Rank Fusion in Neural Ranking Model

Delft University of Technology
Author: Gayeon Jee (G.Jee@tudelft.nl)
Supervisor: Jurek Leonhardht
Responsible Professor: Avishek Anand

1 Background

Information Retrieval (IR) is about retreiving relevant documents (candidates) given a query and ranking them by relevance. Some ranking model types:

- **Sparse/Lexical Models** - retrieval by term matching; simple but generally worse than dense models as it misses context
- **Dense/Semantic Models** - retrieval by creating embeddings and evaluate similarity based on distance; captures contextual information but requires a lot of time and resource
- **Hybrid Models** - combines results of sparse and dense models
- **Retrieve-and-rerank** - retrieves candidates using sparse models then reranks using dense models

Retrieve-and-rerank with Fast-Forward Indexes is an approach to devise an efficient neural ranking model motivated by [1] (Fig. 1).

Rank Fusion Functions merge the result of lexical and semantic scores to rerank the documents.

Depending on the domain, certain scores information are more useful than the other. Rank fusion functions control how and to what extent each score influences the final rank.

Different types of rank fusion functions:

- **Parametric** vs. **Non-parametric**
- **Score-based** vs. **Rank-based**
- **Voting Rule**

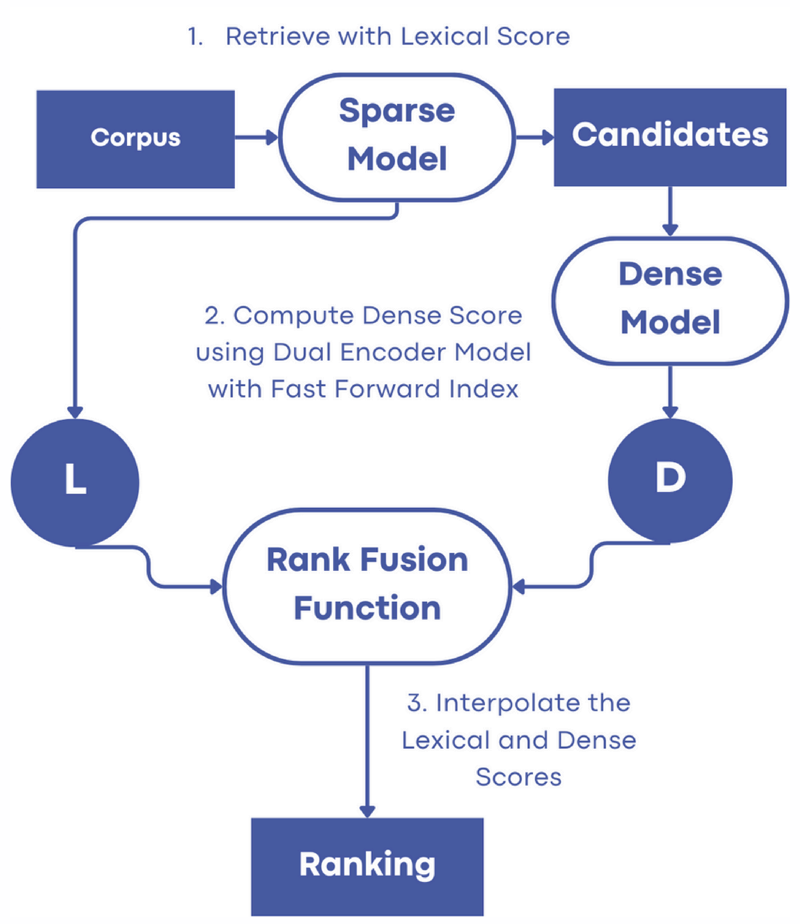


Figure 1: Fast-Forward Index Framework

2 Research Question

"What is the impact of the rank fusion function?"

1. How does the **rankings change in relation to semantic and lexical ranks** using different rank fusion functions?
2. How does using different rank fusion functions impact the **ranking effectiveness** in different domains?
3. How does using different rank fusion functions impact the **latency** in different domains?

