CHAPTER 1

INTRODUCTION

Reading is a fundamental concept for an individual to learn and master due to the importance of acquiring information. Indonesia, one of the fourth largest populations in the world, surprisingly has low reading literacy. Based on [1] Indonesia's reading literacy performance fluctuated after its peak in 2009. This fact is supported by [2] Indonesian students are good at understanding single text but weak at understanding multiple texts. Indonesian students are good at finding information, evaluating, and reflecting information, but weak at understanding the information. Although Indonesia's access to education improves every year, reading literacy remains a challenge within the country, the latest PISA test in 2022 confirmed this fact [1]. One of the solutions that can mitigate this problem is facilitating active learning. Active learning involves the process of reconstructing knowledge stored in the brain, so it is stored within long-term memory [3]. One form of activity is formal questioning such as during exams, or informal questioning carried out independently by students [3]. Specific types of questions that will be utilized in this study will be explained in the next paragraph.

Questions play a significant role in the teaching-learning process [4]. One type of question is Multiple Choice questions (MCQ). Multiple-choice tests are considered one of the most successful and long-lasting educational assessment methods in which, rather than writing the answer to the question, the respondent has to choose the right answer among multiple options [5]. This format is most suitable for measuring knowledge of any cognitive demand (recall, comprehension, or application). Moreover, they are very good at measuring the application of knowledge and skills that are intended to reflect ability [6]. However, MCQ creation is a complex process involving different components components: A Question, correct answer, and incorrect options [6]. The given question must be contextually relevant based on a given text, while the options must be answerable and involve creating incorrect answer options to distract the learners. Preparing high-quality MCQs manually is time-consuming and labor-intensive [7]. Thus underlines the need for automation. In addition, the current studies on MCQs are built based on integrating two components, Question generation and distractor generation. An overview of studies on Question Generation in Indonesia, Distractor Generation Globally, and Multiple-choice Question Generation progress will be mentioned. All mentioned studies are based on a deep learning approach.

Currently, the trends in Indonesian NLP are toward question generation (QG) for the Indonesian language. Muis et al. [8] utilized sequence-to-sequence learning with translated datasets, achieving moderate success despite dataset limitations. Building on this, Fuadi et al. [9] improved QG by leveraging a native Indonesian dataset and the mT5 model, show-

ing that native-language datasets can improve automatic metrics performance. [10] tried to compare the multilingual and monolingual models, and found the multilingual generates better BLEU and ROUGE-L score. However, none of these studies conduct human-level performance analysis. Despite the acceleration of question generation in the Indonesian language showing promising results, the current distractor generation in low-resources languages, like Indonesian has not yet been extensively researched. Furthermore, there is no specific research that covers the deep learning approach for distractor generation in the Indonesian language. However, to provide context for the research, the latest distractor generation studies in the English language will be mentioned.

Distractor Generation (DG) is not a relatively new field. The latest research to generate semantic-rich distractors using deep learning is started by Gao et.al [11]. The authors proposed a Hierarchical Encoder-Decoder along with custom attentions and components to tackle the problem. Their result shows that the average BLEU 1,2,3,4 and ROUGE-L are 26.93, 13.57, 8.00, 5.21, and 14.54 respectively. The main limitation of the proposed method is some generated distractors are semantically away from the article. Zhou et al [12] solved the problem in [11] . The authors proposed Co-Attention and Hierarchical attention to generate coherent-long distractors and a semantic similarity loss function. The results yield better performance compared with [11] in automatic metrics and human-level performance. The main limitation of this research is the authors focus on the relation between distractors and passages. The 2 previous studies' essential problems do not consider the relation between passage, question, and correct answer information when generating distractors.

Qiu et.al [13] has the same purpose as what [11], [12] have done. The proposed method difference is in the additional information modeling and modules to utilize the newly added information. The results are better than the previous methods. The main limitation of this research is the author does not do additional error analysis to the generated distractors. Although the previously literature are introducing architecture modification changes. Research from [14] utilized the mT5 model. Their results show that the proposed model mT5 is able to outperform previous research. The main limitations of this research are not conducting any human-evaluation and error analysis results.

