ABSTRACT

During the COVID-19 pandemic's impact on education, the shift to online learning has become a new norm. Automatic evaluation of short essay responses has become crucial in the context of online education, yet it remains a significant challenge. This research focuses on the development of the Automatic Short Answer Scoring (ASAS) system using approaches such as Maximum Marginal Relevance (MMR) and paraphrasing through three customized deep learning models.

MMR demonstrated an accuracy of 88.5% in creating key answer variations, which were subsequently used as references in the ASAS evaluation. Subsequently, the generation of reference answers was conducted through paraphrasing utilizing the BART, GPT-2, and GPT-2 template models, resulting in a maximum of 40 key answers, including the original key answer.

The analysis results indicate a significant improvement in Root Mean Square Error (RMSE), correlation, and Mean Absolute Error (MAE) with the addition of paraphrase generation from the three models (GPT, BART, GPT Template). The results are a 15,67% enhancement in RMSE, 25% in correlation, and a reduction of 14% in MAE signify a substantial enhancement in the quality of answer variations and similarity to evaluator assessments. Statistical tests using the Wilcoxon signed rank test also confirm a significant difference between the average outcomes of the MMR method and the usage of three paraphrase models.

Keywords: Automatic Short Answer Scoring, Maximum Marginal Relevance, Paraphrase Generation, Text Similarity, BART, GPT.

INTISARI

Dalam situasi pendidikan yang terpengaruh oleh pandemi COVID-19, transisi ke pembelajaran daring telah menjadi kebiasaan baru. Penilaian otomatis terhadap jawaban esai singkat menjadi penting, tetapi juga menjadi sebuah tantangan. Penelitian ini berfokus pada pengembangan *Automatic Short Answer Scoring* (ASAS) dengan menggunakan pendekatan pembangkit jawaban seperti *Maximum Marginal Relevance* (MMR) dan parafrase melalui tiga model *deep learning* yang telah disesuaikan.

MMR menunjukkan akurasi sebesar 88,5% dalam menciptakan variasi referensi jawaban. Selanjutnya, dilakukan generasi referensi jawaban melalui parafrase menggunakan model BART, GPT-2, dan GPT-2 template, menghasilkan hingga 40 referensi jawaban termasuk yang asli. Seluruh referensi jawaban tersebut akan diperhitungkan untuk penilaian sistem uraian singkat otomatis.

Hasil analisis menunjukkan peningkatan signifikan pada nilai *Root Mean Square Error* (RMSE), korelasi, dan *Mean Absolute Error* (MAE) dengan penambahan generasi parafrase tiga model (GPT, BART, GPT Template). Perbaikan sebesar 15,67% pada RMSE, 25% pada korelasi, dan pengurangan 14% pada MAE mengindikasikan peningkatan substansial dalam kualitas variasi jawaban dengan kemiripan dengan penilaian evaluator. Hasil uji statistika dengan *Wilcoxon Signed Rank Test* juga memastikan perbedaan signifikan antara rata-rata hasil Metode MMR dengan penggunaan tiga model parafrase.

Kata kunci -- *Automatic Short Answer Scoring*, *Maximum Marginal Relevance*, *Paraphrase Generation*, *Text Similarity*, BART, GPT.