Towards Two-Step Fine-Tuned Abstractive Summarization for Low-Resource Language Using Transformer T5

Salhazan Nasution[1], Ridi Ferdiana[2], Rudy Hartanto[3]
Department of Electrical and Information Engineering,
Faculty of Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia[1,2,3]
Department of Informatics Engineering-Faculty of Engineering, Universitas Riau, Pekanbaru, Indonesia[1]

***Abstract*—This study explores the potential of two-step fine-tuning for abstractive summarization in a low-resource language, focusing on Indonesian. Leveraging the Transformer-T5 model, the research investigates the impact of transfer learning across two tasks: machine translation and text summarization. Four configurations were evaluated, ranging from zero-shot to two-step fine-tuned models. The evaluation, conducted using the ROUGE metric, shows that the two-step fine-tuned model (T5-MT-SUM) achieved the best performance, with ROUGE-1: 0.7126, ROUGE-2: 0.6416, and ROUGE-L: 0.6816, outperforming all baselines. These findings demonstrate the effectiveness of task transferability in improving abstractive summarization performance for low-resource languages like Indonesian. This study provides a pathway for advancing natural language processing (NLP) in low-resource language through two-step transfer learning.**

Keywords*—*Abstractive summarization; low-resource language; Transformer T5; transfer learning

I. Introduction

The rapid advancement of technology has accelerated the flow of information, with an increasing number of people opting to consume information digitally rather than through printed media [1]. However, as the volume of information available on the Internet continues fto grow, humans face the challenge of processing and understanding this information within a limited amount of time, while their reading speed is inherently constrained to an average of 300 words per minute [2]. Automatic text summarization systems are, therefore, essential to facilitate faster information retrieval and comprehension.

Text summarization can be categorized into two types: manual summarization and automatic summarization. Manual summarization involves the direct effort of humans, which is inherently labor-intensive, time-consuming, and costly. Consequently, automation is required to produce summaries more efficiently and at a lower cost [3].

Text summarization is a method for generating concise, accurate, and digestible summaries from lengthy textual documents. This technique is commonly encountered in everyday life, such as in news headlines, meeting minutes, movie synopses, or book reviews. The objective of text summarization is to produce a summary from a collection of articles or documents that retains the essential information while being significantly shorter than the original text [4].

Text summarization is a subfield of Natural Language Processing (NLP) and typically employs two main approaches: extractive summarization and abstractive summarization. Extractive summarization involves selecting words, phrases, or sentences directly from the source document, ranking their relevance, and assembling them into a summary [5]. In contrast, abstractive summarization generates summaries by creating new sentences or phrases that convey the same meaning as the source document [6].

Abstractive summarization aids readers in better understanding an article because it generates a new summary by paraphrasing the content. This approach is considered the most ideal, as it holds significant potential for producing human-like summaries [7]. By rephrasing and condensing the source material, abstractive summarization achieves a level of coherence and fluency that extractive methods often lack, making it a highly valuable tool in text summarization research and applications.

Research on text summarization in Indonesia has been conducted across various domains. Studies employing extractive methods include summarization of news articles [8][9][10][11], snippets for search engine results [12], book synopsis summarization [13] using ranking methods such as Maximal Marginal Relevance (MMR) [14] and Text Rank [15], as well as meeting minute summarization [16][17]. In contrast, abstractive summarization research for the Indonesian language remains limited due to resource constraints. Indonesian is categorized as a low-resource language due to the limited availability of datasets [18]. Although Indonesian is recognized as the fourth most widely used language on the Internet, research progress in NLP for this language has been slow due to a lack of resources and datasets [19].

To date, there are only two large-scale datasets available for Indonesian text summarization: IndoSum [20], consisting of 19,000 articles and summaries, and Liputan6 [21], containing 200,000 articles with summaries. Combining these datasets can enhance the model's knowledge by exposing it to more data, patterns, and variations. However, dataset integration carries potential risks, particularly if one dataset is less accurate. Therefore, thorough pre-processing and in-depth data analysis of both datasets are necessary before merging them, along with cross-validation to ensure optimal results.

Research in the field of summarization has seen rapid advancements. Some studies have implemented Bidirectional

Gated Recurrent Unit (BiGRU) to generate summaries of Indonesian-language journals [22]. Other approaches include using Genetic Semantic Graphs [23], Abstract Meaning Representation (AMR) graphs [7], and Semantic Role Labeling (SLR) [24] for summarizing news articles. Wijayanti et al. [25] investigated the performance of pre-trained BERT models and evaluated them using the ROUGE metric [26]. The experimental results revealed that English pre-trained models could generate summaries from Indonesian articles. The utilization of Indonesian pre-trained models in conjunction with the IndoSum dataset, a benchmark for Indonesian text summarizing, produced more effective summaries. Nevertheless, issues such as fragmented sentences, erroneous phrases, and redundant vocabulary were still noted.

The efficacy of transfer learning provides a substantial benefit compared to earlier methodologies. Large Language Models (LLMs) like Transformer-T5 [27], pre-trained on vast datasets, may be tailored for many purposes, including machine translation and summarization. T5, leveraging its transfer learning capabilities, may be fine-tuned and retrained with more specialized datasets for various languages. This adaptability enables T5 to overcome the constraints in Indonesian text summarizing research by integrating machine translation and summarization, so enhancing the model's knowledge and ultimately elevating the accuracy and quality of Indonesian text summaries.

The rapid proliferation of digital content has highlighted the necessity for automatic text summary to enhance information accessibility. Summarization can be classified into extractive and abstractive methods. Extractive summarization directly selects sentences from the source text, whereas abstractive summarization rephrases and condenses the content, resulting in human-like summaries. In spite of its ability to produce coherent and meaningful summaries, abstractive summarization remains a difficult endeavor, particularly for low-resource languages such as Indonesian.

Existing research on Indonesian summarization is largely extractive due to the lack of large-scale, high-quality datasets required for abstractive models. However, advancements in Transformer-based models, particularly the Text-to-Text Transfer Transformer (T5), have introduced a new paradigm for natural language processing (NLP). T5 unifies multiple NLP tasks into a text-to-text framework, enabling it to perform tasks such as summarization and machine translation under the same architecture. Moreover, its transfer learning capabilities allow T5 to adapt to new tasks and languages with minimal data.

This study addresses the research gap in Indonesian abstractive summarization by evaluating the effectiveness of two-step fine-tuning. Using the T5 model, the research investigates the impact of first fine-tuning on machine translation and then on summarization. The evaluation, conducted using the ROUGE metric, provides insights into how task transferability enhances summarization performance in low-resource languages. This study contributes to advancing NLP for Indonesian by presenting an effective method to overcome data scarcity and achieve high-quality abstractive summarization.

The rest of this paper is organized as follows: Section II presents the motivating scenario, discussing the challenges of abstractive summarization in low-resource languages and the rationale behind using the Transformer-T5 model. Section III reviews related work, highlighting previous studies on Indonesian text summarization and transfer learning approaches. Section IV details the methodology, including the two-step fine-tuning approach, dataset descriptions, experimental design, and evaluation metrics. Section V presents the results, comparing different fine-tuning strategies and their impact on summarization performance. Section VI discusses the findings, analyzing the advantages of the proposed approach, the challenges encountered, and potential directions for future research. Finally, Section VII concludes the study by summarizing key insights and contributions.

