Dense passage retriever pada tugas pencarian pertanyaan serupa dengan data pertanyaan forum kesehatan = Dense passage retriever for similar questions retrieval task on consumer health forum questions data

Mahardika Krisna Ihsani, author

Abstrak

Penelitian pada data berbahasa Inggris menemukan bahwa Dense Passage Retriever atau DPR mempunyai keterbatasan dalam hal menangani kondisi out-of-distribution data termasuk out-of-domain data. Saat ini, data latih berbahasa Indonesia yang bisa digunakan untuk melatih DPR cukup terbatas. Semua data latih tersebut berasal dari domain umum yang jika digunakan untuk melatih DPR mungkin menghasilkan performa yang rendah pada data uji dengan domain spesifik. Penelitian ini membandingkan antara performa DPR yang dilatih pada data latih dengan domain berbeda dengan domain data uji dan performa sparse retriever model untuk mengetahui apakah fenomena performa DPR yang tidak terlalu baik pada kondisi out-of-domain data juga terjadi pada bahasa Indonesia. Selain itu, penelitian ini mengevaluasi dua pendekatan untuk memperbaiki performa DPR dan mengatasi permasalahan keterbatasan data latih yakni pendekatan untuk memasukkan informasi exact-term matching kepada DPR dan pendekatan untuk mencoba melatih DPR pada beberapa jenis synthetic dataset berbahasa Indonesia. Hasil eksperimen menunjukkan bahwa performa DPR yang tidak terlalu baik pada data uji out-of-domain juga terjadi pada bahasa Indonesia yang ditunjukkan dengan skor evaluasi DPR yang relatif rendah terhadap skor evaluasi sparse retriever model. Selain itu, salah satu metode pemasukan informasi exact-term matching pada DPR yakni hybrid DPR-sparse retriever model menghasilkan skor BPref yang cenderung lebih baik dibandingkan skor BPref DPR pada seluruh eksperimen. Hasil pengujian pendekatan pelatihan DPR dengan synthetic dataset menunjukkan bahwa DPR yang dilatih dengan synthetic dataset pada penelitian ini menghasilkan skor BPref yang mengimbangi skor BPref DPR yang dilatih dengan data latih yang memang bisa digunakan untuk melatih DPR. Investigasi lebih lanjut pada hasil pengujian tersebut menunjukkan bahwa proses fine-tuning dan faktor domain data latih mungkin bisa mempengaruhi performa DPR. Selain itu, panjang token data latih dan faktor ukuran data latih tidak mempunyai korelasi terhadap performa DPR.
.....Researches on English data found that Dense Passage Retriever (DPR), a neural information retrieval model, has limitation on handling out-of-distribution data, including out-of-domain data. Information retrieval datasets in Indonesian that can be used for training DPR are quite scarce. All of those datasets are open-domain which may produce low model performance when the DPR tested on certain domain-specific dataset. This research compared the DPR performance to sparse retriever model performance to check whether DPR’s lack of performance when it’s tested on out-of-domain also can occur on Indonesian dataset. This research also tested two approaches that might improve DPR performance on that condition and also might overcome the training data scarcity problem that consist of methods to embed exact-term matching information into DPR and DPR fine-tuning on several Indonesian synthetic training datasets. The experiment result shows that DPR’s lack of performance on out-of-domain data also occur in Indonesian dataset which can be shown that all evaluation scores produced by DPR which is trained on out-of-domain training data are lower than any sparse retriever model’s evaluations scores. Result shows that hybrid DPR-sparse retriever model produced relatively higher BPref than DPR BPref. Additionally, result shows that

DPR which is fine-tuned on synthetic datasets that were used on this research produced relatively in-par BPref score in compare to BPref score that is produced by DPR which is fine-tuned on training datasets that are inherently can be used to fine-tune DPR. Further investigation on the synthetic dataset training approach results found that fine-tuning process and training data's domain may affect DPR performance. Additionally, training data token length and training data size don't have correlation with the DPR performance according to this experiment.