Finally, the research for MCQ generation, combining Question Generation and Distractor Generation to work together has also grown. In recent years, there has been a growth toward the development of the MCQ generation. Vachef et al. [15] employed T5 transformer models to generate multi-task question generation and multi-task distractors generation, but their system lacked comprehensive human evaluation. Rodriguez et al. [16] also utilizes T5 for multi-task question generation, but single-task distractor generation. Moreover, Rodriguez et.al have done human testing to validate the quality and difficulty level of their proposed MCQs. While these studies provide frameworks, they primarily target English-language MCQs and currently research in MCQ in the Indonesian language using deep learning remains a challenge.

1.1 Statement of the problems

Based on previous Indonesian studies, distractor generation and multiple-choice question (MCQ) generation remain a challenging area in the Indonesian language. Most existing research primarily focuses on question generation, which is only one component of MCQ construction. However, these QG studies have several limitations. Many rely on overlap-based metrics but do not conduct in-depth evaluations of result quality. Additionally, human-level judgment is rarely implemented, despite its importance, as highlighted by [11], [12], [13], [16], and [17] suggests. Furthermore, most studies in this domain depend on translated datasets, which may not fully capture the linguistic nuances of the Indonesian language.

While several challenges exist in Indonesian MCQ generation, this study focuses on addressing several key issues. Current research efforts predominantly emphasize QG, whereas distractor generation using deep learning remains underexplored. Moreover, error analysis of both QG and DG components has not been throughly investigated. Since these 2 components are essential for constructing MCQs, this study aims to measure a deep-learning based MCQ generation system for the Indonesian language and investigated the components of MCQ through error analysis.

Additionally, different approaches to MCQ generation, such as those proposed by [15] and [16], introduce further complexity. This study explores various MCQ configurations in the Indonesian language, which combine Single-Task Question Generation, Multi-Task Question Generation, Single-Task Distractor Generation, and Multi-Task Distractor Generation, all of which will be evaluated using human-level performance assessments.

1.2 Conceptual Framework/Paradigm

The conceptual framework helps to understand the key factors related to this research and shows how they interact with each other. This study, focuses on developing, and evaluating the factoid Indonesian multiple-choice questions using text-to-text transformers identifies and explains the following key elements:

1. Text data
   Text data refers to the input text from SQuAD and RACE datasets. These datasets serve as the primary input for the predictive model. In this study, the T5.

2. T5 Model
   T5 refers to Text-to-Text transfer Transformers developed by Google Research. It is a transformer-based architecture designed for a variety of natural language processing tasks. T5 is named T5, because all NLP tasks are cast into a text input to text output format.

3. Human Evaluation
   Human evaluation is the process of assessing the quality of NLP outputs through human judgment. It is often necessary because automated metrics (e.g., BLEU, ROUGE) do not always correlate well with human perceptions of quality.

1.3 Research Problem

Based on the statement of the problem section, the overall problem statements are outlined as follows:

1. How effective are T5-based MCQ configurations in generating factoid Indonesian multiple-choice questions from translated datasets, as assessed by human-level performance metrics?

2. How do error analysis reflect the relevancy of the Question Generations and Distractor Generations for Single-task and Multi-task?

1.4 Objective and Hypotheses

1.4.1 Objectives

The objectives of this research are as follows:

1. To evaluate the effectiveness of T5-based MCQs configuration in producing factoid Indonesian multiple-choice questions from translated datasets, using human-level performance metrics as evaluation criteria.

2. To compare the relevancy results of the multi-task and single task approaches in Question Generators and Distractor Generators that are generated by the T5 model.

1.4.2 Hypotheses

Premise 1: Previous research indicates that T5 can be used to generate question generation [16], [9], and [10], or distractor generation [16], [14]. Moreover, the latest research in MCQ [16], and [15] used T5. Given its capabilities in generating both questions and distractors, and considering that prior research [16] has evaluated MCQ outputs through human judgment, we hypothesize that when Indonesian MCQs generated by the T5 model are assessed against human-level performance, they will exhibit good quality, and easy difficulty as the result by [16].