II. Motivating Scenario

Abstractive summarization, especially in a low-resource language like Indonesian, requires a nuanced understanding of the language's syntax, semantics, and cultural context. The Transformer-T5 model, with its unified text-to-text approach, is well-suited for such tasks, but it must first acquire foundational linguistic knowledge in the target language to perform effectively.

For the T5 model to generate coherent and human-like abstractive summaries in Indonesian, it must be fine-tuned on datasets that introduce it to the language. This step is vital as the model's pre-training on general datasets may not include sufficient exposure to Indonesian. By enabling the model to learn the intricacies of Indonesian through a targeted fine-tuning process, we establish a critical foundation for downstream tasks like summarization.

The most efficient approach to instructing a large language model in a new language is to fine-tune it using a high-quality parallel dataset for machine translation. In particular, a parallel corpus of English and Indonesian can be a valuable source of grammatical structures, idiomatic expressions, and linguistic patterns which are diverse. This fine-tuning phase guarantees that the model can accurately comprehend Indonesian texts and establishes the foundation for abstractive summarization.

In this study, we implement a two-step fine-tuning strategy to reconcile the disparity between task-specific training and foundational language comprehension. Initially, the T5 model is refined on English-Indonesian machine translation duties to enhance its linguistic proficiency. The model is subsequently fine-tuned on an Indonesian abstractive summarization dataset to enhance its summarization capabilities. This sequential fine-tuning approach is intended to optimize the model's performance by utilizing transfer learning to overcome the obstacles presented by the low-resource nature of the Indonesian language.

III. Related Work

Recent studies have explored the use of pre-trained models for Indonesian text summarization as shown in Table I. The previous work [25] fine-tuned BertSumAbs with an Indonesian pre-trained BERT model on the IndoSum dataset, achieving ROUGE-1: 0.67, ROUGE-2: 0.54, and ROUGE-L: 0.65, outperforming English pre-trained BERT models. However, challenges such as meaningless words and repetitive phrases persist. Similarly, [36] examined IndoBERT checkpoints and

TABLE I. COMPARISON OF ROUGE SCORES ACROSS MODELS

No.	Research	Model	Dataset	R-1	R-2	R-L
1	[28]	BART (Augmented)	Liputan6	0.4093	0.3409	0.4037
2	[29]	IndoBERT	Liputan6	0.4235	0.2415	0.3544
3	[30]	ALBERT	IndoSum	0.4528	0.4077	0.4439
4	[31]	EASum (IndoBART + Extractive)	IndoSum	0.3600	0.2300	0.3500
5	[32]	IndoBART (Augmented)	Liputan6	0.3822	0.2079	0.3124
6	[33]	mT5 (Augmented)	Liputan6	0.3856	0.2329	0.3263
7	[34]	BERT2GPT	IndoSum	0.6226	0.5613	0.6043
8	[35]	T5-Large	Wikipedia	0.4738	0.2927	0.4115
9	[36]	BERTSumAbs + IndoBERT-LEM + GPT Decoder	IndoSum	0.6920	0.6135	0.6836
10	[37]	CTRLSum (mBART50)	Liputan6	0.4341	0.2406	0.3963
11	[38]	Transformer (CLS + MWE)	IndoSum	0.4247	0.2226	0.3809
12	[25]	BertSumAbs (IndoBERT)	IndoSum	0.6700	0.5400	0.6500
13	**Our Research**	Transformer-T5 (two-step transfer learning)	IndoSum	**0.7126**	**0.6416**	**0.6816**

found that BERTSumAbs with IndoBERT-LEM and a GPT-like decoder performed best, achieving ROUGE-1: 0.6920, ROUGE-2: 0.6135, and ROUGE-L: 0.6836, highlighting the importance of decoder selection in transformer-based summarization. Beyond BERT models, [30] explored ALBERT, a lightweight variant of BERT, fine-tuning it on IndoSum. The best model achieved ROUGE-1: 0.4528, ROUGE-2: 0.4077, and ROUGE-L: 0.4439, demonstrating a viable alternative for resource-constrained environments while maintaining competitive accuracy.

Some studies have combined extractive and abstractive approaches to improve summarization performance. The author in [31] introduced EASum, a model that first generates extractive summaries using Doc2Vec and cosine similarity, then refines them using IndoBART. On IndoSum, it achieved ROUGE-1: 0.36, ROUGE-2: 0.23, and ROUGE-L: 0.35, while on Liputan6, it performed lower than mBART and IndoGPT. A different hybrid approach was explored in [34], which combined BERT as an encoder with GPT-2 as a decoder. The model, trained on IndoSum, achieved ROUGE-1: 0.6226, ROUGE-2: 0.5613, and ROUGE-L: 0.6043, outperforming BertSum-based models in bi-gram accuracy.

Data augmentation techniques have been widely applied to improve model generalization. The author in [32] utilized synonym replacement-based augmentation and fine-tuning on Liputan6, where the best-performing IndoBART model achieved ROUGE-1: 0.3822, ROUGE-2: 0.2079, and ROUGE-L: 0.3124. Similarly, [33] applied backtranslation-based augmentation on mT5, achieving ROUGE-1: 0.3856, ROUGE-2: 0.2329, and ROUGE-L: 0.3263, demonstrating improvements in coherence and informativeness. Another augmentation-based study, [28], fine-tuned BART on Liputan6, IndoSum, and ChatGPT-augmented summaries, generating over 36,000 new instances. The best model reached ROUGE-1: 0.4093, ROUGE-2: 0.3409, and ROUGE-L: 0.4037, highlighting the effectiveness of multi-dataset fine-tuning.

Cross-lingual and multilingual approaches have also been investigated to expand summarization capabilities. The author in [37] introduced CTRLSum, a keyword-controlled summarization method fine-tuned on mBART50, which achieved ROUGE-1: 0.4341, ROUGE-2: 0.2406, and ROUGE-L: 0.3963 on Liputan6. For cross-lingual summarization (CLS), [38] proposed an end-to-end CLS model integrating multilingual word embeddings (MWE). The model, trained on a translated IndoSum dataset, achieved ROUGE-1: 0.4247, ROUGE-2: 0.2226, and ROUGE-L: 0.3809, improving cross-lingual representation alignment.

Large language models (LLMs) have recently been explored as few-shot summarization tools. The author in [29] evaluated ChatGPT with prompt tuning on Liputan6, comparing it with IndoBART, IndoBERT, and mBART50. While IndoBERT performed best on Liputan6 (ROUGE-1: 0.4235, ROUGE-2: 0.2415, ROUGE-L: 0.3544), ChatGPT had lower automatic evaluation scores but excelled in human evaluation, suggesting superior fluency but lower adherence to reference summaries. Benchmarking efforts have assessed various transformer-based models. The author in [35] compared T5-Large, Pegasus-XSum, and ProphetNet-CNNDM on Wikipedia datasets in English and Indonesian, finding that T5-Large achieved the best ROUGE scores (ROUGE-1: 0.4738, ROUGE-2: 0.2927, ROUGE-L: 0.4115).

