Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12873
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Paraphrase Generation for Reading Comprehension

Faishal Januarahman[1)*], Ade Romadhony[2)]
[1,2)]School of Computing, Telkom University, Indonesia
[1)]faishaljanuarahman@student.telkomuniversity.ac.id, [2)]aderomadhony@telkomuniversity.ac.id,

Submitted : Aug 11, 2023 | **Accepted** : Aug 14, 2023 | **Published** : Oct 1, 2023

Abstract: Reading comprehension is an assessment that tests readers understanding of a concept from the given text. The testing process is conducted by providing questions related to the content within the context of the text. The purpose of this research is to create new question variations from existing questions, and one of the methods to achieve this is by paraphrasing questions through the task of paraphrase generation. This can help ensure that readers have fully grasped a concept of a text. This study employs a traditional approach known as the thesaurus-based approach, in which the process involves substituting synonyms using the Indonesian Thesaurus dictionary. The data used consists of a list of Indonesian language reading comprehension assessment questions ranging from elementary to high school levels. To measure the quality of the generated paraphrased questions, two evaluation processes are conducted which are automatic evaluation with the scores ranging from 0-1 and human evaluation with score ranging from 1-4. The automatic evaluation includes the BLEU-4 metric, resulting in a score of 0.044, and the ROUGE-L metric, resulting an F1-score of 0.421. As for human evaluation, the obtained relevancy score is 2.533, and the fluency score is 3.186. The results from both evaluation metrics indicate that the generated paraphrased questions exhibit diverse new word choices but tend to have slightly different meanings compared to the reference questions.

Keywords: BLEU; Human Evaluation; Paraphrase Generation; ROUGE; Reading Comprehension; Thesaurus

INTRODUCTION

Reading Comprehension is one of the assessment methods that tests the readers understanding of concepts in a text. This method trains the reader to explore a text and identify its main intent, which will later be tested through questions (Soemantri, 2011). These questions are usually designed with answers that are already available from the text. As a result, the available question types become more diverse, depending on the information present in the text. Based on this, the aim of this study is to implement paraphrase generation, which intends to generate new paraphrased questions from existing ones. These questions have different word structures but still retain their original meanings. One of the benefits obtained through this approach is the ability to ensure that the reader has fully grasped a concept of a text (Rathod, Tu, & Stasaski, 2022).

There have been several studies conducted that have adopted paraphrase generation. In the research by (Bolshakov & Gelbukh, 2004), they performed paraphrase generation using the synonym substitution method with the assistance of WordNet to select synonym candidates for replacement, and internet statistics to choose the most suitable synonym candidates based on the surrounding words. Synonym substitution itself is a traditional approach that does not involve neural models in the paraphrase generation process (Zhou & Bhat, 2021). In the neural models approach, there have been several studies,

*name of corresponding author

including one by (Prakash et al., 2016) which used Seq2Seq to implement stacked residual LSTM networks for paraphrase generation. This approach has demonstrated an advanced ability in natural language generation. However, on the other hand, there is a weakness in terms of the lack of diversity in the words in the generated paraphrase outputs (Lin & Wan, 2021). Specifically for the Indonesian language, research by (Barmawi & Muhammad, 2019) has conducted paraphrase generation using the contextual synonym substitution method. They addressed the weaknesses of the NGM-based method proposed by Gadag (Gadag & Sagar, 2016) with the aim of enhancing the naturalness of the paraphrased sentences.

This study adopts the method of synonym substitutions by utilizing synonyms from the Indonesian Thesaurus (Sugono et al., 2008). The data used consists of a list of reading comprehension questions tested at the elementary to high school levels. The paraphrasing process involves checking whether the words present in the input question are included in the Indonesian Thesaurus, and then matching their tags. If both conditions are satisfied, the word is selected as a candidate for substitution. The substitution process includes selecting the first word that appears in the synonym list in the thesaurus, which eventually becomes the output of the paraphrase. To aid in this paraphrasing process, the input questions are tokenized and subjected to part-of-speech tagging using a source tagset taken from the research by (Dinakaramani, Fam, Luthfi, & Manurung, 2014).

