Abstract
Paraphrasing can be defined as the expression of a text in different diction while referring to the same meaning. An automatic paraphrase generation system plays a crucial role in Natural Language Processing (NLP). In previous studies, the resulting paraphrase dataset was extracted using a translation machine with the assumption that text pairs certainly have semantic similarities. So, the filter used is only on the difference in the variety of diction. As a result, the resulting dataset tends to be unsatisfactory in terms of lexical diversity and semantic similarity. Therefore, this research aims to generate a paraphrase dataset by utilizing the Abstractive Summarization task on the Liputan6 dataset. The summary text of humans in the Liputan6 dataset will be paired with the summary text of the system. After that, the text pair will be filtered based on the average of semantic similarity using BERTScore and lexical diversity using inverseSacreBLUE. The filtration process used has proven successful in increasing lexical diversity compared to previous studies which showed an increase in the inverseSacreBLEU score from 57.42 to 72.76. The dataset resulting from liputan6 (146,030 data) is nearly 40 times smaller than the previous study (5,753,296 data), but has higher semantic similarity and lexical diversity values. This shows that the quality of the resulting dataset is better than previous studies.

Keywords: paraphrase generation, semantic similarity, lexical diversity