ChatGPT

Hybrid Retrieval: Lexical (BM25) vs Embedding Strategies

Information retrieval can combine **sparse lexical matching** (e.g. BM25) with **dense semantic matching** (e.g. word2vec/BERT embeddings) to improve accuracy. In one strategy, the system first retrieves a broad set of candidates by semantic similarity (cosine on embeddings) and then refines their ranking using BM25 ("embedding-first, then lexical re-rank"). In the other, it retrieves by BM25 and then re-ranks those by vector similarity ("lexical-first, then semantic re-rank"). Intuitively, embedding-based recall can pull in documents with query synonyms or related concepts, while BM25 ensures matching key terms precisely [1] [2].

Embedding-First (Vector → BM25 Re-rank)

This approach computes an initial cosine-similarity ranking between the query vector and document vectors (using word2vec, sentence embeddings, etc.), then re-scores those top-K hits with BM25. Theoretically, it **maximizes recall of semantically related documents** even if they share few exact terms, then uses BM25 to boost those with actual term overlap. For example, Cao *et al.* (2024) propose a "dense-to-hybrid" two-stage scheme: Stage 1 is a purely dense (embedding) search; Stage 2 refines the results by adding BM25-based scoring [3] [4]. This can capture synonyms and related concepts that BM25 alone misses. Empirically, embedding-first methods often find unique relevant documents: one analysis showed that a purely semantic search (top-1000 by cosine) uncovered on average 8 relevant docs per query that BM25 missed (on Robust04 news data) [5].

However, these semantic-only rankings can also introduce noise (irrelevant semantic matches) [1]. In practice, purely embedding retrieval often underperforms BM25 in precision or standard IR metrics. For example, Galke *et al.* (2017) found that an IDF-weighted word-vector centroid model (embedding-only) matched BM25 on short-text fields but was weaker on longer documents [6]. Similarly, Zhao *et al.* (2017) showed that combining a BM25 baseline with a Word-Mover's-Distance–based semantic score significantly increased NDCG (e.g. +23% at rank 20) [7]. These results imply that embedding-first retrieval can boost ranking when combined properly, but often needs BM25's precision.

Lexical-First (BM25 → Vector Re-rank)

In contrast, the BM25-first strategy starts with exact-term retrieval, then reorders those candidates by semantic similarity. The intuition is that BM25 gives a strong, precise "seed set" (precise lexical matches), and embedding re-ranking brings semantic nuance without losing exact matches. For instance, Yan *et al.* (2018) likened BM25 to "precise memory" and neural embeddings to "associative memory" of the corpus [2]. They implemented a sequential scheme: first retrieve seeds by BM25, then expand/rerank them by finding nearest neighbors in vector space.

Empirical studies often find **BM25-first yields higher baseline precision and recall**. Yan *et al.* reported that BM25 alone already had strong recall@1000, and "neural (semantic) alone" was significantly worse [1]. Their combined methods (BM25 seeds plus vector expansion) consistently improved recall: the sequential BM25→vector scheme (SeqSearch) outperformed a parallel merge strategy (and pure BM25) on recall [1]. On standard news corpora (Robust04, WT2G), BM25→embedding boosting lifted recall@1000 by a few points (e.g. from 68% to ~72%) [1]. It also modestly increased MAP and NDCG. Notably, Yan *et al.* found that even expanding only 25% of the seeds by vector neighbors gave nearly the same benefit [8], suggesting most gains came from a subset of high-quality seeds.

Empirical Comparisons

- **Retrieval Effectiveness:** In controlled experiments, **hybrid retrieval (combining both signals) consistently beat either alone**. For example, Rayo *et al.* (2025) in a regulatory-document task report BM25 recall@10=0.761, semantic-only=0.810, and a *fused* BM25+semantic approach =0.833. Similarly, MAP@10 rose from ~0.624 (BM25) or 0.629 (semantics) to 0.702 for the hybrid [9]. This demonstrates that lexical and semantic methods capture complementary relevance. [9] [7]

- **NDCG/Precision:** Zhao *et al.* (2017) added a vector-based semantic score (using Word Mover's Distance on titles) to BM25. They observed **large gains in ranking quality**: hybrid NDCG@5/10/20 were ~23–25% higher than BM25 alone [7]. This shows embedding signals can improve top-ranked accuracy if properly blended. Conversely, embedding-only models often underperform: the same study saw semantic-only precision (and NDCG) consistently below BM25's [10].

