ABSTRACT

TRANSFORMER-BASED MARIANMT FOR ENGLISH-INDONESIA TRANSLATION WITH FINE-TUNING DATA TECHNOLOGY, EDUCATION, AND DESIGN (TED) TALKS

By

Adinda Putri Hermawan

In the era of globalization, the need for accurate automatic translators is increasing to support cross-language communication, especially from English to Indonesian. This study aims to evaluate the effectiveness of fine-tuning on the transformer-based MarianMT model using formal TEDTalks data, and analyze the influence of question marks in question sentences on translation quality. The methods used include data pre-processing, division of the dataset into five scenarios of training proportion and validation, model hyperparameter adjustment, and application of fine-tuning techniques with early-stopping and dropout to avoid overfitting. Evaluation was performed using the BLEU score on the test data. The results show that fine-tuning with in-domain data can improve translation accuracy, indicated by the highest BLEU gain of 32.14. This study also shows that the removal of question marks affects the meaning of translation, especially in short and medium sentences. The 90:10 (training:validation) data proportion scenario produced the best performance with good loss value stability. Thus, a combination of fine-tuning and proper pre-processing settings can produce more accurate and contextually appropriate translations.

Keywords: Neural Machine Translation, Transformer, MarianMT, Fine-Tuning, TEDTalks, Language Formality, BLEU Score

ABSTRAK

MARIANMT BERBASIS TRANSFORMER UNTUK PENERJEMAHAN BAHASA INGGRIS-INDONESIA DENGAN FINE-TUNING DATA *TECHNOLOGY, EDUCATION, AND DESIGN* (TED) *TALKS*

Oleh

Adinda Putri Hermawan

Pada era globalisasi, kebutuhan akan penerjemah otomatis yang akurat semakin meningkat untuk mendukung komunikasi lintas bahasa, khususnya dari bahasa Inggris ke bahasa Indonesia. Penelitian ini bertujuan untuk mengevaluasi efektivitas *fine-tuning* pada model MarianMT berbasis *transformer* menggunakan data TEDTalks yang bersifat formal, serta menganalisis pengaruh tanda tanya dalam kalimat tanya terhadap kualitas terjemahan. Metode yang digunakan meliputi *pre-processing* data, pembagian dataset ke dalam lima skenario proporsi pelatihan dan validasi, penyesuaian *hyperparameter* model, serta penerapan teknik *fine-tuning* dengan *early-stopping* dan *dropout* untuk menghindari *overfitting*. Evaluasi dilakukan menggunakan BLEU *score* pada data uji. Hasil penelitian menunjukkan bahwa *fine-tuning* dengan data *in-domain* dapat meningkatkan akurasi terjemahan, ditunjukkan oleh perolehan BLEU tertinggi sebesar 32,14. Penelitian ini juga menunjukkan bahwa penghapusan tanda tanya memengaruhi makna terjemahan, khususnya pada kalimat pendek dan sedang. Skenario proporsi data 90:10 (pelatihan:validasi) menghasilkan kinerja terbaik dengan stabilitas nilai loss yang baik. Dengan demikian, kombinasi *fine-tuning* dan pengaturan *pre-processing* yang tepat dapat menghasilkan terjemahan yang lebih akurat dan sesuai konteks.

Kata-kata kunci: *Neural Machine Translation*, *Transformer*, MarianMT, *Fine-Tuning*, TEDTalks, Keformalan Bahasa, BLEU *Score*.