To measure the quality of the paraphrased question results, several evaluation processes are carried out. One of these is the BLEU metric evaluation, which is utilized to gauge how well the paraphrased sentences correspond to the reference. Although originally developed for assessing machine translation systems, this method is also commonly used for paraphrase generation tasks. Another automatic evaluation method used is the ROUGE metric evaluation, which was initially designed for text summarization tasks and operates on a recall-based evaluation metric. Given that automatic evaluation metrics primarily emphasize n-gram overlaps, this study also incorporates human evaluation to judge the quality of the paraphrased questions. Human annotators are tasked with assessing both relevancy and fluency. Relevancy revolves around the semantic similarity between the reference and paraphrased questions, while fluency primarily pertains to the ease of reading and comprehending the meaning of the paraphrased questions.

LITERATURE REVIEW

There have been several studies that have adopted paraphrase generation methods according to their specific needs. This method can be implemented in various tasks such as question answering (Dong, Mallinson, Reddy, & Lapata, 2017), machine translation (Thompson & Post, 2020), and semantic parsing (Cao et al., 2020). Referring to (Zhou & Bhat, 2021), paraphrase generation approaches can generally be divided into two categories which is traditional approaches and neural approaches.

Neural approaches have been widely used in recent years. With the advancement of neural networks such as Seq2Seq, this model was first utilized for paraphrase generation in the research conducted by (Prakash et al., 2016). In this research, they explored deep learning models for paraphrase generation, specifically focusing on LSTM by incorporating residual connections between LSTM layers. This approach allowed for deeper training of the LSTM. In the research (Lin & Wan, 2021), a method called back-translation guided multi-round paraphrase generation was proposed. They combined neural paraphrase models with back-translation to generate paraphrases in a multi-round process. Despite the Seq2Seq-based models showcasing advanced capabilities, their output mostly involves altering a few words from the original question (Lin & Wan, 2021).

Traditional approaches are methods that do not involve neural models at all and are mostly carried out manually. Rule-based paraphrase generation is one method within this approach, built upon manually crafted paraphrase rules or patterns, or automatically gathered ones. Based on previous research that adopted the rule-based method (Quirk, Brockett, & Dolan, 2004), it is evident that if the paraphrase patterns used are complex and lengthy, they can impact the performance and limitations of the resulting paraphrased sentences. Besides rule-based, there is also the thesaurus-based approach, which generates paraphrases by substituting words in the reference sentence with their synonym lists extracted from a thesaurus or WordNet. This method is relatively straightforward, but it does have limitations in the diversity of paraphrases generated (Zhao et al., 2009).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12873 p-ISSN : 2541-044X

In this study, the thesaurus-based method is adopted by performing synonym substitution in the reference questions with their synonym lists extracted from the Indonesian Thesaurus (Sugono et al., 2008). Thesaurus-based paraphrasing itself is a simple and effective approach, since the rules are written by humans and there is ease of access to the thesaurus synonym dictionary. However, on the other hand, the weakness of this approach is that it cannot generate different types of paraphrases and relies solely on synonym substitution. In the initial phase, this process matched the words and tags in the reference questions with those present in the thesaurus. Then, if a match was found, the word replacement process was carried out using the list of synonym words that appeared first in the thesaurus. Based on this, the success of this process heavily relies on the performance of tokenization and part-of-speech tagging to classify the words in the reference questions.

METHOD

The objective of this study is to generate new paraphrased questions from the questions used in reading comprehension assessments. The dataset utilized in this study comprises reading comprehension assessment questions ranging from elementary school to high school levels. As depicted in Figure 1, this study commences by conducting preprocessing on the used data. The first step involves converting the question text to lowercase through case folding. Subsequently, tokenization is applied to the question text to transform it into word units. Following this, a POS-tagging process is executed to categorize words according to their part-of-speech classes.

