



PERANCANGAN DAN IMPLEMENTASI SISTEM TWO-STAGE RETRIEVAL BERBASIS BM25 DAN NEURAL RERANKING PADA DATASET SQUAD 1.1

Riva Malika Alea Firdaus ¹⁾

¹⁾ Sistem Informasi, Fakultas Teknik, Universitas Negeri Surabaya, Surabaya, Indonesia

Email: riva.22146@mhs.unesa.ac.id

Abstract

This study designs and implements a two-stage retrieval system combining BM25-based candidate generation with neural reranking, applied to the SQuAD 1.1 dataset reframed for passage retrieval. Development covers data preparation, corpus construction through selective chunking, BM25 candidate generation, and integration of a modular reranking component into an interactive user interface. System evaluation was conducted through black-box testing across multiple functional scenarios. Result show that the corpus and relevance mapping were successfully built, BM25 achieved high candidate recall, and all interface features functioned as designed. These findings confirm that the proposed pipeline can be implemented end-to-end into a functional application, providing a validated basis for further quantitative evaluation of reranking performance.

Keywords: Two-Stage Retrieval; BM25; Neural Reranking; Sistem Information Retrieval; SQuAD 1.1

Abstrak

Penelitian ini merancang dan mengimplementasikan sistem two-stage retrieval yang menggabungkan candidate generation berbasis BM25 dengan neural reranking, diterapkan pada dataset SQuAD 1.1 yang disesuaikan ke skema passage retrieval. Pengembangan mencakup persiapan data, pembangunan korpus melalui selective chunking, candidate generation BM25, serta integrasi komponen reranking modular ke dalam antarmuka pengguna interaktif. Evaluasi sistem dilakukan melalui black-box testing pada skenario fungsional. Hasil menunjukkan korpus dan pemetaan relevansi berhasil dibangun, BM25 mencapai candidate recall tinggi, dan seluruh antarmuka berfungsi sesuai rancangan. Temuan ini membuktikan bahwa pipeline yang diusulkan dapat diimplementasikan secara menyeluruh menjadi aplikasi fungsional, menjadi dasar tervalidasi untuk evaluasi kinerja reranking secara kuantitatif pada penelitian selanjutnya.

Kata Kunci: Two-Stage Retrieval; BM25; Neural Reranking; Sistem Information Retrieval; SQuAD 1.1



PENDAHULUAN

Pertumbuhan data teks tidak terstruktur yang pesat menjadikan sistem information retrieval (IR) sebagai komponen penting dalam berbagai aplikasi berbasis teks (Hambarde & Proença, 2023). Pendekatan leksikal seperti BM25 masih banyak digunakan karena efisiensinya, namun memiliki keterbatasan dalam menangkap kesesuai makna ketika terdapat perbedaan kata antara query dan dokumen meskipun relevan secara semantik (Chakraborty et al., 2023; Guo et al., 2022) untuk mengatasi hal ini, berkembang arsitektur two-stage retrieval yang memisahkan proses pencatatan menjadi tahap candidate generation berbasis BM25 dan tahap neural reranking menggunakan model transformer, baik melalui mekanisme full-interaction maupun late-interaction yang menghitung representasi query dan passage secara terpisah sebelum digabungkan melalui operasi pencocokan seperti MaxSim (Ahmad et al., 2019; Chakraborty et al., 2023; Khattab & Zaharia, 2020)

Meskipun penelitian neural ranking terus berkembang, aspek rekayasa end-to-end-nya mulai dari penyusunan korpus dari data mentah, penanganan dokumen panjang melalui strategi chunking, penentuan ukuran sampel kandidat yang teruji secara statistik, hingga integrasi ke dalam sistem yang dapat digunakan langsung, relatif belum banyak dibahas secara rinci, padahal penyusunan korpus yang tepat sangat menentukan kualitas retrieval (Kiss et al., 2025). Berdasarkan hal tersebut, penelitian ini merancang dan mengimplementasikan sistem two-stage retrieval yang mengintegrasikan candidate generation

berbasis BM25 dengan modul neural reranking modular pendukung lebih dari satu mekanisme, menggunakan studi kasus dataset SQuAD 1.1 yang disesuaikan ke skema passage retrieval, mulai dari persiapan data, pembangunan korpus dan qrels, candidate generation tervalidasi statistik, hingga integrasi ke antarmuka pengguna interaktif yang diuji secara fungsional.

METODE PENELITIAN

Penelitian ini menggunakan dataset SQuAD 1.1 (87.599 entri train, 10.570 entri validation) yang diperoleh melalui Hugging Face Dataset. Data diproses melalui pemilihan kolom penting, transformasi jawaban, pembersihan nilai kosong/duplikat, serta validasi jawaban terhadap context, sehingga diperoleh 87.592 entri train dan 10.568 entri validation yang siap diolah.

Context yang telah bersih dideduplikasi menjadi korpus passage, dengan strategi selective chunking berdasarkan distribusi panjang token (threshold 180 token) menggunakan sliding window 150 token dan stride 75 token, mengacu pada pendekatan chunking teks panjang (Kiss et al., 2025) sehingga menghasilkan 26.727 passage (train) dan 3.047 passage (validation). Pemetaan relevansi (qrels) dibentuk melalui pencocokan rentang token jawaban terhadap passage. kandidat awal passage dihasilkan menggunakan BM25 pada sampel 1.000 query per split (ukuran divalidasi rumus Cochran) dengan 20 kandidat teratas untuk train dan 50 kandidat teratas untuk validation.