Despite the advancements in Indonesian abstractive summarization, most studies either focus on direct fine-tuning of transformer-based models or augmenting datasets to improve performance. However, few works explore multi-task learning approaches that leverage machine translation as a pre-training step for summarization. Existing studies, such as those using IndoBART, mT5, and BERT2GPT, demonstrate the effectiveness of pre-trained models but primarily fine-tune them only on summarization tasks.

This paper addresses this gap by proposing a two-step fine-tuning approach where a model is first fine-tuned on a machine translation task, followed by fine-tuning for abstractive summarization. The hypothesis is that exposing the model to translation tasks helps it learn better language representations and text reformation techniques, which could lead to more coherent, fluent, and information-rich summaries. Unlike previous works that rely solely on pre-trained language models or data augmentation, this research explores whether a structured pre-training strategy with translation improves the performance of Indonesian abstractive summarization models.

IV. METHODS

A. Text-to-Text Transfer Transformer (T5)

The Text-to-Text Transfer Transformer (T5) is a cutting-edge natural language processing (NLP) model developed by Google Research. It is structured to encapsulate all NLP tasks within a cohesive text-to-text format, wherein both the input and output are expressed as sequences of text. This approach enables T5 to do several tasks, including translation, summarization, categorization, and question answering, with a consistent architecture and training methodology.

- Text-to-Text Format: T5 differentiates itself from conventional models, which are task-specific (such as categorization or generation), by reinterpreting every activity as a text creation challenge. For instance:
    - Translation: Input = “translate English to Indonesian: Where are you?”, Output = “Di mana kamu?”.
    - Summarization: Input = “summarize: The article discusses...”, Output = “Key points of the article...”.
    - Classification: Input = “classify sentiment: This movie was amazing!”, Output = “positive”.
- Unified Architecture: T5 utilizes the Transformer architecture, a deep learning model based on self-attention mechanisms. It employs an encoder-decoder design:
    - The encoder processes the input text to generate contextual representations.
    - The decoder generates the output text token by token.
- Pre-Training with Transfer Learning: T5 is pre-trained on a large corpus using a task called span corruption, where spans of text in the input are masked, and the model is trained to predict the masked spans. This approach helps the model learn a rich understanding of language, which can then be fine-tuned for specific downstream tasks.
- Scalability: T5 comes in various sizes, from small to large (e.g. T5-Small, T5-base, T5-Large, T5-XXL), allowing flexibility depending on computational resources and task complexity.
- Fine-Tuning: After pre-training, T5 can be fine-tuned on specific tasks with labeled data. This step ensures that the model learns task-specific nuances while leveraging the general language knowledge from pre-training.

The versatility of T5 makes it particularly suitable for tasks like summarization and machine translation, as it treats both tasks uniformly in the text-to-text format. This study leverages T5’s capabilities in a multi-stage process: first fine-tuning it on machine translation tasks to leverage its language understanding capabilities, followed by additional fine-tuning on summarization tasks to optimize its performance for generating concise and coherent summaries. The model’s capacity to generalize across tasks facilitates zero-shot testing, permitting assessment without task-specific fine-tuning.

B. Datasets

1) Summarization task: The dataset employed in this study is IndoSum [20]. IndoSum establishes a novel standard for text summarizing in Bahasa Indonesia. The IndoSum dataset comprises articles obtained from Indonesian online news outlets and contains almost 200 times the number of articles compared to prior research on Indonesian-language articles [39]. To be more precise, IndoSum includes 18,762 articles and their respective summaries, which are divided into six categories: entertainment, inspiration, sports, celebrity, headlines, and technology.

A number of well-known online news portals, including CNN Indonesia, Kumparan, Suara.com, Antaranews, and others, provided the data for IndoSum. The mean article length in this dataset is 292 words, with the biggest article being 1,228 words and the shortest item consisting of 36 words. The summaries average 58 words in length, with the greatest being 86 words and the smallest including 32 words.

2) Machine translation task: Datasets with parallel English-Indonesian sentences were utilized for pre-training, including OpenSubtitle [40], TED 2018 [41], TED 2020 [42], News Commentary [43], and CCMatrix [44]. These datasets were chosen especially for the purpose of training the T5 model for translating from English to Indonesian because of their complimentary qualities and applicability to the field of text translation.

The OpenSubtitle dataset, comprising a compilation of movie and television subtitles, offers a valuable repository of informal, conversational language. This dataset is very beneficial for training models to process daily language, colloquialisms, and conversational writing. The extensive range of themes and contexts guarantees that the model is exposed to multiple linguistic styles and structures.

The TED 2018 and TED 2020 datasets are composed of transcripts from TED Talks, which are renowned for their formal and informative communication style. Ideally, these datasets are utilized to train models that will translate academic, technical, and professional content. Additionally, they assist the model in dealing with a variety of subjects, as TED Talks encompass a broad spectrum of fields, including technology, science, and the arts and culture.

The News Commentary dataset emphasizes news stories and commentary, with a formal tone and terminology typically seen in journalistic literature. This dataset enables the model to manage domain-specific language, organized arguments, and subjective content characteristic of news sources.

Finally, the CCMatrix dataset constitutes a substantial compilation of parallel sentences derived from web-crawled data. The extensive size and varied content provide thorough coverage of linguistic structures and terminology, rendering it essential for enhancing the model’s generalization capacity across distinct text kinds. For the CCMatrix dataset, we utilized just 600,000 data points due to constraints of computational resources and processing duration.

We sought to establish a strong and adaptable training basis for the T5 model by integrating these datasets. This selection guarantees the model’s exposure to a wide array of text styles, tones, and domains, facilitating its effective performance in multiple translation scenarios, encompassing informal, formal, and technical contexts. The meticulous selection of datasets demonstrates our aim to develop a translation model that can tackle the intricacies and diversity present in actual English-to-Indonesian translation projects.

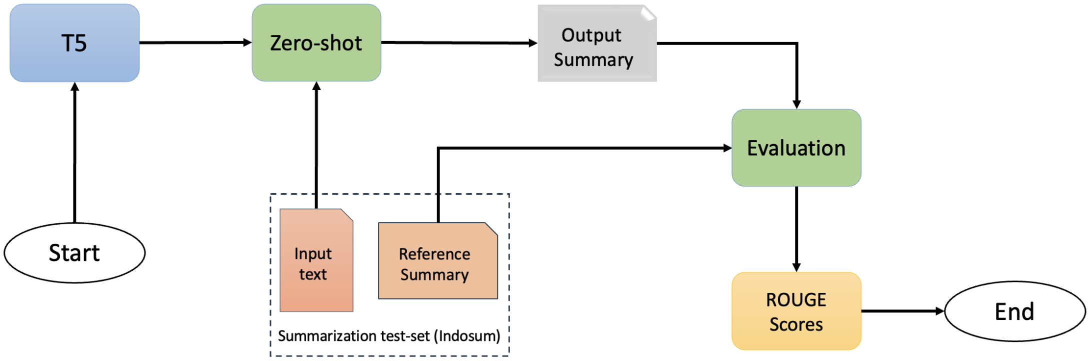

Fig. 1. T5 (Zero-shot).