3 Methodology

The general setup and variables of the experiment elaborated:

- Models chosen: BM25, TF-IDF-based sparse model, and TCT-ColBERT, BERT-based dual encoder dense model
- Dataset for evaluation [2]:

DATASET	DOMAIN	TASK
MS MACRO PsgTREC DL '19	Misc	Passage Retrieval
MS MACRO PsgTREC DL '20	Misc	Passage Retrieval
BEIR FiQA-2018	Finance	Question Answering
BEIR NFCorpus	Bio-Medical	Bio-Medical IR
BEIR QUORA	Quora	Duplicate Question Retrieval
BEIR DBPedia	Wikipedia	Entity-Retrieval
BEIR FEVER	Wikipedia	Fact Checking
BEIR ArguAna*	Misc	Argument Retrieval
BEIR CQADupStack (English)*	Misc	Argument Retrieval
BEIR Scifact*	Scientific	Fact Checking
BEIR SCIDOCS*	StackEx.	Duplicate-Qeustion Retrieval

- *These datasets do not have a dev set used for validation so tested in zero-shot fashion
- Chosen various types of rank fusion functions
  - Score-based fusion - inputs the scores directly
    - Convex Rank Fusion - parametric; identity, min-max normalization, z-score normalization [3]

$$f_{CONVEX}(q,d) = \alpha f_{SEM}(q,d) + (1-\alpha) f_{LEX}(q,d)$$

  - Rank-based fusion - inputs the ranks
    - Reciprocal Rank Fusion - parametric (2 parameters) [3]

$$f_{RRF}(q,d) = \frac{1}{\alpha + r_{LEX}(q,d)} + \frac{1}{\alpha + r_{SEM}(q,d)}$$

    - CombMNZ 0 non-parametric [4]

$$combMNZ(d_k) = 2 \times \{(|L| - r_{LEX}(q,\ d) + 1) + (|L| - r_{SEM}(q,\ d) + 1)\}$$

    - Inverse Square Rank Fusion - non-parametric [5]

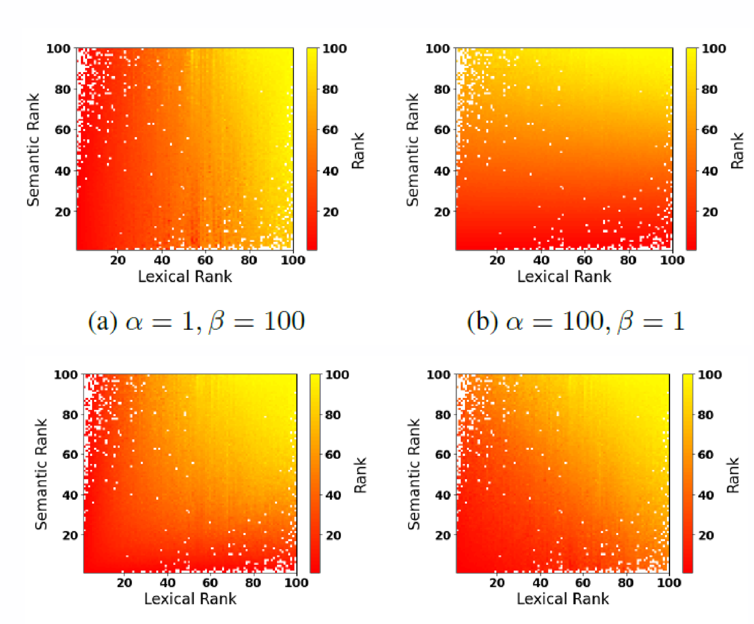
$$f_{ISR}(q,d) = 2 \times \left(\frac{1}{r_{LEX}(q,d)^2} + \frac{1}{r_{SEM}(q,d)^2}\right)$$

  - Voting Rule (rank-based fusion)
    - Condorcet Fuse [6] - considers pairwise preference relationship; uses min-max normalized convex rank fusion as a tie breaker. The convex function is validated.

Sparse and FF indexes are built in davance. The pipeline is equivalent to the framework described in Figure 1.

- **Ranking in Relation to Semantic and Lexical Scores** - graph a **heatmap** where the **lexical and dense ranks are the axes** and **final rank after interpolation is the hue**
- **Ranking Effectiveness**
  - Metrics used for the experiment: **nDCG@10, RR@10, MAP@100**
  - **Validate** on datasets with a dev set for **parametric functions**
  - Run the experiment with each rank fusion function
- **Latency**
  - Datasets used are **Arguana** and **QUORA**
  - Sample 100 queries and retrieve 100 candidates
  - Using **timeit module**, measure latency of the interpolation and metrics computation stage with each rank fusion function. For each experiment, the pipeline is ran for multiple times. Several rounds of these runs are computed. The average of the fastest run reported.