After the preprocessing process, the paraphrase generation is conducted using the synonym substitution method, which retrieves synonym dictionaries from the Indonesian Thesaurus. Once the paraphrase process is completed, an evaluation is carried out to assess the quality of the paraphrased question results. The evaluation uses two methods which is automatic evaluation and human evaluation. BLEU and ROUGE-L metrics are utilized for the automatic evaluation process. Both of these metrics primarily assess the n-gram overlap between the original and paraphrased questions. Human evaluation is performed to assess the relevancy and fluency of the paraphrased question results.

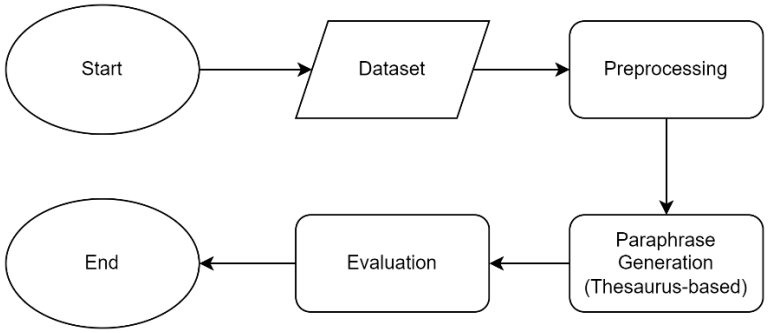

Figure 1. Flowchart Process of Paraphrase Generation with Thesaurus-based

Dataset

The data used in this study consists of a list of Indonesian reading comprehension assessment questions ranging from elementary school to high school levels. The overall quantity of data was 620 rows, each consisting of text and questions related to the context of the text itself. The topics of the text and questions covered include general knowledge, natural phenomena in Indonesia, legendary stories, fables, folklore, and also knowledge of Indonesian arts and culture. However, since this study focuses on paraphrase generation for the questions, only the question column is utilized, and the total number of questions used is 300.

Preprocessing

The preprocessing stage is one of the implemented phases in this study. The first step in this stage involves performing case folding, which is the process of converting all characters in the text to lowercase. The outcomes of the case folding process can be observed in Table 1. After case folding process, the next step is the tokenization process.In this process, a question is divided into smaller units

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12873 p-ISSN : 2541-044X

which are words, as they will be utilized to find their synonyms. As shown in Table 2, the tokenization procedure separates the question sentence into individual words.

The final step conducted in the preprocessing process is performing POS-tagging, which aims to categorize words according to their part-of-speech classes. The POS-tagging process is assisted by a source tagset taken from the research by (Dinakaramani, Fam, Luthfi, & Manurung, 2014). In their research, they designed a part-of-speech tagset for the indonesian language, resulting in a POS tagset consisting of 23 manually assigned tags. For this study, the tags used have been adapted to match the tags present in the Indonesian Thesaurus (Sugono et al., 2008). The outcomes of the POS-tagging process can be observed in Table 3.

Table 1. Example of Case Folding Process

Before	After
Kapal layar dan kapal feri merupakan jenis transportasi...	kapal layar dan kapal feri merupakan jenis transportasi...
Suatu drama dikatakan komedi bila...	suatu drama dikatakan komedi bila...
Teknik membaca yang digunakan untuk teks yang sulit, yaitu	teknik membaca yang digunakan untuk teks yang sulit, yaitu ....