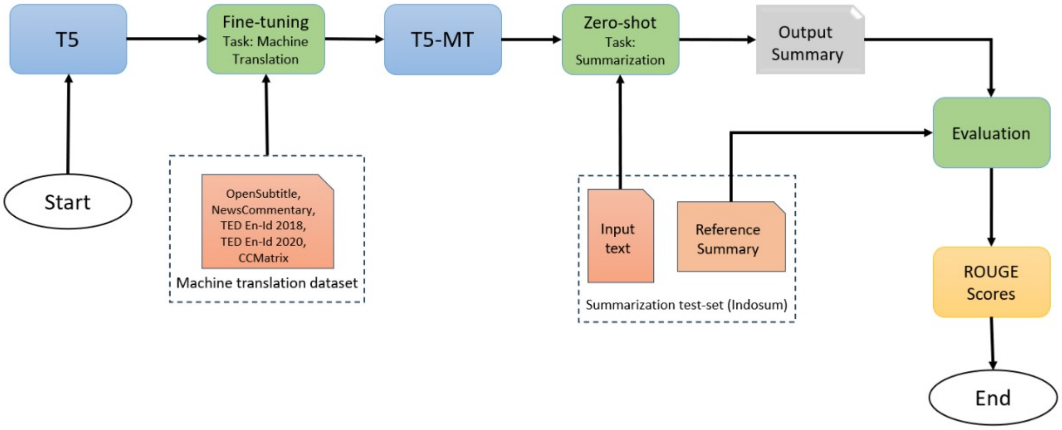
T5
Fine-tuning
Task: Machine Translation
T5-MT
Zero-shot
Task: Summarization
Output Summary
Evaluation
Start
OpenSubtitle, NewsCommentary, TED En-Id 2018, TED En-Id 2020, CCMatrix
Machine translation dataset
Input text
Reference Summary
Summarization test-set (Indosum)
ROUGE Scores
End

Fig. 2. T5-MT (Fine-tuning).

C. *Experimental Design*

A two-step fine-tuning approach is implemented in this study to assess the efficacy of the T5 model on Indonesian abstractive summarization. Four experimental configurations are included in the design, which has two tasks: machine translation and summarization.

1) Experiment 1: T5 Zero-shot: Fig. 1 illustrates the workflow for a zero-shot summarization task that employs the T5 (Text-to-Text Transfer Transformer) model. Starting with the pre-trained T5 model, which has not been refined on the specific summarization dataset, the procedure commences. The zero-shot approach involves explicitly testing the model on summarization tasks without any additional training. This demonstrates the T5 model's ability to generalize to unseen tasks by utilizing its pre-trained knowledge.

The summarization task employs the IndoSum dataset, comprising two primary elements: input text (the texts designated for summarization) and reference summaries (the authoritative summaries utilized for assessment). The T5 model analyzes the input text and produces a summary, which is subsequently compared to the reference summary to assess its quality.

The assessment procedure employs the ROUGE metric, which quantifies the overlap between the generated summaries and the reference summaries based on unigrams (ROUGE-1), bigrams (ROUGE-2), and the longest common subsequences (ROUGE-L). The resultant ROUGE scores quantify the T5 model's performance on the summarization job. The workflow concludes with the reporting of the ROUGE scores as the final result. This picture clearly delineates the process from job start to evaluation, highlighting the zero-shot capabilities of T5 in summarization tasks.

2) Experiment 2: T5-MT Fine-tuning: The T5 model is employed to perform two sequential tasks: fine-tuning for machine translation and zero-shot summarization. The workflow is illustrated in Fig. 2. It initiates with the T5-base model, which is refined on a machine translation task. This fine-tuning process uses diverse datasets such as OpenSubtitles, NewsCommentary, TED En-Id 2018, TED En-Id 2020, and CCMatrix, which consist of parallel text pairs for English-to-Indonesian translation. The outcome of this step is a specialized T5 model for machine translation, labeled as T5-MT.

The fine-tuned T5-MT model is then tested on a summarization task using the IndoSum dataset in a zero-shot manner, meaning the model is applied to summarization without additional task-specific fine-tuning. The summarization dataset includes input text (the articles to be summarized) and cor-

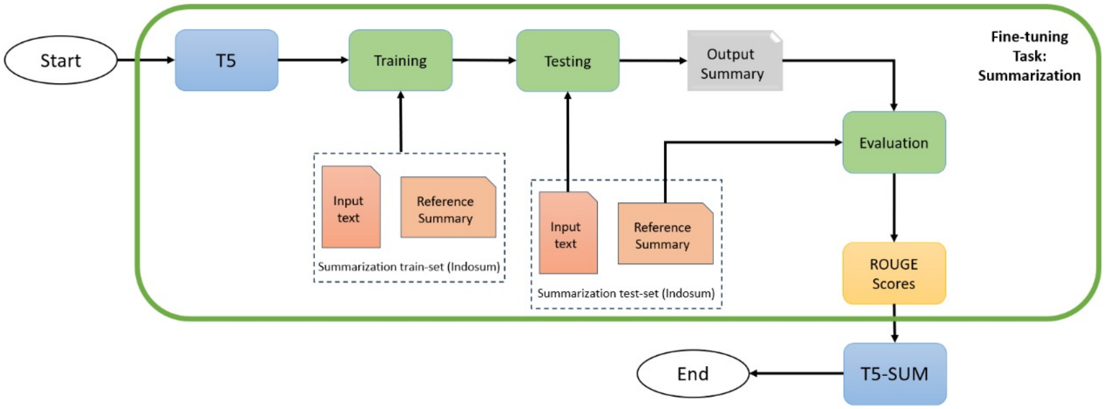

Fig. 3. T5-SUM (Fine-tuning).

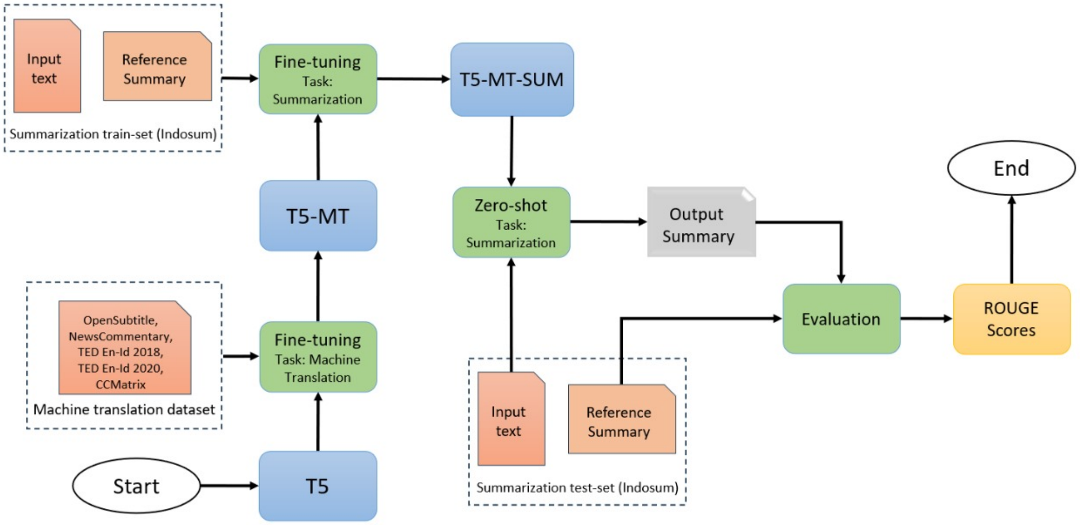
Input text
Reference Summary
Summarization train-set (Indosum)
Fine-tuning Task: Summarization
T5-MT-SUM
End
T5-MT
Zero-shot Task: Summarization
Output Summary
OpenSubtitle, NewsCommentary, TED En-Id 2018, TED En-Id 2020, CCMatrix
Machine translation dataset
Fine-tuning Task: Machine Translation
Evaluation
ROUGE Scores
Input text
Reference Summary
Summarization test-set (Indosum)
Start
T5

