

Neural Re-Ranker untuk Mengidentifikasi Pertanyaan Serupa pada Forum Kesehatan Berbahasa Indonesia = Neural Re-Rankers to Identify Duplicate Questions in Indonesian Health Forums

Ezra Pasha Ramadhansyah, author

Deskripsi Lengkap: <https://lib.ui.ac.id/detail?id=9999920552170&lokasi=lokal>

Abstrak

Sistem perolehan pertanyaan serupa diimplementasikan pada banyak situs tanya jawab, khususnya pada forum tanya jawab kesehatan. Implementasi dari sistem pencarian pertanyaan serupa dapat beragam seperti text based retriever dan neural ranker. Permasalahan utama dari neural ranker adalah kurangnya penelitian dalam bahasa Indonesia untuk modelnya, khususnya untuk yang menggunakan BERT sebagai model untuk deteksi pertanyaan serupa. Pada penelitian ini akan dicari tahu sejauh apa neural re-ranker BERT dapat memperbaiki kualitas ranking dari text-based retriever jika diterapkan fine-tuning pada model. Model yang digunakan oleh penelitian berupa BERT dan test collection yang digunakan merupakan dataset forum kesehatan yang disusun oleh Nurhayati (2019). Untuk mengetahui sejauh mana model berbasis BERT dapat berguna untuk re-ranking, eksperimen dilakukan pada model pre-trained multilingualBERT, IndoBERT, StevenWH, dan distilBERT untuk melihat model yang terbaik untuk di-fine-tune. Penelitian juga mengusulkan dua metode fine-tuning yakni attention mask filter dengan IDF dan frozen layer dengan melakukan freezing pada beberapa layer di dalam BERT. Model dan metode ini kemudian diuji pada beberapa skenario yang telah ditentukan. Hasil dari eksperimen menunjukkan bahwa re-ranker dapat meningkatkan kualitas text based retriever bila di-fine-tune dengan metode dan skenario tertentu.

Beberapa model memberikan hasil yang lebih baik dengan dataset forum kesehatan dan dengan text based retriever BM25 dan TF-IDF. Model multilingualBERT dan metode fine-tuning layer freezing memberikan hasil yang terbaik dari semua kombinasi. Kenaikan tertinggi terdapat pada kombinasi BM25 dan multilingualBERT dengan layer freezing dengan kenaikan sebesar 0.051 dibandingkan BM25.

.....The system of acquiring similar questions is implemented on many Question and Answering sites, including health forums. Implementations of similar question search systems can vary, such as text-based retrievers and neural rankers. The main issue with neural rankers is the lack of research in Indonesian language for neural ranker models, especially those using BERT. This study aims to investigate how far BERT as a neural re-ranker can improve the ranking quality of a text-based retriever when applied with fine-tuning. The model used in this research is BERT, and the test collection used is a health forum dataset compiled by Nurhayati (2019). To answer the research question, experiments were conducted on multiple pre-trained models: multilingual BERT, IndoBERT, StevenWH, and distilBERT to identify the best model for fine-tuning. This study also proposes two new fine-tuning methods: attention mask filter with IDF thresholding and frozen layer by freezing some layers within BERT. These models and methods were then tested under predefined scenarios. The experiment results show that the re-ranker can enhance the quality of the text-based retriever when fine-tuned with specific methods and scenarios. These models perform especially well using the health form dataset as well as using the text based retrievers BM25 and TF-IDF. Out of all models, multilingualBERT performed the best with frozen layer fine-tuning performing as the best fine-tuning method. The most significant increase of all combinations is the combination of BM25 and multilingualBERT with frozen layer fine-tuning with a 0.051 increase compared to the baseline BM25.