Table 2. Example of Tokenization Process

Before	After
kapal layar dan kapal feri merupakan jenis transportasi...	[‘kapal’, ‘layar’, ‘dan’, ‘kapal’, ‘feri’, ‘merupakan’, ‘jenis’, ‘transportasi’]
suatu drama dikatakan komedi bila...	[‘suatu’, ‘drama’, ‘dikatakan’, ‘komedi’, ‘bila’]
teknik membaca yang digunakan untuk teks yang sulit, yaitu	[‘teknik’, ‘membaca’, ‘yang’, ‘digunakan’, ‘untuk’, ‘teks’, ‘yang’, ‘sulit’, ‘yaitu’]

Table 3. Example of POS-Tagging Process

Before	After
[‘kapal’, ‘layar’, ‘dan’, ‘kapal’, ‘feri’, ‘merupakan’, ‘jenis’, ‘transportasi’]	[(‘kapal’, ‘n’), ('layar', 'n'), ('dan', 'p'), ('kapal', 'n'), ('feri', 'n'), ('merupakan', 'v'), ('jenis', 'n'), ('transportasi', 'n')]
[‘suatu’, ‘drama’, ‘dikatakan’, ‘komedi’, ‘bila’]	[('suatu', 'num'), ('drama', 'n'), ('dikatakan', 'v'), ('komedi', 'n'), ('bila', 'SC')]
[‘teknik’, ‘membaca’, ‘yang’, ‘digunakan’, ‘untuk’, ‘teks’, ‘yang’, ‘sulit’, ‘yaitu’]	[('teknik', 'n'), ('membaca', 'n'), ('yang', 'SC'), ('digunakan', 'v'), ('untuk', 'IN'), ('teks', 'n'), ('yang', 'SC'), ('sulit', 'a'), ('yaitu', 'SC')]

Thesaurus

A thesaurus is one of the data sources used in this study for the synonym substitution process. A thesaurus itself is a dictionary that provides a set of words with related meanings. In this study, the source is taken from the Indonesian Thesaurus by the Pusat Bahasa, Departemen Pendidikan Nasional (Sugono et al., 2008). The data format of this thesaurus is in JSON format, with keys containing

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12873 p-ISSN : 2541-044X

Indonesian words. Each key has an object value with several more keys within it. The first key represents the tag of the corresponding Indonesian word source. In the used thesaurus, the list of tags includes adjectives, adverbs, figurative expressions, nouns, numerals, conjunctions, pronouns, and verbs. The second key represents synonyms and holds a list of synonym words for the given indonesian word. The last key represents antonyms and contains a list of antonym words for the respective word.

Paraphrase Generation

Paraphrasing is the primary focus of this study by using the thesaurus-based approach, which involves the process of paraphrasing by performing synonym substitution for certain words in the reference questions. As depicted in Figure 2, the data used consists of words and their corresponding tags obtained from the preprocessing process. Next, it will be checked whether these words are available in the Indonesian Thesaurus. If a word is not found, the paraphrasing process will halt and move on to the next word, if available. However, if the word is found, it will proceed to the tag matching process.

The matched tag represents the tag resulting from preprocessing and the tag of the word found in the thesaurus. If the tags do not match, the paraphrasing process will stop. However, if the tags match, the word will be selected as a candidate for replacement. The process of selecting synonymous words involves choosing the first word that appears in the synonym list for that word in the thesaurus. After selecting the synonym, the word in the reference question will be replaced by the chosen synonyms word, ultimately resulting in a new paraphrased question, which is the main output of this study.