Fig. 4. T5-MT-SUM (Two-step Fine-tuning).

responding reference summaries (gold-standard summaries). The model processes the input text and generates an output summary.

The generated summaries are evaluated using the ROUGE metric, which computes the similarity between the output summaries and the reference summaries. The evaluation provides scores for ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and ROUGE-L (longest common subsequence). These scores serve as a measure of the summarization performance. The workflow concludes with the ROUGE scores summarizing the model's effectiveness. This diagram showcases a pipeline approach, where a model fine-tuned for one task (machine translation) is tested for generalization in another task (summarization), illustrating the versatility of T5.

3) Experiment 3: T5-SUM Fine-tuning: Fig. 3 diagram represents the workflow for fine-tuning the T5 model specifically for a summarization task. It begins with the pre-trained T5 model, which undergoes fine-tuning using the IndoSum dataset. The dataset contains input text (articles or passages to be summarized) and their corresponding reference summaries (gold-standard summaries). The fine-tuning process adapts the T5 model to perform summarization tasks effectively, creating a specialized model referred to as T5-SUM.

The process is divided into two main phases: training and testing. During the training phase, the model learns patterns and relationships between the input text and the reference summaries from the training set of the IndoSum dataset. Once trained, the model is tested using the test set of the same dataset, where it generates summaries (labeled as Output Summary) for new, unseen input texts.

The ROUGE metric is employed to assess the summaries that are generated, comparing the overlap between the summaries and the reference summaries. The model is evaluated using ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and ROUGE-L (longest common subsequences). The process culminates in the ROUGE scores, which serve as an indicator of the summarization task's performance of the fine-tuned T5-SUM model. This diagram effectively emphasizes the significance of dataset preparation and evaluation metrics by illustrating a structured pipeline for training and evaluating a summarization model.

4) Experiment 4: T5-MT-SUM Two-Step Fine-tuning: Fig. 4 shows a two-step fine-tuning approach for modifying the T5 model to perform summarization tasks. It starts with the pre-trained T5 model, which is then fine-tuned on a machine translation task with datasets including OpenSubtitles, NewsCommentary, TED En-Id 2018, TED En-Id 2020, and CCMatrix. This step results in the T5-MT model, which is designed for machine translation.

During the second phase, the T5-MT model is subjected to an additional fine-tuning process for the summarizing job utilizing the IndoSum dataset, comprising input text and their respective reference summaries. Through further refinement, the T5-MT model is modified for summarizing tasks, resulting in the final model known as T5-MT-SUM. This two-step fine-tuning ensures that the model takes advantage of its machine translation knowledge before being improved for summarization, potentially enhancing performance.

The T5-MT-SUM model is subsequently evaluated on the IndoSum test set, producing output summaries for novel input texts in a zero-shot summarizing task. The produced summaries are assessed against the reference summaries utilizing the ROUGE metric, which measures summary quality through unigram, bigram, and longest common subsequence overlaps (ROUGE-1, ROUGE-2, and ROUGE-L, respectively). The procedure culminates with the ROUGE scores, which encapsulate the model's efficacy in the summarization task. This figure clearly illustrates the sequential fine-tuning method, demonstrating how knowledge transfer between tasks can improve performance.

D. Evaluation Metrics

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a metric that is intended to assess the quality of summaries by corresponding them to one or more human-generated reference summaries. It measures the overlap between the candidate summary and reference summaries in terms of n-grams, sequences, or word pairs.

- ROUGE-1: Measures unigram overlap.
- ROUGE-2: Measures bigram overlap.
- ROUGE-L: Measures the longest common subsequence between generated and reference summaries.

ROUGE-N evaluates the n-gram overlap between a candidate summary and reference summaries. It is defined as:

$$\text{ROUGE-N} = \frac{\sum_{S\in\text{Reference Summaries}} \sum_{\text{gram}_n\in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S\in\text{Reference Summaries}} \sum_{\text{gram}_n\in S} \text{Count}(\text{gram}_n)} \tag{1}$$

Where:

- $\text{Count}_{\text{match}}(\text{gram}_n)$: The number of n-grams in both the candidate summary and the reference summary.
- $\text{Count}(\text{gram}_n)$: The total number of n-grams in the reference summary.
- n: The length of the n-gram (e.g. 1 for unigrams, 2 for bigrams).

1) ROUGE-1 (Unigram overlap): ROUGE-1 measures the overlap of unigrams between the reference and candidate summaries. It is defined as:

$$\text{ROUGE-1} = \frac{\text{Number of Overlapping Unigrams}}{\text{Total Number of Unigrams in the Reference Summary}} \tag{2}$$

Example:

- Reference Summary: "I really like reading novels".
- Candidate Summary: "I like reading novels every day".
- Unigrams in Reference: I, really, like, reading, novels.
- Unigrams in Candidate: I, like, reading, novels, every, day.
- Overlapping Unigrams: I, like, reading, novels (4 overlaps).

The total number of unigrams in the reference summary is 5. Thus:

$$\text{ROUGE-1} = \frac{4}{5} = 0.8 \tag{3}$$

2) ROUGE-2 (Bigram Overlap): ROUGE-2 measures the overlap of bigrams (pairs of consecutive words) between the reference and candidate summaries. It is defined as:

$$\text{ROUGE-2} = \frac{\text{Number of Overlapping Bigrams}}{\text{Total Number of Bigrams in the Reference Summary}} \tag{4}$$

Example:

- Reference Bigrams: I really, really like, like reading, reading novels.
- Candidate Bigrams: I like, like reading, reading novels, novels every, every day.
- Overlapping Bigrams: like reading, reading novels (2 overlaps).