Data latih dibentuk dari query yang passage positifnya ditemukan dalam kandidat BM25 dengan satu



hard negative per query (Karpukhin et al., 2020). Modul reranking dirancang modular untuk mendukung mekanisme full-interaction dan late-interaction (Khattab & Zaharia, 2020) yang dilatih menggunakan backbone bert-base-uncased pada GPU Tesla T4. Model yang telah dilatih diintegrasikan ke dalam pipeline retrieval-reranking dan antarmuka pengguna berbasis web, yang memungkinkan pengguna memasukkan query, mengatur jumlah kandidat (top-k), serta melihat hasil pemeringkatan pada dua tab sesuai mekanisme yang diimplementasikan. Sistem diuji menggunakan black-box testing pada 15 skenario, mencakup kondisi normal dan tidak normal, untuk memastikan seluruh fitur berjalan sesuai rancangan.

HASIL DAN PEMBAHASAN

Tahap candidate generation BM25 menghasilkan candidate recall sebesar 75,5% pada data train (755 dari 1.000 query awal berhasil menemukan positive passage dalam kandidat top-20) dan rata-rata diatas 85% pada data validation (contoh interaksi pertama: 877 dari 1.000 query, candidate recall 87,7%, top-50). Tingginya candidate recall pada validation menunjukkan tahap retrieval pertama mampu menyediakan kandidat yang representatif untuk tahap reranking, sementara sebagian kecil kegagalan diduga disebabkan vocabulary mismatch pada pendekatan leksikal BM25.

Model reranking yang telah dilatih berhasil diintegrasikan ke dalam antarmuka pengguna berbasis web (GAMBAR 1), yang menyediakan kolom input query, pengaturan jumlah kandidat (top-k), serta dua tab hasil

pemeringkatan ulang sesuai mekanisme reranking yang diimplementasikan, lengkap dengan info box yang menampilkan status positive passage dan indikator efisiensi sistem sebagai bagian dari fitur transparansi antarmuka. Pengujian fungsional menggunakan metode black-box testing pada 15 skenario, mencakup kondisi normal (quer valid, konfigurasi top-k wajar) maupun tidak normal (query kosong, query diluar dataset, nilai top-k ekstrem), menunjukkan bahwa seluruh skenario (15/15) dinyatakan lulus, membuktikan seluruh fitur sistem berfungsi sesuai rancangan tanpa kegagalan.



Gambar 1. Antarmuka input dan konfigurasi sistem

KESIMPULAN

Penelitian ini berhasil merancang dan mengimplementasikan sistem two-stage retrieval yang mengintegrasikan tahap candidate generation berbasis BM25 dengan modul neural reranking modular yang mendukung mekanisme full-interaction dan late-interaction, menggunakan studi kasus dataset SQuAD 1.1. seluruh tahapan pipeline, mulai dari persiapan data, pembangunan korpus dan qrels melalui selective chunking, hingga candidate generation BM25, berjalan dengan candidate recall diatas 75% pada data train dan diatas 85%



pada data validation, menunjukkan tahap retrieval mampu menyediakan bagi tahap reranking. Sistem yang dihasilkan berhasil diintegrasikan ke dalam antarmuka pengguna interaktif, dan seluruh 15 skenario pengujian black-box dinyatakan lulus, membuktikan bahwa seluruh fitur berfungsi sesuai rancangan. Keberhasilan mengintegrasikan dua mekanisme reranking ke dalam satu pipeline yang sama menunjukkan bahwa arsitektur sistem yang dirancang bersifat fleksibel dan dapat diperluas untuk mekanisme reranking lain dimasa mendatang. Penelitian selanjutnya dapat memanfaatkan sistem ini sebagai fondasi untuk melakukan evaluasi kinerja reranking secara kuantitatif, termasuk pengujian pada dataset lain serta penambahan mekanisme reranking baru.

DAFTAR PUSTAKA

Ahmad, A., Constant, N., Yang, Y., & Cer, D. (2019). ReQA: An Evaluation for End-to-End Answer Retrieval Models. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 137–146. <https://doi.org/10.18653/v1/D19-5819>

Chakraborty, T., La Gatta, V., Moscato, V., & Sperli, G. (2023). Information retrieval algorithms and neural ranking models to detect previously fact-checked information. *Neurocomputing*, 557. <https://doi.org/10.1016/j.neucom.2023.126680>

Guo, J., Cai, Y., Fan, Y., Sun, F., Zhang, R., & Cheng, X. (2022). Semantic Models for the First-Stage Retrieval: A Comprehensive Review. *ACM Transactions on*

Information Systems, 40(4), 1–42. <https://doi.org/10.1145/3486250>

Hambarde, K. A., & Proença, H. (2023). Information Retrieval: Recent Advances and Beyond. *IEEE Access*, 11, 76581–76604. <https://doi.org/10.1109/ACCESS.2023.3295776>

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>

Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 39–48. <https://doi.org/10.1145/3397271.3401075>

Kiss, C., Nagy, M., & Szilágyi, P. (2025). Max–Min semantic chunking of documents for RAG application. *Discover Computing*, 28(1). <https://doi.org/10.1007/s10791-025-09638-7>