible. In future studies, the effects of different words on summarization will be examined. It is also seen in the human evaluation results that the proposed model is quite successful compared to the baseline

models in summarizing a scientific article. The experimental results demonstrate that the proposed model outperforms the baseline methods.

References

[Aksenov, D., et al. 20] Aksenov, D., Moreno-Schneider, J., Bourgonje, P., Schwarzenberg, R., Hennig, L., & Rehm, G., (2020). Abstractive text summarization based on language model conditioning and locality modeling. *arXiv preprint arXiv:2003.13027*.

[Al-Zahrani et al., 15] Al-Zahrani, A. M., Mathkour, H., & Abdalla, H. I. (2015). PSO-Based Feature Selection for Arabic Text Summarization. *J. Univers. Comput. Sci*., 21(11), pages 1454-1469.

[Alomari et al., 22] Alomari, A., Idris, N., Sabri, A. Q. M., & Alsmadi, I., (2022). Deep reinforcement and transfer learning for abstractive text summarization: *A review, Computer Speech & Language*, 71(101276).

[Altmami et al., 20] Altmami, N. I., & Menai, M. E. B., (2020). Automatic summarization of scientific articles: A survey, *Journal of King Saud University-Computer and Information Sciences*, 34(4), pages 1011-1028.

[Bajaj, A., et al. 21] Bajaj, A., Dangati, P., Krishna, K., Ashok Kumar, P., Uppaal, R., Windsor, B., Brenner, E., Dotterrer, D., Das, R., McCallum, A., (2021). A. Long document summarization in a low resource setting using pretrained language models. *arXiv preprint arXiv:2103.00751*.

[Borah, M., et al. 22] Borah, M., Dadure, P., & Pakray, P., (2022). Comparative analysis of T5 model for abstractive text summarization on different datasets.

[Cachola et al., 20] Cachola, I., Lo, K., Cohan, A., & Weld, D. S., (2020). TLDR: Extreme summarization of scientific documents, *arXiv preprint arXiv:2004.15011*.

[Cai et al., 22] Cai, X., Liu, S., Yang, L., Lu, Y., Zhao, J., Shen, D., & Liu, T., (2022). COVIDSum: A linguistically enriched SciBERT-based summarization model for COVID-19 scientific papers, *Journal of Biomedical Informatics*, 127(103999).

[Cao, S., and Wang, L. 23] Cao, S., and Wang, L., (2023). AWESOME: GPU Memory-constrained Long Document Summarization using Memory Mechanism and Global Salient Content, *arXiv preprint arXiv:2305.14806*.

[Chalkidis, I., et al. 20] Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I., (2020), LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

[Devlin, J., et al. 18] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Etemad, A. G., et al. 21] Etemad, A. G., Abidi, A. I., & Chhabra, M., (2021). Fine-Tuned T5 for Abstractive Summarization. *International Journal of Performability Engineering*, 17(10).

[Feijo, D. D. V., and Moreira, V. P. 23] Feijo, D. D. V., and Moreira, V. P., (2023). Improving abstractive summarization of legal rulings through textual entailment. *Artificial intelligence and law*, 31(1), pages 91-113.

[Fleiss, J. L. 71] Fleiss, J. L., (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5), 378.

[Frisoni, Giacomo, et al. 23] Frisoni, G., Italiani, P., Moro, G., Bartolini, I., Boschetti, M. A., & Carbonaro, A., (2023). Graph-Enhanced Biomedical Abstractive Summarization Via Factual Evidence Extraction. *SN Computer Science*, pages 4.5 500.

[Grail, Q., et al 21] Grail, Q., Perez, J., & Gaussier, E. (2021). Globalizing BERT-based transformer architectures for long document summarization. *In Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics*, pages 1792-1810.

[Gupta et al., 10] Gupta, V., & Lehal, G. S., (2010). A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3), pages 258–268.

[Gupta, S., and Gupta, S. K. 19], Gupta, S., and Gupta, S. K. (2019). Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121, pages 49-65.

[Huang et al., 21] Huang, L., Cao, S., Parulian, N., Ji, H., & Wang, L., (2021). Efficient attentions for long document summarization. *arXiv preprint arXiv:2104.02112*.

[Huang et al., 22] Huang, Z., & Xie, Z., (2022).A patent keywords extraction method using TextRank model with prior public knowledge, *Complex & Intelligent Systems*, 8(2), pages 1-12.

[Koncel-Kedziorski, R., et al. 19] Koncel-Kedziorski, R., Bekal, D., Luan, Y., Lapata, M., & Hajishirzi, H. (2019). Text generation from knowledge graphs with graph transformers. *arXiv preprint arXiv: 1904.02342*.

[Kondadadi et al., 21] Kondadadi, R., Manchanda, S., Ngo, J., & McCormack, R., (2021), Optum at MEDIQA 2021: Abstractive summarization of radiology reports using simple BART finetuning, *In Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 280-284.

[Krishna et al., 23] Krishna, K., Gupta, P., Ramprasad, S., Wallace, B. C., Bigham, J. P., & Lipton, Z. C., (2023). USB: A Unified Summarization Benchmark Across Tasks and Domains, *arXiv preprint arXiv:2305.14296*.

[La Quatra and Cagliero, 22] La Quatra, M., & Cagliero, L., (2022). Transformer-based highlights extraction from scientific papers. *Knowledge-Based Systems*, 252(109382).