The total number of bigrams in the reference summary is 4. Thus:

$$\text{ROUGE-2} = \frac{2}{4} = 0.5 \tag{5}$$

3) ROUGE-L: ROUGE-L measures the longest common subsequence (LCS) between a candidate summary and a reference summary, capturing sentence-level structure. The precision, recall, and F-measure based on LCS are defined as:

$$P_{\text{LCS}} = \frac{\text{LCS}(X, Y)}{|Y|} \tag{6}$$

$$R_{\text{LCS}} = \frac{\text{LCS}(X, Y)}{|X|} \tag{7}$$

$$F_{\text{LCS}} = \frac{(1 + \beta^2) \cdot P_{\text{LCS}} \cdot R_{\text{LCS}}}{\beta^2 \cdot P_{\text{LCS}} + R_{\text{LCS}}} \tag{8}$$

Where:

- $\text{LCS}(X, Y)$: The length of the longest common subsequence between the reference summary X and the candidate summary Y.
- $|X|$: The length of the reference summary.
- $|Y|$: The length of the candidate summary.
- β: A weighting parameter, often set to 1 to equally weight precision and recall.

Example:

- Reference Summary: “I really like reading novels”.
- Candidate Summary: “I like reading novels every day”.
- LCS: “I like reading novels” (4 words).
- Length of Candidate: 6, Length of Reference: 5.

Calculate precision and recall:

$$P_{\text{LCS}} = \frac{4}{6} = 0.6667, \quad R_{\text{LCS}} = \frac{4}{5} = 0.8 \tag{9}$$

Calculate F-measure with $\beta = 1$:

$$F_{\text{LCS}} = \frac{(1 + 1^2) \cdot 0.6667 \cdot 0.8}{1^2 \cdot 0.6667 + 0.8} = 0.7273 \tag{10}$$

V. Results

The ablation study was conducted to evaluate the impact of fine-tuning at different stages on the Transformer-T5 model’s performance in Indonesian abstractive summarization. Four configurations of the model were tested, as illustrated in Fig. 5, with the evaluation focusing on ROUGE-1, ROUGE-2, and ROUGE-L scores.

1) T5-base (Zero-shot): ROUGE-1: 0.4458, ROUGE-2: 0.3381, ROUGE-L: 0.3961. This configuration represents the baseline performance of the pre-trained T5 model without task-specific fine-tuning. The scores reflect the limited effectiveness of the zero-shot approach for a low-resource language like Indonesian, where the model has no prior exposure to Indonesian summarization or translation tasks.

2) T5-MT (Machine Translation fine-tuning): ROUGE-1: 0.6224, ROUGE-2: 0.5220, ROUGE-L: 0.5793. This model was fine-tuned on a machine translation task using the CCMatrix dataset (300K parallel sentences). The improvement over the zero-shot baseline demonstrates the transferability of knowledge learned in machine translation to abstractive summarization.

3) T5-SUM (Summarization fine-tuning): ROUGE-1: 0.7106, ROUGE-2: 0.6393, ROUGE-L: 0.6790. Fine-tuning directly on the IndoSum dataset for summarization significantly enhanced performance compared to T5-MT. This result indicates that task-specific training is critical for improving the model’s summarization capabilities.

4) T5-MT-SUM (Two-Step fine-tuning): In this configuration, the model was first fine-tuned on the machine translation task (OpenSubtitle, TED 2018, TED 2020, NewsCommentary, CCMatrix 600K) and subsequently fine-tuned on the summarization task (IndoSum). This two-step approach achieved the best performance with ROUGE-1: 0.7126, ROUGE-2: 0.6416, and ROUGE-L: 0.6816, underscoring the benefits of sequential task transferability.

VI. Discussion

The ablation study was designed to investigate how different fine-tuning strategies affect the T5 model’s performance on Indonesian abstractive summarization. By systematically varying the stages of fine-tuning, the study aimed to uncover the relative contributions of machine translation and task-specific summarization fine-tuning to the final performance.

A. Ablation Study

1) Importance of related tasks: Fine-tuning on machine translation (T5-MT) significantly boosted performance compared to the zero-shot baseline. The ROUGE-1 improvement from 0.4458 to 0.6224 suggests that the knowledge gained in machine translation helps the model understand Indonesian syntax and semantics better.

2) Critical role of task-specific training: Fine-tuning directly on the summarization task (T5-SUM) led to further substantial gains, with ROUGE-1 reaching 0.7106. This highlights that task-specific training is essential for generating coherent and meaningful summaries.

3) Advantages of two-step fine-tuning: The incremental improvement of T5-MT-SUM over T5-SUM (ROUGE-1: 0.7126 vs. 0.7106) demonstrates the synergy of combining task-specific fine-tuning with knowledge from a related task like machine translation. While the improvement is small, it underscores the potential of two-step fine-tuning as a systematic approach to boosting model performance.

4) Implications for low-resource summarization: The findings from this ablation study provide valuable insights for Summarization research in low-resource languages:

- Leveraging Related Tasks: Transfer learning from machine translation to summarization is an effective strategy to overcome data scarcity in low-resource settings.
- Sequential Fine-Tuning: Two-step fine-tuning enhances the adaptability of pre-trained models by progressively refining their knowledge through related tasks.

B. Comparison with Existing Summarization Models

The current paper introduces a two-step transfer learning approach using Transformer-T5, where the model is first fine-tuned on a machine translation task before being fine-tuned for abstractive summarization on the IndoSum dataset. This structured pre-training strategy demonstrates significant improvements over previous works as shown in Table I. In

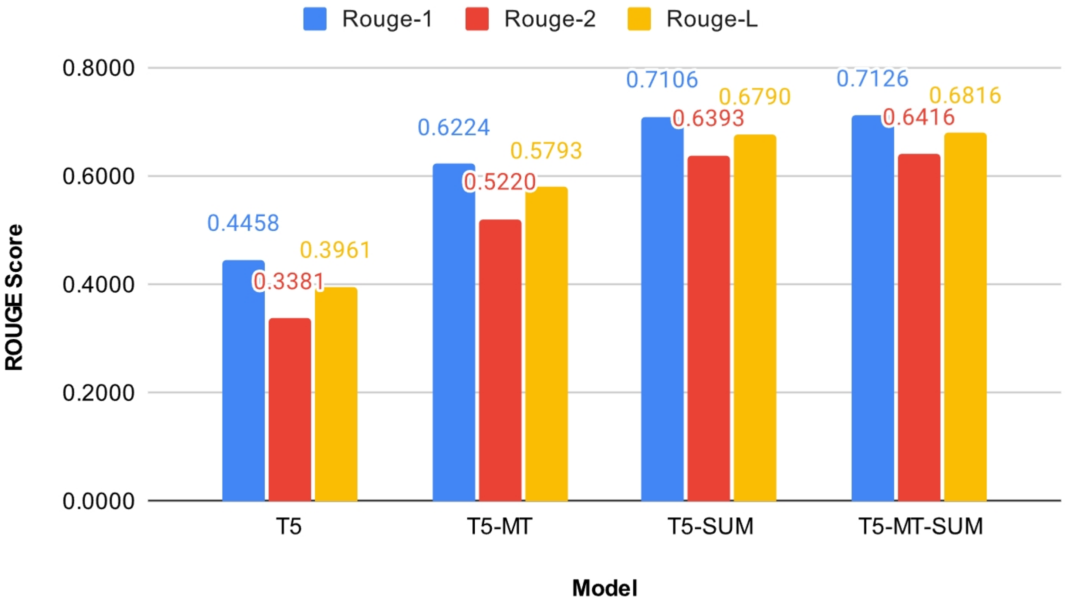

Fig. 5. Experiment result for two-step fine-tuned abstractive summarization.

terms of performance, the proposed model achieves ROUGE-1: 0.7126, ROUGE-2: 0.6416, and ROUGE-3: 0.6816, making it the highest-performing model across all three metrics. It surpasses the previous best result from BERTSum-Abs + IndoBERT-LEM + GPT Decoder (ROUGE-1: 0.6920, ROUGE-2: 0.6135, ROUGE-3: 0.6836), as well as BERT2GPT (ROUGE-1: 0.6226, ROUGE-2: 0.5613, ROUGE-3: 0.6043), demonstrating the effectiveness of combining translation-based pre-training with summarization fine-tuning.

