

Pengembangan Model Pencarian Semantik Bahasa Indonesia Menggunakan Bidirectional Encoder Representations from Transformer (BERT) = Development of Indonesian Semantic Search Model Using Bidirectional Encoder Representations from Transformer (BERT)

Aryoshi Wicaksono, author

Deskripsi Lengkap: <https://lib.ui.ac.id/detail?id=9999920564691&lokasi=lokal>

Abstrak

Penelitian ini bertujuan untuk mengembangkan sistem pencarian semantik berbasis model Bidirectional Encoder Representations from Transformers (BERT) untuk mendukung bahasa Indonesia dan Inggris. Sistem ini dirancang untuk memahami konteks kueri secara multibahasa dengan memanfaatkan teknik knowledge distillation. Model “teacher” dilatih pada tugas question-answering dengan bahasa Inggris, sementara model “student” dilatih untuk menangkap konteks multibahasa. Dataset mMARCO dan Mr. TyDi digunakan untuk pelatihan dan pengujian. Hasil penelitian menunjukkan bahwa model berhasil mempelajari embedding multibahasa dengan akurasi translasi hingga 99%. Kualitas pemeringkatan teks dievaluasi menggunakan nilai metrik reciprocal rank (RR), recall (R), dan normalized discounted cumulative gain (nDCG). Pada metrik recall, nilai tertinggi dicapai oleh MS MARCO (EN) dengan $R@1000$ sebesar 0,8085. Untuk metrik precision, MS MARCO (EN) unggul pada $P@10$ dengan 0,3090. Pada metrik nDCG, nilai tertinggi juga diraih MS MARCO (EN) pada $NDCG@1000$ sebesar 0,2709, sedangkan Mr. TyDi (ID) mencatat 0,2340 pada $NDCG@100$. Untuk metrik MRR, Mr. TyDi (ID) mencapai hasil terbaik dengan $MRR@3$ sebesar 0,1585 dan $MRR@1000$ sebesar 0,1796.

.....This study aims to develop a semantic search system based on the Bidirectional Encoder Representations from Transformers (BERT) model to support both Indonesian and English languages. The system is designed to understand multilingual query contexts by utilizing the knowledge distillation technique. The teacher model was trained on an English question-answering task, while the student model was trained to capture multilingual contexts. The datasets used for training and testing include mMARCO and Mr. TyDi. The results demonstrate that the model successfully learned multilingual embeddings with translation accuracy reaching up to 99%. The quality of text ranking was evaluated using the metrics reciprocal rank (RR), recall (R), and normalized discounted cumulative gain (nDCG). For Recall, the highest value was achieved by MS MARCO (EN) with $R@1000$ of 0.8085. In terms of Precision, MS MARCO (EN) excelled at $P@10$ with a score of 0.3090. Regarding nDCG, the highest value was also achieved by MS MARCO (EN) at $NDCG@1000$ with 0.2709, while Mr. TyDi (ID) recorded 0.2340 at $NDCG@100$. For MRR, Mr. TyDi (ID) performed best, reaching $MRR@3$ of 0.1585 and $MRR@1000$ of 0.1796.