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) [**?**] 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__O1JnX1V4JuZOzy7jlK78J?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 methods.

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.

4 Discussion

The results show clear improvements when moving from sparse (TF–IDF) to dense retrieval. The cross-encoder adds significant gains by capturing query–document interactions. Finally, the ensemble approach (RRF) yields the best performance across all metrics, confirming that hybrid methods effectively leverage complementary strengths.

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. The approach demonstrates the effectiveness of hybrid IR pipelines for challenging domains like patents. Future work could explore larger transformer re-rankers, domain-specific pretraining, and additional evaluation metrics such as nDCG.

References

[1] Vladimir Karpukhin, Barlas Oğuz, 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.