Compared to data augmentation-based models, such as BART (Augmented) and mT5 (Augmented), which improved summarization performance but remained below ROUGE-1: 0.41, the proposed two-step approach proves to be more effective. Instead of solely relying on augmented training data, this study shows that a structured pre-training approach enhances the model's ability to generate coherent and fluent summaries. Additionally, the study outperforms multilingual and cross-lingual models, such as CTRLSum and Transformer (CLS + MWE), which achieved lower ROUGE scores despite their ability to generate summaries in multiple languages. The findings suggest that a monolingual, targeted approach with pre-training on a related NLP task (machine translation) yields better results than cross-lingual training.

Overall, this research sets a new benchmark for Indonesian abstractive summarization, showing that fine-tuning on machine translation before summarization enhances summary coherence and quality. The structured two-step learning strategy proves more effective than direct fine-tuning, data augmentation, and hybrid extractive-abstractive methods, establishing Transformer-T5 (two-step transfer learning) as the best-performing model for Indonesian text summarization compared to previous work.

C. Challenges and Limitations

Despite the promising results, the study faced several challenges:

- Marginal gains: The improvement from T5-SUM to T5-MT-SUM, while consistent, was relatively small. This suggests diminishing returns with additional fine-tuning stages and calls for further exploration of factors such as dataset size and quality.
- Evaluation metrics: ROUGE scores, while widely used, have limitations in capturing summary coherence and informativeness. Incorporating human evaluations could provide a more comprehensive assessment of summary quality.

D. Future Directions

This study paves the way for several promising research directions. First, future work could explore optimizing the two-step fine-tuning process by refining dataset selection or introducing pretraining techniques tailored to low-resource languages. Incorporating larger or more diverse datasets could further improve the model's ability to generalize across different summarization scenarios.

Second, qualitative evaluation of generated summaries through human assessments should be conducted to complement the ROUGE metric. This would provide deeper insights into summary coherence, readability, and informativeness, which are crucial for practical applications.

Third, expanding the approach to multilingual fine-tuning could assess its generalizability across other low-resource languages. The combination of cross-lingual transfer learning and domain-specific training could offer a robust framework for NLP applications in diverse linguistic contexts.

Finally, future research could explore adapting the two-step fine-tuning methodology for other NLP tasks, such as question answering, sentiment analysis, or content generation. These extensions would highlight the broader applicability of this approach and contribute to advancing natural language processing in low-resource settings.

VII. Conclusions

This research explored the effectiveness of a two-step fine-tuning approach using the Transformer-T5 model for abstractive text summarization in Bahasa Indonesia, a low-resource language. Four configurations were evaluated: zero-shot (T5-Base), single-task fine-tuning for machine translation (T5-MT), single-task fine-tuning for summarization (T5-SUM), and two-step fine-tuning combining machine translation and summarization tasks (T5-MT-SUM). The two-step fine-tuning approach achieved the best performance, with ROUGE-1: 0.7126, ROUGE-2: 0.6416, and ROUGE-L: 0.6816, demonstrating the benefits of task transferability in improving summarization quality.

The findings indicate that leveraging knowledge from a related task, such as machine translation, can enhance the performance of abstractive summarization models. Task-specific fine-tuning was also shown to play a critical role in generating coherent and accurate summaries. The results validate the potential of transfer learning to overcome the challenges posed by data scarcity in low-resource languages like Bahasa Indonesia.

Acknowledgment

This research is supported by Universitas Gadjah Mada.

References

[1] A. Mitchell, J. H. Holcomb, and R. Weisel, "State of the News Media 2016," *WD info*, p. 118, 2016. [Online]. Available: http://www.journalism.org/2016/06/15/state-of-the-news-media-2016/#

[2] S. Primativo, D. Spinelli, P. Zoccolotti, M. De Luca, and M. Martelli, "Perceptual and Cognitive Factors Imposing "Speed Limits" on reading rate: A study with the rapid serial visual presentation," *PLoS ONE*, vol. 11, no. 4, pp. 1–25, 2016.

[3] O. Klymenko, D. Braun, and F. Matthes, "Automatic text summarization: A state-of-the-art review," *ICEIS 2020 - Proceedings of the 22nd International Conference on Enterprise Information Systems*, vol. 1, no. Iceis, pp. 648–655, 2020.

[4] G. Sharma and D. Sharma, "Automatic Text Summarization Methods: A Comprehensive Review," *SN Computer Science*, vol. 4, no. 1, pp. 1–18, 2023. [Online]. Available: https://doi.org/10.1007/s42979-022-01446-w

[5] Y. J. Kumar, O. S. Goh, H. Basiron, N. H. Choon, and P. C. Suppiah, "A review on automatic text summarization approaches," pp. 178–190, apr 2016. [Online]. Available: https://thescipub.com/abstract/jcssp.2016.178.190

[6] A. Khan, N. Salim, and H. Farman, "Clustered genetic semantic graph approach for multi-document abstractive summarization," *2016 International Conference on Intelligent Systems Engineering, ICISE 2016*, vol. 77, no. 18, pp. 63–70, 2016.

[7] V. Severina and M. L. Khodra, "Multidocument Abstractive Summarization using Abstract Meaning Representation for Indonesian Language," in *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*. IEEE, sep 2019, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8904449/

[8] I. R. Musyaffanto, G. Budi Herwanto, and M. Riasetiawan, "Automatic extractive text summarization for indonesian news articles using maximal marginal relevance and non-negative matrix factorization," *Proceedings - 2019 5th International Conference on Science and Technology, ICST 2019*, pp. 1–6, 2019.

[9] D. Annisa and M. L. Khodra, "Query-based summarization for Indonesian news articles," *Proceedings - 2017 International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2017*, 2017.

[10] R. Reztaputra and M. L. Khodra, "Sentence structure-based summarization for Indonesian news articles," *Proceedings - 2017 International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2017*, pp. 0–5, 2017.

[11] R. B. Hutama, A. R. Barakbah, and A. Helen, "Indonesian news auto summarization in infrastructure development topic using 5W+1H consideration," *Proceedings - International Electronics Symposium on Knowledge Creation and Intelligent Computing, IES-KCIC 2017*, vol. 2017-Janua, pp. 258–264, 2017.