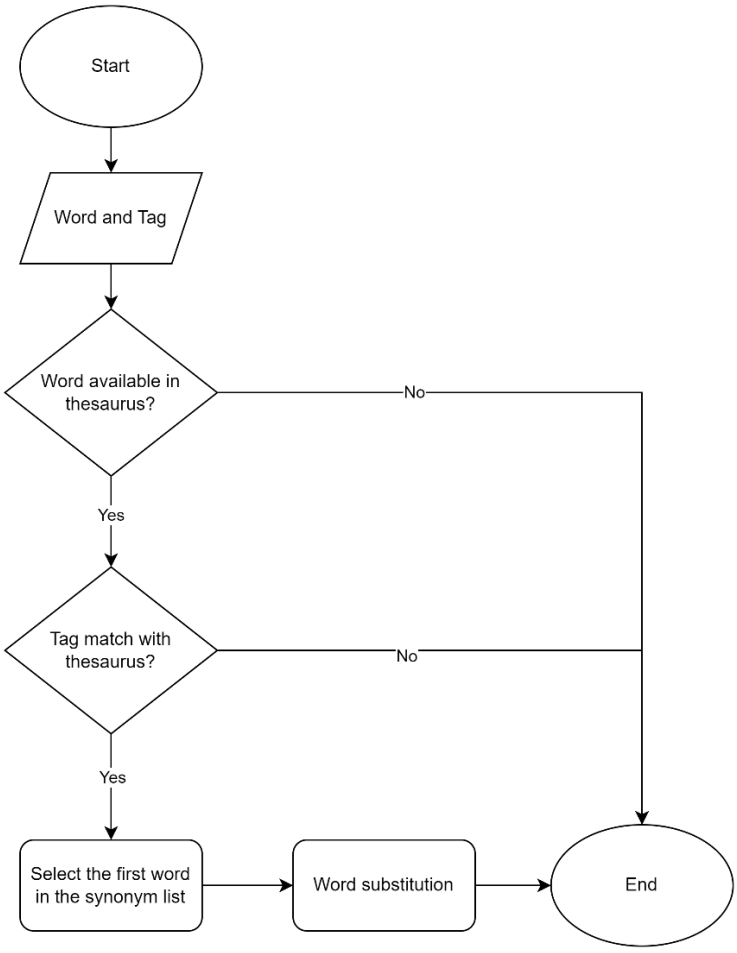

Figure 2. Flowchart of The Synonym Substitution Process

BLEU-4

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12873 p-ISSN : 2541-044X

In this study, one of the automatic evaluation methods used is the BLEU metric. BLEU itself is an evaluation technique originally used to assess machine translation (MT) tasks (Papineni, Roukos, Ward, & Zhu, 2002). However, in several recent studies that have adopted paraphrase generation, BLEU evaluation has also been used. This is because, fundamentally, this evaluation method measures the level of word overlap between the generated sentences and the reference sentences (Zhou & Bhat, 2021) (Lin & Wan, 2021).

BLEU primarily focuses on the calculation of n-gram precision. As seen in equation (1), precision calculation involves dividing the number of shared n-grams between the generated sentence and the reference sentence by the total number of n-grams present in the generated sentence. BLEU-4 represents the final outcome of this evaluation, as shown in equation (2). This involves calculating the geometric average of each precision generated, ranging from unigrams to 4-grams. The reason for limiting the calculation to 4-grams is to avoid the potential impact of word order being randomized between the generated and reference sentences, which could result in different sentence meanings.

$$N - gram\ precision = \frac{Clip(Num\ n{-}gram\ matches)}{Num\ n{-}grams\ in\ generated} \quad (1)$$

$$BLEU - 4\ Score = \sqrt[4]{p1 \times p2 \times p3 \times p4} \quad (2)$$

ROUGE-L

ROUGE is one of the automatic evaluation metrics used in this study, alongside BLEU. Initially, ROUGE was used to evaluate text summarization tasks (C. Lin, 2004), which also shares similarities with BLEU in measuring word overlap but focuses on both precision and recall values. While BLEU computes metrics based on various n-grams, in this study, the ROUGE-L variant is utilized. This variant calculates scores based on the longest common subsequence (LCS). LCS emphasizes finding the longest sequence of words that appear in both the generated and reference sentences. The sequence of words doesn't necessarily need to be contiguous but what matters is the order of their appearance.

