Comparative analysis of neural IR models

1. STATE OF THE ART IN NEURAL IR

Context: What is Information Retrieval?

Information Retrieval (IR) is the task of finding relevant documents or passages in response to a user query. Traditional approaches (like BM25) rely on term matching—documents scoring higher when they share many words with the query. Neural IR methods instead learn representations that capture semantic similarity beyond surface-level term overlap, enabling more flexible matching of paraphrases and related concepts.

The Neural IR Landscape:

Neural models for IR differ primarily in *how* and *when* they compute query-document similarity:

Dual-encoder architectures compute query and document embeddings independently, then compare them. This encompasses both dense approaches (e.g., Dense Passage Retrieval) that map text into continuous vectors, and sparse approaches (e.g., SPLADE) that learn which terms are important. The advantage is efficiency: embeddings can be pre-computed and indexed, enabling fast retrieval at scale. The tradeoff is that independent encoding misses direct interaction between query and document text.

Early interaction models (cross-encoders) jointly process the query and document together before computing a relevance score (e.g., MonoBERT). This allows richer interaction patterns but is computationally expensive, making them practical only for re-ranking a smaller set of candidate documents.

Late interaction models (e.g., ColBERT) compute embeddings independently like dual-encoders but allow token-level interaction during scoring. This balances efficiency with the ability to capture fine-grained relevance signals better than dense dual-encoders alone.

Large Language Models (LLMs) can be applied to ranking through prompting and in-context learning, without task-specific training. Models like RankGPT demonstrate that LLMs can infer relevance from demonstrations, opening questions about how pre-training scale and instruction-tuning affect IR performance.

The field is characterized by a fundamental tradeoff: efficiency versus effectiveness. Current research investigates which model families best balance this tradeoff across different scenarios and resource constraints.

2. OBJECTIVES

Main objective: Conduct a systematic comparative analysis of neural IR models across the spectrum—from sparse and dense dual-encoders through late and early interaction models to LLM-based approaches. The intern will:

- Implement or integrate multiple representative models from each family
- Evaluate performance on standard benchmarks (TREC, MS MARCO, BEIR, or others)
- Analyze tradeoffs: accuracy, latency, memory footprint, scalability, indexing requirements
- Generate detailed comparison reports with visualizations and insights on which model families excel in different scenarios (retrieval depth, domain specificity, query complexity)
- Document reproducible evaluation protocols

Secondary objective (time permitting): Propose and prototype novel approaches that could bridge identified gaps or combine strengths of different model families. Potential directions include: hybrid sparse-dense retrieval, efficient late interaction variants, or improved prompting strategies for LLM-based ranking.

REFERENCES

Dual Encoders:

- RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering (Ding et al., 2020)
- Towards Effective and Efficient Sparse Neural Information Retrieval (Formal et al., 2024)

Cross-encoders (Early Interaction):

- Nogueira, R., & Cho, K. (2019). "Passage Re-ranking with BERT"
- Formal, R., Piwowarski, B., & Clinchant, S. (2021). "Towards a Better Metric for Evaluating Question Generation Systems"

Late Interaction Models:

- ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction (Santhanam et al., 2021)

LLM-based Ranking:

- RankZephyr: Effective and Robust Zero-Shot Listwise Reranking is a Breeze! (Pradeep et al., 2023)

Contact

Benjamin Piwowarski benjamin.piwowarski@cnrs.fr