4 Results and Discussion



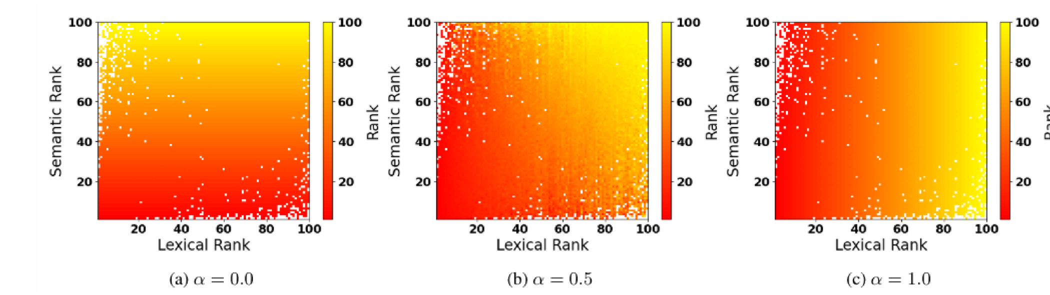
(a) $\alpha = 1, \beta = 100$ (b) $\alpha = 100, \beta = 1$ (c) $\alpha = 1, \beta = 1$ (d) $\alpha = 1000, \beta = 1000$

Figure 2: Reciprocal Rank Fusion

Reciprocal Rank Fusion (Figure 2)

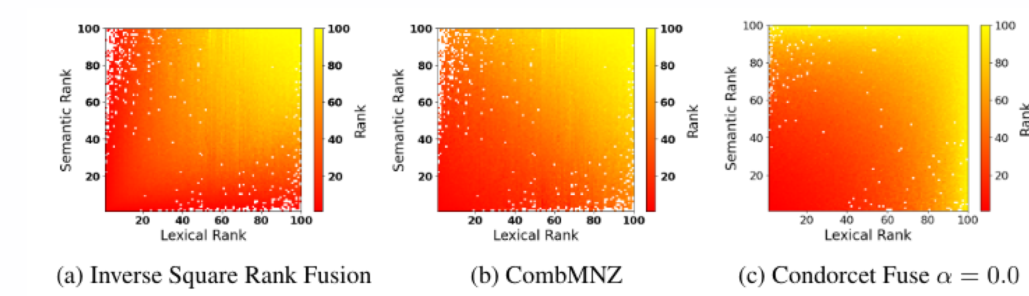
- Larger alpha and smaller beta: lexical < semantic
- Smaller alpha and larger beta: lexical > semantic
- Greater the parametric value, it mitigates the effect of the higher ranks. Thus, even if alpha and beta are equivalent, the ranking interpolation changes depending on the value.

4 Result and Discussion (Continued)



(a) $\alpha = 0.0$ (b) $\alpha = 0.5$ (c) $\alpha = 1.0$

Figure 3: Convex Rank Fusion



(a) Inverse Square Rank Fusion (b) CombMNZ (c) Condorcet Fuse $\alpha = 0.0$

Figure 4: Inverse Square Reciprocal, CombMNZ, Condorcet Fuse

Convex Rank Fusion (Figure 3)

- Larger alpha: lexical > semantic
- Smaller alpha: lexical < semantic
- Linear gradient

For **non-parametric functions** (Figure 4), the lexical and semantic scores always have equal weight.

- **Inverse Square Rank Reciprocal** : a high rank in one list dominates the other rank
- **CombMNZ**: additive of the ranks and no further manipulation
- **Condorcet Fuse**: a low rank in one list dominates the other rank due to the nature of preference relationship

	MS MARCO	FiQA	NFCorpus	QUORA	DBPedia	FEVER
Convex	0.0	0.1	0.2	0.1	0.1	0.0
Convex (Min-Max)	0.2	0.5	0.6	0.5	0.4	0.1
Convex (Z Score)	0.1	0.3	0.4	0.4	0.4	0.1
Reciprocal	(100, 1)*	(1, 1)	(60, 60)	(1, 1)	(80, 20)	(100, 1)
Condorcet Fuse	0.3	0.5	0.1	0.5	0.3	0.2

Table 1: Validation result

Validation Result (Table 1)

- Generally, better performance with more contribution of the **dense score**
- Datasets with **balance** between the two scores: FiQA (Arguana, CQADupStack), NFCorpus (SCIDOCS, Scifact), QUORA
- Datasets that in **favor of dense scores**: MS MARCO, QUORA, DBPedia, FEVER

Ranking Effectiveness Result (Table 2, 3)

- **Convex rank fusion** and their normalization variants yield the **best** ranking effectiveness
- **Reciprocal rank fusion** is the **second best** as it has the highest score excluding the convex functions
- **Non-parametric** approaches **worse** than the parametric approaches in general. However:
  - There is smaller difference in the scores for the **balanced datasets**
  - ISR has similar perforamance as reciprocal when the reciprocal's alpha and beta values does not have a large contrast
  - Condorcet Fuse has similar performance as CombMNZ