As seen in equation (3), to calculating the recall value in this metric involves dividing the length ratio of the LCS by the number of unigrams in the reference sentence. To calculate precision, as seen in equation (4), the length ratio of the LCS is divided by the number of unigrams in the generated sentence. The final outcome reported in this metric evaluation is the F1-score, which factors in both precision and recall values, as shown in Equation (5).

$$ROUGE - L\ (recall) = \frac{LCS(generated,reference)}{Num\ words\ in\ reference} \quad (3)$$

$$ROUGE - L\ (precision) = \frac{LCS(generated,reference)}{Num\ words\ in\ generated} \quad (4)$$

$$ROUGE - L\ (F1 - Score) = 2 \times \left(\frac{precision \times recall}{precision + recall}\right) \quad (5)$$

Human Evaluation

In this study, a human evaluation process was conducted to measure the quality of the paraphrased questions. A sample of 30 questions was chosen, resulting in a total of 60 questions, including both the original and paraphrased questions for evaluation. Five undergraduate students were selected to assess each sample. The aspects evaluated in this process were relevancy and fluency. Relevancy assessed the semantic similarity between the generated questions and the reference questions, while fluency evaluated the ease of reading and comprehending the meaning of the paraphrased questions. A rating scale ranging from 1 (indicating the worst condition) to 4 (indicating the best condition) was used for this human evaluation process.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12873
e-ISSN : 2541-2019
p-ISSN : 2541-044X

RESULT

As previously mentioned, the output of this study is paraphrased questions generated using the thesaurus-based approach through synonym substitution. To evaluate the paraphrased question results, two evaluation methods were used which is automatic evaluation on a scale of 0 to 1, and human evaluation on a scale of 1 to 4. As shown in Table 4, the outcomes of both evaluation methods are presented along with their respective attributes. BLEU-4 reflects the precision calculations for unigrams to 4-grams, and the resulting precision values are geometrically averaged to produce a score of 0.044. Regarding ROUGE-L, three values are provided which precision, recall, and F1-score. Precision calculates the ratio of the number of LCS (longest common subsequence) elements in the generated paraphrased question that also appear in the reference question. Recall calculates the ratio of the number of LCS elements in the reference question that are also appear in the generated paraphrased question. F1-score, on the other hand, is the harmonic mean of precision and recall values. The F1-score indicates high or low the recall and precision values obtained. In this study, the F1-score achieved is 0.421.

The results of human evaluation represent the quality of meaning or semantics in the generated paraphrased questions. Relevancy measures the semantic similarity, while fluency gauges the ease of comprehending the meaning of the generated paraphrased questions. As seen in Table 4 and Figure 3, fluency tends to have higher values compared to relevancy. This suggests that the generated questions are easily understandable, even though their meanings slightly differ from the reference questions.

Table 4. Result of Automatic and Human Evaluation

Automatic Evaluation				Human Evaluation	
BLEU-4	ROUGE-L			Relevancy	Fluency
	Precision	Recall	F1-Score		
0.044	0.414	0.428	0.421	2.533	3.186

