Pengembangan Abstractive-Extractive Text Summarization dengan BART untuk Teks Berita Bahasa Indonesia = Development of Abstractive-Extractive Text Summarization with BART for Indonesian News Text

Michael Harditya, author

Abstrak

Penelitian ini melakukan pengembangan integrasi metode perangkum abstraktif dengan metode ekstraktif dalam merangkum teks berita yang melebihi input maksimal dari model machine learning. Penggabungan metode abstraktif dan ekstraktif menciptakan rangkuman yang lebih natural tanpa kehilangan makna semantiknya, serta menyelesaikan keterbatasan jumlah input maksimal dari model machine learning yang digunakan pada metode abstraktif. Bagian abstraktif dibuat menggunakan model machine learning yang menggunakan arsitektur Transformer, yaitu model BART. Bagian ekstraktif menggunakan algoritma gabungan untuk melakukan pembobotan tiap kalimat menggunakan term frequency – inverse document frequency (TF-IDF), konjungsi antar kalimat, dan peletakan kalimat pada paragraf yang dapat diidentifikasi menggunakan algoritma pemrograman. Dataset yang digunakan adalah benchmark IndoSum, yaitu dataset bahasa Indonesia untuk merangkum teks, sehingga dapat dievaluasikan dengan model pada penelitian yang serupa. Beberapa pengujian dilakukan pada model BART dan tokenizer, dengan nilai ROUGE Score menunjukan adanya peningkatan pada tokenizer bahasa Indonesia ketimbang bahasa Inggris. Hasil evaluasi pada finetuning model BART mendapatkan nilai ROUGE Score sebesar 0,725 untuk ROUGE-1, 0,635 untuk ROUGE-2, 0,699 untuk ROUGE-L, dan 0,718 untuk ROUGE-Lsum, menjadikan model BART lebih tinggi pada beberapa model lainnya pada riset terkait. Human evaluation dilakukan pada hasil integrasi, menunjukan hasil yang baik untuk morfologi, semantik, dan kenaturalan rangkuman, namun masih buruk untuk kesalahan pengetikan.

.....This research develops the integration of abstractive summarization methods with extractive methods in summarizing news texts that exceed the maximum input from the machine learning model. Combining abstractive and extractive methods creates a more natural summary without losing its semantic meaning, and resolves the limitations of the maximum number of inputs from the machine learning model used in the abstractive method. The abstractive part was created using a machine learning model that uses the Transformer architecture, namely the BART model. The extractive section uses a combined algorithm to weight each sentence using term frequency - inverse document frequency (TF-IDF), conjunctions between sentences, and placement of sentences in paragraphs that can be identified using a programming algorithm. The dataset used is the IndoSum benchmark, namely an Indonesian language dataset for summarizing text, so that it can be evaluated with models in similar research. Several tests were carried out on the BART model and tokenizer, with the ROUGE Score showing an increase in the Indonesian language tokenizer compared to English. The evaluation results of finetuning the BART model obtained a ROUGE Score of 0.725 for ROUGE-1, 0.635 for ROUGE-2, 0.699 for ROUGE-L, and 0.718 for ROUGE-Lsum, making the BART model higher than several other models in related research. Human evaluation was carried out on the integration results, showing good results for morphology, semantics and naturalness of summaries, but still poor results for typing errors.