Patent Reranking with Dense and Cross Encoders

Md Naim Hassan Saykat

Abstract

This project explores reranking methods for patent retrieval by combining dense retrieval with transformer-based cross-encoders. The pipeline integrates traditional TF–IDF baselines, dense embeddings (BGE), and cross-encoder re-ranking, further enhanced through Reciprocal Rank Fusion (RRF). Experiments on a patent dataset demonstrate improvements in retrieval quality measured by Mean Average Precision (MAP), Recall@k, and Mean Rank.

Keywords: Information Retrieval, Dense Retrieval, Cross-Encoder, Reranking, Patents, Reciprocal Rank Fusion.

1 Introduction

Patent search is a critical task in information retrieval due to the large volume of documents and the need for precise relevance judgments. Classical term-based methods such as TF–IDF offer efficiency but suffer from lexical mismatch. Neural embedding models such as Dense Passage Retrieval (DPR) [1] and transformer-based re-rankers like BERT [2] have emerged as powerful solutions for capturing semantic relevance. Recent advances in open-source toolkits such as Hugging Face Transformers [3] have made it easier to integrate dense retrievers and cross-encoders into end-to-end IR pipelines. This project explores how these approaches can be combined and adapted for the patent retrieval domain.

This work implements and evaluates a reranking pipeline for patents, using a combination of:

- TF–IDF (baseline),
- Dense retrieval with BGE embeddings,
- Cross-encoder re-ranking with a fine-tuned BERT model,
- Reciprocal Rank Fusion (RRF) for combining results.

2 Methodology

2.1 Dataset

The dataset consists of patent queries, relevance mappings, and document features:

- `train_queries.json` – training queries
- `test_queries.json` – test queries used for evaluation
- `train_gold_mapping.json` – relevance judgments
- `documents_features.json` – patent document features

Due to GitHub file-size limits, the dataset is hosted externally on Google Drive.[1]

[1] https://drive.google.com/drive/folders/1Oy4Gp1KVO__01JnX1V4JuZ0zy7jlK78J?usp=sharing

2.2 Models

- **TF–IDF Baseline:** Sparse vector retrieval for initial ranking.
- **Dense Retriever (BGE):** Embedding-based retrieval producing dense representations of queries and documents, inspired by prior dense retrieval work [1].
- **Cross-Encoder:** A BERT-based pairwise scoring model trained on query–document pairs, following the passage re-ranking paradigm [2].
- **Ensemble (RRF):** Reciprocal Rank Fusion combining dense retriever and cross-encoder outputs.

2.3 Evaluation Metrics

We report:

- Mean Average Precision (MAP),
- Recall@10,
- Mean Rank.

Implementation of models and training pipelines was carried out using the Hugging Face Transformers library [3].

3 Experiments

3.1 Training Setup

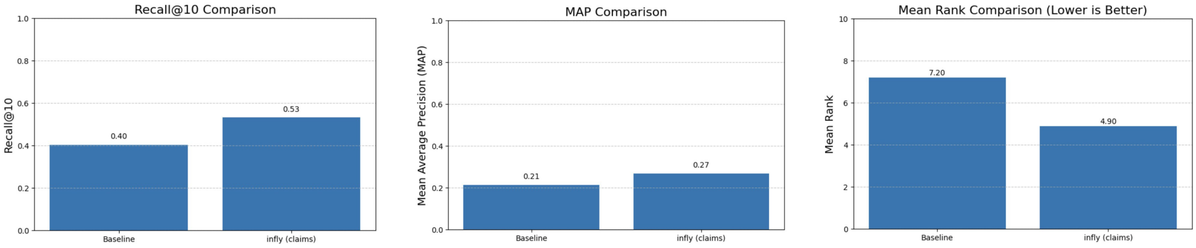
The cross-encoder was trained using the `cross_encoder_reranking_train.py` script with 3 epochs and batch size 16. Dense embeddings were precomputed using the BGE model. Evaluation scripts computed MAP, Recall@10, and Mean Rank.

3.2 Results

Table 1 shows the retrieval performance across models. We also visualize the trends in Recall@10, MAP, and Mean Rank in Figure 1 for clearer comparison.

Model	**MAP**	**Recall@10**	**Mean Rank**
Dense Retriever (infly/inf-retriever-v1-1.5b)	0.2140	0.4046	7.20
Cross-Encoder Re-ranker	0.2424	0.4426	6.35
Ensemble (Dense + Cross-Encoder, RRF)	**0.2681**	**0.5321**	**4.90**

Table 1: Comparison of retrieval models on the patent dataset.



(a) Recall@10 Comparison (b) MAP Comparison (c) Mean Rank Comparison

Figure 1: Performance comparison across metrics: (a) Recall@10, (b) MAP, and (c) Mean Rank.

4 Discussion

The results in Table 1 and Figure 1 highlight clear performance gains when moving from sparse to neural methods. The dense retriever improves over the baseline with a MAP of 0.2140 and Recall@10 of 0.4046, demonstrating the effectiveness of dense embeddings for semantic matching. The cross-encoder re-ranker further boosts performance, achieving MAP 0.2424 and Recall@10 0.4426, by explicitly modeling query–document interactions. Finally, the ensemble approach with Reciprocal Rank Fusion (RRF) yields the strongest results across all metrics (MAP 0.2681, Recall@10 0.5321, Mean Rank 4.90), confirming that hybrid methods effectively leverage complementary strengths. These findings indicate that while dense retrieval provides strong semantic coverage, combining it with re-ranking strategies maximizes both recall and ranking quality.

5 Conclusion

We developed and evaluated a patent reranking pipeline combining TF–IDF, dense retrievers, and cross-encoders, with further improvements from Reciprocal Rank Fusion (RRF). The results show that dense retrievers provide strong semantic matching, while cross-encoders further improve ranking by modeling fine-grained query–document interactions. The ensemble approach achieved the best overall performance (MAP 0.2681, Recall@10 0.5321, Mean Rank 4.90), confirming the effectiveness of hybrid pipelines that integrate retrieval and re-ranking strategies.

This study highlights the value of combining complementary IR techniques to address patent search, where both semantic coverage and precise ranking are crucial. Future work could explore larger transformer-based re-rankers (e.g., DeBERTa, Longformer), domain-specific pretraining, and metrics such as nDCG and precision@k to further validate performance.

References

[1] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics, 2020.

[2] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.

[3] Thomas Wolf, Lysandre Debut, Victor Sanh, et al. Transformers: State-of-the-art natural language processing, 2020.