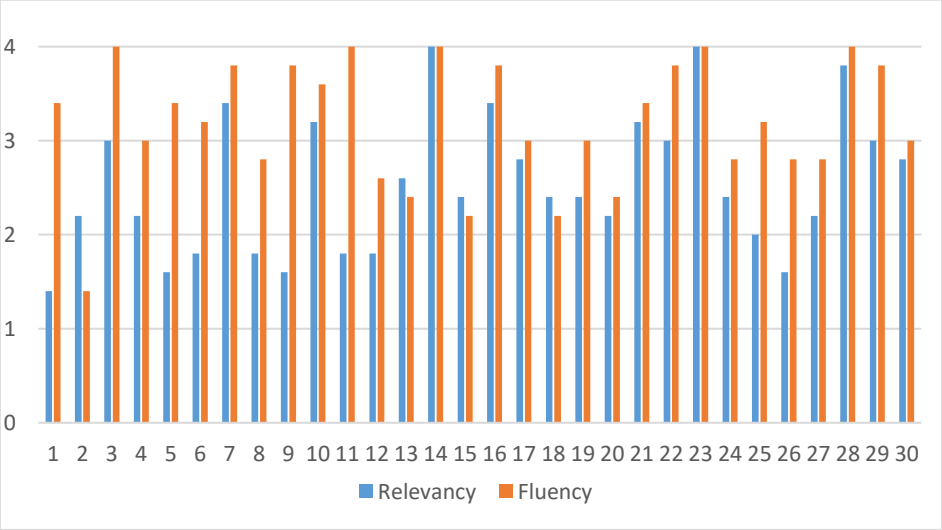

Figure 3. Result of Human Evaluation Using Questionnaire Method for Thirty Questions

DISCUSSIONS

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12873 p-ISSN : 2541-044X

Both BLEU and ROUGE are evaluation metrics that both rely on n-gram overlap to assess between generated questions and the reference questions. Since both metrics were originally developed for evaluating machine translation tasks (Papineni, Roukos, Ward, & Zhu, 2002) and text summarization tasks (C. Lin, 2004), they focus on the similarity of words appearing in both the reference and generated sentences. However, as paraphrasing aims to change sentence structures while preserving the original meaning, high scores on these metrics do not necessarily indicate the quality of the generated paraphrased questions. The outcomes of both metrics will have low values if there are many non-matching n-grams between the generated and reference sentences. Conversely, the metrics will yield high values if numerous matching n-grams are found. Given that paraphrasing involves changing sentence structures which also includes changing words, the lower scores on these metrics indicate that the generated sentences have achieved the goal of paraphrasing which is by changing some words. In this study, the BLEU-4 score achieved is 0.044, and the ROUGE-L score is 0.421, both of which tend to be low values. However, despite the low scores, it doesn't solely imply that the quality of the generated questions is good. This is because the objective of paraphrasing is not only to change sentence structures but also to maintain the meaning in alignment with the reference questions.

To assess this, as previously mentioned, a human evaluation process was conducted to evaluate relevancy and fluency. Using a scale of 1-4, it was found that the relevancy score was lower at 2.533, while the fluency score was 3.186. This demonstrates that the generated paraphrased questions, while tending to produce easily understandable meanings, still exhibit differences from the reference questions in terms of relevancy, as indicated by the evaluation results. Those problem caused by the limitations of this study which is only adapt synonym substitution method. Paraphrasing, on the other hand, can be achieved through various methods. As seen in Table 5, there is a comparison of the output results between the highest and lowest scores for relevancy and fluency. Although the generated paraphrased questions exhibit diversity in terms of new words, their meanings still tend to differ from the reference questions due to the limitations of the POS tagging model and the Indonesian Thesaurus used. Both of these aspects could be further developed in future research by adding some new rules other than synonym substitution in the traditional paraphrasing approach.

Table 5. Output Example of The Highest and Lowest Human Evaluation Score

Original Question	Paraphrased Question	Human Evaluation	
		Relevancy	Fluency
untuk apa gagak memunguti kerikil?	buat apa gagak memulung batu?	**3.8**	**4**
Dari soal sebelumnya, buatlah peta pikiran lengkapnya !	dari pertanyaan sebelumnya, buatlah atlas anggapan lengkapnya !	1.8	2.6

CONCLUSION

This study implements the task of paraphrase generation for reading comprehension questions ranging from elementary to high school levels, with the goal of introducing question variations that enable readers to better grasp a given reading text. The paraphrase generation process in this study uses a traditional approach known as the thesaurus-based approach. In this process, synonym substitution is performed using synonyms retrieved from the Indonesian Thesaurus dictionary. As seen from the evaluation section, the results of this study demonstrate that the generated paraphrased questions successfully incorporate new word variations that differ from the reference questions. However, from a human perspective, the paraphrased questions do not always perfectly align in meaning with the original questions, as indicated by a relevancy score of 2.533. Since the primary aim of paraphrasing is to retain the original meaning, achieving this objective may involve selecting synonymous words that are contextually appropriate for the given question in future research. Moreover, by adding some new rules such as adjustments to word order or the conversion of active and passive sentences could also be considered.