- **Document/Query Length:** Galke *et al.* (2017) found embedding methods are **particularly effective for short texts**. In their benchmarks, an IDF-weighted word-vector average matched or beat TF-IDF/BM25 on *titles* (short fields), but fell behind on longer text [6]. Intuitively, embeddings shine when the query/document are too short to give reliable term statistics. In contrast, for longer documents (full text), BM25's exact term signals dominate.

Domain-Specific Findings

Many recent works confirm these trends in specialized corpora. In a biomedical/clinical setting, combining BM25 with semantic techniques (query expansion or embedding distances) significantly improved recall and ranking [10] [7]. In regulatory/legal domains, Rayo *et al.*'s fine-tuned semantic model combined with BM25 outperformed each alone [9]. However, as Yan *et al.* note, the **scope for improvement depends on how strong the BM25 baseline already is**. For datasets where BM25 already achieves high recall (e.g. well-tuned, jargon-rich corpora), adding vectors yields smaller gains [11]. In queries/domains with vocabulary mismatch or many synonyms, vector-first or hybrid approaches show larger advantages.

Which Strategy When?

- **Embedding-First (VSM → BM25)** tends to **maximize recall**, especially useful when queries are short or have many semantically-related terms. It can retrieve relevant documents BM25 misses (improving coverage), at the cost of more noise. It also parallels modern "dense retrieval" pipelines. Cao *et al.*'s dense-first strategy is one example of leveraging embeddings for initial recall [3] [4]. Use this when term mismatch is a major issue.

- **Lexical-First (BM25 → VSM)** generally gives **higher precision and more stable baseline**, since BM25 anchors on exact terms. Re-ranking by embeddings then refines the order (giving synonyms a slight boost) without losing core hits. Empirically, this often outperforms doing the opposite: Yan *et al.* found BM25-first (their SeqSearch) beat embedding-first (ParSearch) on recall

[12]. Use this when exact term recall is critical or BM25 is already strong (e.g. technical queries, long documents).

- **Hybrid Scoring** (combining scores rather than strict two-stage) is also popular: e.g. linear fusion of BM25 and embedding scores yielded the strongest results in many studies [9] [7]. In practice, a weighted sum (or learn-to-rank) that balances BM25 and cosine often yields robust performance across domains.

In summary, combining BM25 with vector-space matching almost always helps. Studies consistently report that a hybrid approach outperforms either alone on standard metrics like Recall, MAP, and NDCG [9] [7]. The optimal pipeline depends on the task: if recall and semantic coverage are paramount, starting with embedding-based recall can help; if precision and exact matching are key, starting with BM25 is safer. Often the best results come from **fusion** or **re-ranking**: for example, BM25-retrieved candidates re-ordered by embedding similarity (or vice versa) achieves strong gains [1] [9].

Key Takeaways: BM25 is a robust baseline (lexical matches) while embeddings add semantic recall. Embedding-first retrieval expands coverage, but adds noise, whereas BM25-first preserves precision. Empirical results (MAP, NDCG) usually favor **hybrid methods** that weight both signals [9] [7]. The relative benefit of each strategy depends on query length, domain specificity, and how strong the lexical baseline already is. In practice, tuning the fusion or re-ranking order for your domain (potentially via held-out data) yields the best performance [13] [7].

Sources: Comprehensive IR studies on lexical vs. semantic retrieval [1] [9] [6] [7]; hybrid search analyses and benchmarks [2] [3] [4] [5]. (Results drawn from IR literature reporting MAP/NDCG/Recall improvements for various BM25/embed strategies.)

---

[1] [2] [5] [8] [11] [12] [13] Beyond Precision: A Study on Recall of Initial Retrieval with Neural Representations
https://arxiv.org/pdf/1806.10869

[3] [4] Efficient and Effective Retrieval of Dense-Sparse Hybrid Vectors using Graph-based Approximate Nearest Neighbor Search
https://arxiv.org/html/2410.20381v1

[6] Word Embeddings for Similarity Scoring in Practical Information Retrieval
https://www.zbw.eu/fileadmin/pdf/forschung/2017-colloquium-galke-word-embeddings.pdf

[7] [10] arxiv.org
https://arxiv.org/pdf/1608.01972

[9] A Hybrid Approach to Information Retrieval and Answer Generation for Regulatory Texts
https://arxiv.org/html/2502.16767v1