Premise 2: Based on [16], and [15], a deep learning approach for MCQ generation can be built using two integrated components: a Question Generator and a Distractor Generator. Research from [15] used a multi-task approach to generate MCQs, while [16] employs a multi-task approach for the Question Generator and a single-task approach for

the Distractor Generator. We hypothesize that a comparative analysis of these approaches in the Indonesian context will yield meaningful insights into the relevancy of generated outputs. Specifically, we expect that the single-task approach will produce more relevant outputs compared to the multi-task approach.

1.5 Assumption

There are 4 assumptions employed in this research. To begin with, The T5 model is assumed to be capable of handling diverse tasks due to its complex architecture. However, alternative architectures (e.g., mT5) were not explored due to computational constraints. Next, the proposed Question Generation are factoid, while the distractors are non-factoid due to the nature of these datasets. Additionally,The dataset used for both QG and DG was obtained through translation, and the observations are assumed to be sufficient for the study's purposes. However, a detailed dataset quality analysis is outside the research scope. The accuracy of the translation tool (Google Translate, 86%) is assumed reasonable based on [18]. No indepth manual verification of translations was conducted, as this falls outside the research scope.

1.6 Scope and Delimitation

A. **Principal Variables**

1. **Independent Variables**
   The independent variables include T5-Base model, translated RACE and SQuAD datasets, and the variations of Question Generators and Distractor Generators used for the MCQs.
2. **Dependent Variables**
   The dependent variables are automatic metric results, error-analysis of Question Generation & Distractor Generation, and the human-level performance results of the MCQs survey.

B. **Locale**
   The research focuses on Indonesian, making it specifically applicable to Indonesian MCQ studies. However, the ideas generated from this research can be extended to another languages.

C. **TimeFrame**
   The study is conducted over a span of 1 year 2 months, from September 2023 to January 2025.

D. **Justification**

1. The research duration includes code writing & model training, translation of the RACE dataset, error analysis, survey creation, survey analysis, thesis book and manuscript writing.

E. **Limitations**

This study has its own limitations, which are described as follows:

1. The use of translated datasets may introduce inaccuracies, potentially limiting the model's ability to fully capture Indonesian linguistic nuances.
2. A detailed evaluation of the translated dataset quality is not conducted, as it is beyond the research scope.
3. T5-Base model is used, which may limit the pre-trained capabilities for the Indonesian language nuances, as this model was trained on the C4 dataset.
4. Alternative models such as mT5 (which is trained on multiple languages, including Indonesian) were not explored due to computational constraints. The mT5-Base model has 550M parameters, requiring significantly more resources than T5-Base [19].
5. No hyperparameters tuning were involved, the hyperparameters may not optimal.

1.7 Significance of the Study

This research is significant in advancing automated Multiple-Choice Question (MCQ) generation, particularly in Natural Language Generation (NLG) for the Indonesian language. Given the limited research and resources available for Indonesian MCQ generation, this study serves as a foundational step toward improving Multiple-Choice Question generation using deep learning in the Indonesian language.

1.8 Main Academic Contributions

The significance contribution of this research lies in its potential to advance the field of automated Multiple-choice question (MCQ) generation, particularly in the Natural Language Generation in the Indonesian language. Specifically, this study holds the following contributions:

1. By implementing the T5 model with translated datasets, this research aims to provide a foundational step for generating MCQs in the Indonesian language. Since the application of deep learning for MCQ generation in Indonesian remains a challenging area, this study serves as a foundational step toward addressing this gap.

2. This study compares multi-task and single-task approaches both on the question and distractor generations, providing insights into which method yields better relevancy for their given tasks and determine whether the error distribution is statistically significance or not. The findings may help another researcher to understand the current capabilities of the T5 model in tackling these problems in the future.

3. This study is also consider the human-level performance, as currently there are no standard metrics in these fields. Rather than relying solely on BLEU, ROUGE-L, and SBERT Scores, this study incorporates previously established human evaluation metrics to assess quality, and difficulty. The results will provide another researcher with an understanding of the human perspective.

4. This study explores different combinations of question generation and distractor generation to examine their impact on MCQ quality and they are assessed by human judgment.

5. This study not only contributes to the development of Indonesian-language educational tools but also contributes to the advancements of NLG research in the Indonesian language.