[12] N. F. Saraswati, Indriati, and R. S. Perdana, "Peringkasan Teks Otomatis Menggunakan Metode Maximum Marginal Relevance Pada Hasil Pencarian Sistem Temu Kembali Informasi Untuk Artikel Berbahasa Indonesia," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, vol. 2, no. 11, pp. 5494–5502, 2018.

[13] A. Indriani, "Maximum Marginal Relevance Untuk Peringkasan Teks Otomatis Sinopsis Buku Berbahasa Indonesia," *SEMNASTEKNOMEDIA ONLINE*, vol. 2, no. 1, pp. 3–5, 2014.

[14] J. Carbinell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," *ACM SIGIR Forum*, vol. 51, no. 2, pp. 209–210, aug 2017. [Online]. Available: https://doi.org/10.1145/3130348.3130369https://dl.acm.org/doi/10.1145/3130348.3130369

[15] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

[16] M. T. Yulyanto and M. L. Khodra, "Automatic extractive summarization on Indonesian parliamentary meeting minutes," *Proceedings - 2017 International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2017*, 2017.

[17] G. H. Rachman and M. L. Khodra, "Automatic rhetorical sentence categorization on Indonesian meeting minutes," in *Proceedings of 2016 International Conference on Data and Software Engineering, ICoDSE 2016*. IEEE, oct 2017, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/7936103/

[18] C. D. F. M. Paul Lewis, Gary F. Simons, *Ethnologue: Languages of the World*, 18th ed. Dallas, TX: SIL International, 2015.

[19] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and A. Purwarianti, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," pp. 843–857, sep 2020. [Online]. Available: http://arxiv.org/abs/2009.05387

[20] K. Kurniawan and S. Louvan, "IndoSum: A New Benchmark Dataset for Indonesian Text Summarization," in *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*. IEEE, nov 2019, pp. 215–220. [Online]. Available: https://ieeexplore.ieee.org/document/8629109/

[21] F. Koto, J. H. Lau, and T. Baldwin, "Liputan6: A Large-scale Indonesian Dataset for Text Summarization," no. 1, pp. 598–608, nov 2020. [Online]. Available: http://arxiv.org/abs/2011.00679https://aclanthology.org/2020.aacl-main.60/

[22] R. Adelia, S. Suyanto, and U. N. Wisesty, "Indonesian Abstractive Text Summarization Using Bidirectional Gated Recurrent Unit," *Procedia Computer Science*, vol. 157, pp. 581–588, 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050919311755

[23] R. S. Devianti and M. L. Khodra, "Abstractive Summarization using Genetic Semantic Graph for Indonesian News Articles," *Proceedings - 2019 International Conference on Advanced Informatics: Concepts, Theory, and Applications, ICAICTA 2019*, 2019.

[24] Y. H. B. Gunawan and M. L. Khodra, "Multi-document Summarization using Semantic Role Labeling and Semantic Graph for Indonesian News Article," *2020 7th International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2020*, 2020.

[25] R. Wijayanti, M. L. Khodra, and D. H. Widyantoro, "Single Document Summarization Using BertSum and Pointer Generator Network," *International Journal on Electrical Engineering and Informatics*, vol. 13, no. 4, pp. 916–930, dec 2021. [Online]. Available: http://ijeei.org/docs-204000029161f4c11fc91b1.pdf

[26] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," pp. 25–26, 2004. [Online]. Available: papers2://publication/uuid/5DDA0BB8-E59F-44C1-88E6-2AD316DAEF85

[27] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, oct 2020. [Online]. Available: http://arxiv.org/abs/1910.10683https://www.jmlr.org/beta/papers/v21/20-074.html

[28] M. Aurelia, S. Monica, and A. S. Girsang, "Transformer-based abstractive indonesian text summarization," *International Journal of Informatics and Communication Technology*, vol. 13, no. 3, pp. 388–399, 2024.

[29] R. A. Rahman and Suyanto, "Performance Analysis of ChatGPT for Indonesian Abstractive Text Summarization," *Proceedings - International Seminar on Intelligent Technology and its Applications, ISITIA*, no. 2024, pp. 477–482, 2024.

[30] A. L. Putra, Sanjaya, M. R. Fachruradzi, and A. Y. Zakiyyah, "Resource Efficient Abstractive Text Summarization in Indonesian with ALBERT," *2024 International Conference on Smart Computing, IoT and Machine Learning, SIML 2024*, pp. 81–85, 2024.

[31] Z. P. Ayudhia and S. Suyanto, "Deeper Investigation on Extractive-Abstractive Summarization for Indonesian Text," *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems, ICETSIS 2024*, no. 2022, pp. 1704–1708, 2024.

[32] M. D. Ferdiansyah and S. Suyanto, "Data Augmentation and Fine-Tuning to Improve IndoBART Performance," *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems, ICETSIS 2024*, pp. 1668–1672, 2024.

[33] A. S. Wijaya and A. S. Girsang, "Augmented-Based Indonesian Abstractive Text Summarization using Pre-Trained Model mT5," *International Journal of Engineering Trends and Technology*, vol. 71, no. 11, pp. 190–200, 2023.

[34] M. Nasari, A. Maulina, and A. S. Girsang, "Abstractive Indonesian News Summarization Using BERT2GPT," *Proceedings - 2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2023*, pp. 369–375, 2023.

[35] D. Puspitaningrum, "A Survey of Recent Abstract Summarization Techniques," 2022, pp. 783–801. [Online]. Available: https://link.springer.com/10.1007/978-981-16-2102-4$_$71

[36] H. Lucky and D. Suhartono, "Investigation Of Pre-Trained Bidirectional Encoder Representations From Transformers Checkpoints For Indonesian Abstractive Text Summarization," *Journal of Information and Communication Technology*, vol. 21, no. 1, pp. 71–94, nov 2022. [Online]. Available: https://e-journal.uum.edu.my/index.php/jict/article/view/13548

[37] R. E. Prasojo and A. A. Krisnadhi, "Controllable Abstractive Summarization Using Multilingual Pretrained Language Model," pp. 228–233, 2022.

[38] A. F. Abka, K. Azizah, and W. Jatmiko, "Transformer-based Cross-Lingual Summarization using Multilingual Word Embeddings for English - Bahasa Indonesia," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 12, pp. 636–645, 2022.

[39] A. Najibullah, "Indonesian Text Summarization based on Naïve Bayes Method," *International Seminar and Conference 2015 : The Golden Triangle (Indonesia-India-Tiongkok)*, pp. 67–78, 2015.

[40] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 2016, pp. 923–929.

[41] Y. Qi, D. S. Sachan, M. Felix, S. J. Padmanabhan, and G. Neubig, "When and why are pre-trainedword embeddings useful for neural machine translation?" *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 2, pp. 529–535, 2018.

[42] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 4512–4525, 2020.

[43] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, pp. 2214–2218, 2012.

[44] H. Schwenk, G. Wenzek, S. Edunov, E. Grave, A. Joulin, and A. Fan, "CCMatrix: Mining billions of high-quality parallel sentences on the web," in *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2021, pp. 6490–6500.