Discussion

- Score-based fusion is better than rank-based fusion since it does not discard the exact scores
- Parametric approaches are better than non-parametric approach due to its flexibility to adjust the weights of lexical and semantic scores

	TREC '19	TREC '20	FiQA-2018	NFCorpus	QUORA	DBPedia	FEVER
BM25 (No Interpolation)	0.480	0.494	0.253	0.322	0.768	0.274	0.427
Convex	0.679	0.641	0.311	0.335	0.841	0.379	0.663
Convex (Min-Max)	0.683	0.655	0.310	0.336	0.842	0.381	0.670
Convex (Z Score)	0.682	0.652	0.310	0.335	0.842	0.378	0.672
Reciprocal	0.679	0.641	0.298	0.331	0.828	0.356	0.663
Condorcet Fuse	0.629	0.592	0.300	0.329	0.821	0.337	0.582
Inverse Square Reciprocal	0.603	0.592	0.294	0.330	0.824	0.352	0.622
CombMNZ	0.623	0.597	0.300	0.329	0.820	0.337	0.579

Table 2: Ranking effective experiment nDCG score for validated datasets

	Arguana	CQADupStack	SCIDOCS	Scifact
BM25 (No Interpolation)	0.342	0.280	0.147	0.672
Convex	0.363	0.319	0.155	0.688
Convex (Min-Max)	0.336	0.318	0.150	0.684
Convex (Z Score)	0.340	0.319	0.157	0.687
Reciprocal	0.357	0.311	0.151	0.668
Condorcet Fuse	0.341	0.306	0.149	0.664
Inverse Square Reciprocal	0.352	0.309	0.149	0.669
CombMNZ	0.344	0.305	0.149	0.666

Table 3: Ranking Effective Experiment nDCG score for zero shot datasets

	Arguana	QUORA
Convex	171	446
Convex (Min-Max)	190	438
Convex (Z Score)	174	445
Reciprocal	174	446
Condorcet Fuse	42108	54190
Inverse Square Reciprocal	177	437
CombMNZ	172	460

Table 4: Latency experiment result

Latency Experiment Result (Table 4)

- All the rank fusion functions have a similar latency except for Condorcet Fuse
- Condorcet Fuse requires iteration through all possible pairs of the documents to establish the preference relationship
- The latency is affected by the size of datasets. However, it is likely to be due to the metrics computation as it requires accessing the actual relevance from the qrels
- Given this result, convex rank fusion is the most effective fusion function that has a good balance between ranking effectivess and latency

5 Conclusion and Future Work

RQ 1. How does the rankings change in relation to semantic and lexical ranks using different rank fusion functions?

- Parametric functions freely manipulate the influence that lexical and semantic scores have
- On the other hand, non-parametric functions put equal weight on them by default

RQ 2. How does using different rank fusion functions impact the ranking effectiveness in different domains?

- Convex > Reciprocal > CombMNZ, ISR, Condorcet Fuse
- Score-based > Rank-based
- Parametric > Non-parametric
- Non-parametric fusion function performance dependent on the domain

RQ 3. How does using different rank fusion functions impact the latency in different domains?

- Latency for interpolation same for all domains
- Condorcet Fuse a lot slower than other functions
- Convex fusion function is the most effective fusion function

Future Work

- Further explore the parameters of the parametric functions
  - Especially reciprocal function which take two parametric values
- Expand on the models experimented

References

[1] Jurek Leonhardt, Henrik Muller, Koustav Rudra, Megha Khosla, Abhijit Anand, and
Avishek Anand. Efficient neural ranking using forward indexes and lightweight encoders.
ACM Trans. Inf. Syst., 2023. Just Accepted
[2] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models, April 01, 2021 2021. Accepted at NeurIPS 2021 Dataset and Benchmark
Track.
[3] Sebastian Bruch, Siyu Gai, and Amir Ingber. An analysis of fusion functions for hybrid retrieval. ACM Transactions on Information Systems, 42(1):1–35, 2024.
[4] Antonio Ju´arez-Gonz´alez, Manuel Montes, Luis Villase˜nor-Pineda, David Pinto, and Manuel P´erez-Couti˜no. Selecting the N-Top Retrieval Result Lists for an Effective Data Fusion, volume 6008. 2010.
[5] Andr´e Mour˜ao, Fl´avio Martins, and Jo˜ao Magalh˜aes. Inverse square rank fusion for multimodal search. 2014.
[6] Shengli Wu and Xiaoqin Zeng. Condorcet Fusion for Blog Opinion Retrieval. 2012.