

REVIEWER'S REPORT

Manuscript No.: **IJAR-58896**

Title: Query-Type Sensitivity and Selective Benefits of Hybrid Lexical-Semantic Retrieval: An Experimental Analysis with BM25 and Multilingual SBERT.

Recommendation:

Accept as it is

Accept after minor revision.....

Accept after major revision ...✓.....

Do not accept (*Reasons below*)

Rating	Excel.	Good	Fair	Poor
Originality		✓		
Techn. Quality		✓		
Clarity		✓		
Significance		✓		

Reviewer's ID: JPR- **098**

Detailed Reviewer's Report

Decision: Accept after Major Revision

Major Revisions

1. Strengthen the experimental dataset and external validity.

The evaluation is based on only 50 queries and 19 documents (950 query-document pairs), which is relatively small for drawing broader conclusions about hybrid information-retrieval systems. The authors should provide a stronger justification for the benchmark size and, if possible, validate the proposed findings on a larger and more diverse document collection and query set.

2. Provide greater detail about the benchmark construction and relevance judgments.

The manuscript states that there are 73 validated PRIMARY positive relevance judgments, but the procedure used to construct, validate, and assign these judgments is insufficiently detailed. The authors should clearly explain who generated the judgments, the annotation criteria, whether multiple annotators were involved, and how disagreements were resolved.

3. Address the lack of statistical significance testing.

The reported improvements are explicitly descriptive, and no inferential statistical tests are conducted within query types. Statistical significance

REVIEWER'S REPORT

and confidence intervals should be considered for the overall comparison, and appropriate paired tests or bootstrap-based confidence intervals should be reported where applicable. This is particularly important for claims concerning the superiority of the hybrid approach.

4. Expand the comparison with contemporary retrieval methods.

The experimental comparison is limited to BM25, one multilingual SBERT configuration, RRF, and the proposed fixed-weight fusion. The authors should consider including stronger modern dense or sparse-dense retrieval baselines, or provide a clear justification for excluding methods such as more recent embedding models, learned sparse retrieval, or late-interaction approaches.

5. Clarify the rationale for selecting $\alpha = 0.5$ as the primary hybrid configuration.

Although the paper correctly states that $\alpha=0.5$ is not claimed to be universally optimal, the selection requires stronger methodological justification. Since the manuscript evaluates an α -grid from 0 to 1, the authors should explain whether $\alpha=0.5$ was predetermined independently of the results and should clearly distinguish exploratory weight sensitivity from model selection.

6. Investigate the effect of score normalization more rigorously.

The hybrid method depends on query-wise min-max normalization before combining BM25 and SBERT scores. Because normalization can substantially influence fusion behavior, the authors should discuss why min-max normalization was selected and, preferably, evaluate whether alternative normalization or calibration methods change the reported conclusions.

7. Improve the analysis of query-type imbalance.

The query taxonomy is substantially unbalanced, with only 1 DEFINITIONAL query and 2 COMPARATIVE queries, compared with 16 MULTI_CONSTRAINT and 13 DIRECT queries. Therefore, conclusions concerning individual query types should be presented very cautiously. The authors should consider sensitivity analysis or confidence intervals and avoid interpreting extremely small subgroups as representative categories.

8. Strengthen the explanation of the apparent PARAPHRASE advantage.

The largest improvement occurs for PARAPHRASE queries ($\Delta =$

REVIEWER'S REPORT

+0.2054), and the manuscript interprets this as evidence that lexical–semantic fusion is particularly useful for alternative wording. However, this interpretation remains observational. The authors should provide additional query-level examples or qualitative analysis demonstrating why the hybrid approach succeeds on these queries and whether the improvement is actually attributable to reduced lexical overlap.

9. Provide more comprehensive error analysis.

The manuscript reports that the hybrid improves 10 queries, ties 38, and performs worse on 2 queries relative to BM25. The two degraded queries deserve detailed analysis. The authors should identify their characteristics and explain why semantic fusion may negatively affect ranking in these cases. Such analysis would significantly strengthen the paper's central claim regarding selective benefit.

10. Moderate and better contextualize the general conclusions.

The conclusions are appropriately bounded, but the paper should consistently emphasize that the findings are benchmark-specific and based on one multilingual SBERT encoder, one RRF configuration, one normalization strategy, and a predefined fusion grid. Additional validation is required before extending the findings to general information-retrieval systems or production environments.

Minor Revisions

1. Correct typographical and formatting inconsistencies.

Several formatting issues are visible, including spacing around equations, table headings, section labels, and words such as “Resultats.” A thorough language and formatting check is recommended.

2. Improve the presentation of the keywords.

The manuscript currently shows an unusual “Key words: -” formatting before the actual keywords. This should be corrected and standardized.

3. Standardize table formatting.

Tables 1–4 should use consistent spacing, alignment, decimal precision, headings, and notation. Table titles should also follow the same formatting convention throughout the manuscript.

4. Clarify notation in the mathematical equations.

The variables used in the min–max normalization, hybrid score, RRF and nDCG equations should be explicitly defined immediately after each

REVIEWER'S REPORT

equation. Equation formatting should also be checked for missing subscripts, parentheses, and mathematical symbols.

5. Correct the RRF formulation and notation presentation.

The RRF equation should be typeset clearly and the meaning of the summation over retrieval systems should be explicitly stated. The relationship between the chosen parameter ($k=60$) and the ranking formulation should also be explained.

6. Improve figure readability.

Figures 1–3 should contain sufficiently large labels, legends, axis descriptions, and readable numerical values. The figures should remain interpretable when printed or viewed at normal manuscript scale.

7. Define abbreviations at first use.

Terms such as BM25, SBERT, RRF, nDCG, DCG and IDCG should be defined at their first occurrence, even if some are widely recognized within information retrieval.

8. Improve consistency in terminology.

The manuscript alternates between terms such as “hybrid,” “hybrid retrieval,” “lexical–semantic fusion,” and “fixed $\alpha=0.5$ hybrid.” The terminology should be standardized where appropriate.

9. Strengthen the reference coverage with recent studies.

The reference list contains important foundational and established retrieval studies, but the literature discussion could be strengthened by incorporating additional recent work on multilingual retrieval, hybrid retrieval, dense retrieval, learned sparse retrieval, and adaptive fusion. The current reference list includes major works from 2009–2024.

10. Revise the manuscript for grammatical and stylistic consistency.

Although the manuscript is generally understandable, several sentences are dense and could be simplified. The authors should perform a final proofreading pass covering grammar, punctuation, capitalization, spacing, section numbering, and consistency of scientific terminology.