[Laskar, M. T. R., et al. 22] Laskar, M. T. R., Hoque, E., & Huang, J. X. (2022). Domain adaptation with pre-trained transformers for query-focused abstractive text summarization. *Computational Linguistics*, 48(2), 279-320.

[Lewis et al., 19] Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L. (2019), Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*, arXiv preprint arXiv:1910.13461.*

[Li et al., 22] Li, P., Lu, W., & Cheng, Q., (2022), Generating a related work section for scientific papers: an optimized approach with adopting problem and method information: *Scientometrics*, 127(8), pages 4397-4417.

[Lin, 04] Lin, C. Y., (2004). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74-81.

[Luan et al., 18] Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H., (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction, *arXiv preprint:1808.09602.*

[Mihalcea et al., 04] Mihalcea R, Tarau P., (2004). Textrank: bringing order into text*, In proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404-411.

[Mishra et al., 22] Mishra, S. K., Saini, N., Saha, S., & Bhattacharyya, P., (2022). Scientific document summarization in multi-objective clustering framework. *Applied Intelligence*, 52(2), pages 1520-1543.

[Moro, G., and Ragazzi, L. 22] Moro, G., and Ragazzi, L., (2022). Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. *In Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10), pages 11085-11093.

[Oh et al., 23] Oh, H., Nam, S., & Zhu, Y., (2023). Structured abstract summarization of scientific articles: Summarization using full-text section information. *Journal of the Association for Information Science and Technology*, 74(2), pages 234-248.

[Pang et al., 22] Pang, B., Nijkamp, E., Kryściński, W., Savarese, S., Zhou, Y., & Xiong, C., (2022). Long document summarization with top-down and bottom-up inference. arXiv preprint arXiv:2203.07586.

[Pu, D., et al. 2022] Pu, D., Hong, X., Lin, P. J., Chang, E., & Demberg, V., (2022). Two-stage movie script summarization: An efficient method for low-resource long document summarization. *In Proceedings of The Workshop on Automatic Summarization for Creative Writing*, pages 57-66.

[Radford C., et al., 18] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I., (2018). Improving language understanding by generative pre-training.

[Raffel et al., 20] Raffel, Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research,* 21(140), pages 1-67.

[Rohde, T., Wu, X., and Liu, Y. 21] Rohde, T., Wu, X., and Liu, Y., (2021). Hierarchical learning for generation with long source sequences. *arXiv preprint arXiv:2104.07545.*

[Saini et al., 23] Saini, N., Reddy, S. M., Saha, S., Moreno, J. G., & Doucet, A. (2023). Multi-view multi-objective clustering-based framework for scientific document summarization using citation context. *Applied Intelligence*, pages 1-25.

[Salton et al., 88] Salton, G., & Buckley, C., (1988), Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), pages 513-523.

[Shen et al., 23] Shen, C., Cheng, L., You, Y., & Bing, L., (2023). A hierarchical encoding-decoding scheme for abstractive multi-document summarization. *arXiv preprint arXiv:2305.08503.*

[Veličković, P., et al., 2017] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y., (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

[Wang et al., 20] Wang, Z., Duan, Z., Zhang, H., Wang, C., Tian, L., Chen, B., & Zhou, M. (2020). Friendly topic assistant for transformer based abstractive summarization, *Empirical Methods in Natural Language Processing (EMNLP),* pages 485-497.

[Wang, Y. Y., et al. 20] Wang, Y. Y., Wu, J. Y., Chou, T. H., Lin, Y. J., & Kao, H. Y. (2020). Enhance Content Selection for Multi-Document Summarization with Entailment Relation. *In 2020 International Conference on Technologies and Applications of Artificial Intelligence (TAAI),* pages 119-124 IEEE.

[Xiao et al., 21] Xiao, W., Beltagy, I., Carenini, G., & Cohan A., (2021). A.: Primer: Pyramid-based masked sentence pre-training for multi-document summarization. arXiv preprint arXiv:2110.08499.

[Xu, K., et al. 18] Xu, K., Hu, W., Leskovec, J., & Jegelka, S., (2018). How powerful are graph neural networks?. *arXiv preprint arXiv:1810.00826.*

[Yadav, H., et al., 23] Yadav, H., Patel, N., & Jani, D., (2023). Fine-Tuning BART for Abstractive Reviews Summarization. *In Computational Intelligence: Select Proceedings of InCITe,* Singapore: Springer Nature Singapore, pages 375-385.

[Yasunaga et al., 19] Yasunaga, M., Kasai, J., Zhang, R., Fabbri, A. R., Li, I., Friedman, D., & Radev, D. R., (2019). Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks: *In Proceedings of the AAAI conference on artificial intelligence,* 33(01), pages 7386-7393.

[Yun, S., et al., 2019] Yun, S., Jeong, M., Kim, R., Kang, J., & Kim, H. J., (2019). Graph transformer networks. *Advances in neural information processing systems*, 32.

[Zhang, H., Liu, X., and Zhang, J. 2022] Zhang, H., Liu, X., and Zhang, J., (2022). Hegel: Hypergraph transformer for long document summarization. *arXiv preprint arXiv:2210.04126*.

[Zhao, L., Xu, W., and Guo, J. 2020] Zhang, H., Liu, X., and Zhang, J., (2020). Improving abstractive dialogue summarization with graph structures and topic words. *In Proceedings of the 28th International Conference on Computational Linguistics,* pages 437-449.