INTISARI

Large language models (LLM) atau model bahasa semakin marak digunakan sebagai sistem tanya jawab, terutama dalam kerangka *retrievel-augmented generation* (RAG) yang menggabungkan pengambilan (*retrieval*) dokumen dari basis data dengan pembangkitan jawaban (*generation*) oleh LLM. Namun, dokumen hasil pengambilan yang panjang membuat proses inferensi menjadi lebih mahal. Berbagai metode kompresi pascapengambilan telah dikembangkan, tetapi belum ada yang dievaluasi secara khusus untuk RAG berbahasa Indonesia. Penelitian ini menganalisis efektivitas tiga metode kompresi pascapengambilan RAG untuk kasus tanya jawab berbahasa Indonesia, yakni Corrective RAG (CRAG), Retrieve-Compress-Prepend (RECOMP), dan extreme-compression RAG (xRAG). RECOMP dilatih ulang dari model bahasa *pretrained*, xRAG dilatih dari dasar, sedangkan CRAG memanfaatkan model *embedding* multibahawa yang sudah tersedia untuk penyaringan segmen.

Evaluasi dilakukan pada 565 pasang data yang berisi pertanyaan, dokumen, dan jawaban, dengan menilai tingkat reduksi, ketepatan jawaban, dan efisiensi waktu eksekusi. Hasilnya ditemukan bahwa model RECOMP varian selektif mencapai tingkat reduksi tertinggi (99,7%), dan lalu diikuti oleh xRAG (99,4%). Namun, keduanya menunjukkan gejala *underfitting* dan banyak mengalami kegagalan dalam menyampaikan informasi penting dari dokumen awal sehingga tidaklah menjadi opsi yang cocok untuk sistem RAG. Sebaliknya, CRAG memiliki tingkat reduksi lebih rendah (38,1-79,3%), tetapi tetap mampu mempertahankan informasi-informasi penting sehingga menjadikannya metode dengan akurasi jawaban tertinggi dengan skor 51,3-67,8%. Akurasi jawaban ini jauh lebih baik daripada RECOMP varian nonselektif (31,9%) ataupun varian selektif (11,0%) serta xRAG (7,3%). CRAG@3 (CRAG yang mengambil 3 segmen teratas setelah pengurutan) terutama mencapai tingkat akurasi jawaban yang relatif sama dengan RAG konvensional (RAG tanpa kompresi pascapengambilan), dengan selisih hanya 0,2 poin persen saja. Waktu eksekusi CRAG ini juga tidak jauh berbeda dibanding RAG konvensional, menjadikannya metode paling stabil dan efektif dalam menyeimbangkan kompresi, ketepatan jawaban, dan efisiensi waktu eksekusi.

Kata kunci : *retrieval-augmented generation*, pengambilan informasi, tanya jawab, perangkuman, model bahasa

ABSTRACT

Large language models (LLMs) are increasingly used as question-answering systems, particularly within the retrieval-augmented generation (RAG) framework, which combines document retrieval from a database with answer generation by an LLM. However, long retrieved documents make the inference process more expensive. Various post-retrieval compression methods have been proposed, but none have been evaluated specifically for Indonesian-language RAG. This study analyzes the effectiveness of three post-retrieval compression methods for Indonesian question answering, namely Corrective RAG (CRAG), Retrieve-Compress-Prepend (RECOMP), and extreme-compression RAG (xRAG). RECOMP is fine-tuned from a pretrained language model, xRAG is trained from scratch, while CRAG leverages an existing multilingual embedding model for segment filtering.

Evaluation was conducted on a test set of 565 paired examples consisting of a question, documents, and an answer, considering reduction rate, answer accuracy, and execution-time efficiency. The results show that selective RECOMP achieves the highest reduction level (99.7%), followed by xRAG (99.4%). However, both exhibit signs of underfitting and frequently fail to convey essential information from the original documents, making them unsuitable for RAG systems. In contrast, CRAG achieves a lower reduction level (38.1–79.3%) but is able to preserve key information, resulting in the highest answer accuracy of 51.3–67.8%. This performance is substantially better than that of non-selective RECOMP (31.9%), selective RECOMP (11.0%), and xRAG (7.3%). CRAG@3, which is the CRAG variant that retains the top three ranked segments, achieves an accuracy level nearly identical to conventional RAG (RAG without postretrieval compression), differing by only 0.2 percentage points. Its execution time is also comparable to conventional RAG, making CRAG the most stable and effective method in balancing compression, answer accuracy, and inference efficiency.

Keywords : retrieval-augmented generation, information retrieval, question answering, summarization, language models