*name of corresponding author

REFERENCES

Barmawi, A. M., & Muhammad, A. (2019). Paraphrasing method based on contextual synonym substitution. Journal of ICT Research and Applications, 13(3), 257. https://doi.org/10.5614/itbj.ict.res.appl.2019.13.3.6

Bolshakov, I. A., & Gelbukh, A. (2004). Synonymous paraphrasing using WordNet and internet. In Lecture Notes in Computer Science (pp. 312–323). https://doi.org/10.1007/978-3-540-27779-8_27

Cao, R., Zhu, S., Yang, C., Liu, C., Ma, R., Zhao, Y., . . . Yu, K. (2020). Unsupervised Dual Paraphrasing for Two-stage Semantic Parsing. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.608

Dinakaramani, A., Fam, R., Luthfi, A., & Manurung, R. (2014). Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus. 2014 International Conference on Asian Language Processing (IALP). https://doi.org/10.1109/ialp.2014.6973519

Dong, L., Mallinson, J., Reddy, S., & Lapata, M. (2017). Learning to Paraphrase for Question Answering. EMNLP 2017. https://doi.org/10.18653/v1/d17-1091

Gadag, A., & Sagar, B. M. (2016). N-gram based paraphrase generator from large text document. International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS). https://doi.org/10.1109/csitss.2016.7779447

Lin, Z., & Wan, X. (2021). Pushing Paraphrase Away from Original Sentence: A Multi-Round Paraphrase Generation Approach. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. https://doi.org/10.18653/v1/2021.findings-acl.135

Lin, C. (2004). ROUGE: a package for automatic evaluation of summaries. Meeting of the Association for Computational Linguistics, 74–81. Retrieved from http://anthology.aclweb.org/W/W04/W04-1013.pdf

Papineni, K., Roukos, S., Ward, T. J., & Zhu, W. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Prakash, A., Hasan, S. A., Lee, K., Datla, V. V., Qadir, A., Liu, J., & Farri, O. (2016). Neural Paraphrase Generation with Stacked Residual LSTM Networks. International Conference on Computational Linguistics, 2923–2934. Retrieved from https://www.aclweb.org/anthology/C16-1275.pdf

Quirk, C., Brockett, C., & Dolan, W. B. (2004). Monolingual machine translation for paraphrase generation. Empirical Methods in Natural Language Processing, 142–149. Retrieved from https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/paraphrase_emnlp_2004_corrected.pdf

Rathod, M., Tu, T., & Stasaski, K. (2022). Educational Multi-Question Generation for Reading Comprehension. In Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022). https://doi.org/10.18653/v1/2022.bea-1.26

Soemantri, A. S. (2011). READING COMPREHENSION PROBLEMS ENCOUNTED BY THE STUDENTS OF HIGHER EDUCATION. JURNAL COMPUTECH & BISNIS, 5(2), 74–80. Retrieved from https://jurnal.stmik-mi.ac.id/index.php/jcb/article/download/69/64

Sugono, D., Sugiyono, Maryani, Y., Meity, D., Qodratillah, T., Budiwiyanto, A., Puspita D., Amalia D., Santoso, T. (2008). Tesaurus Bahasa Indonesia Pusat Bahasa. Departemen Pendidikan Nasional Indonesia

Thompson, B. J., & Post, M. (2020). Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.18653/v1/2020.emnlp-main.8

Zhao, S., Lan, X., Liu, T., & Li, S. (2009). Application-driven statistical paraphrase generation. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. https://doi.org/10.3115/1690219.1690263

Zhou, J., & Bhat, S. (2021). Paraphrase Generation: a survey of the state of the art. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/2021.emnlp-main.414

.

*name of corresponding author