IMT SCUOLA ALTI STUDI LUCCA

Multidimensional Relevance in Cross-Encoder Re-ranking to Combat Health Misinformation

Marco Viviani

University of Milano-Bicocca
Department of Informatics, Systems, and Communication
IKR3 LAB

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA — DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

TOFFEe - TOols for Fighting FakEs / October 15-16, 2024, Sacrestia - Scuola IMT Alti Studi Lucca

Ikr3 Lab

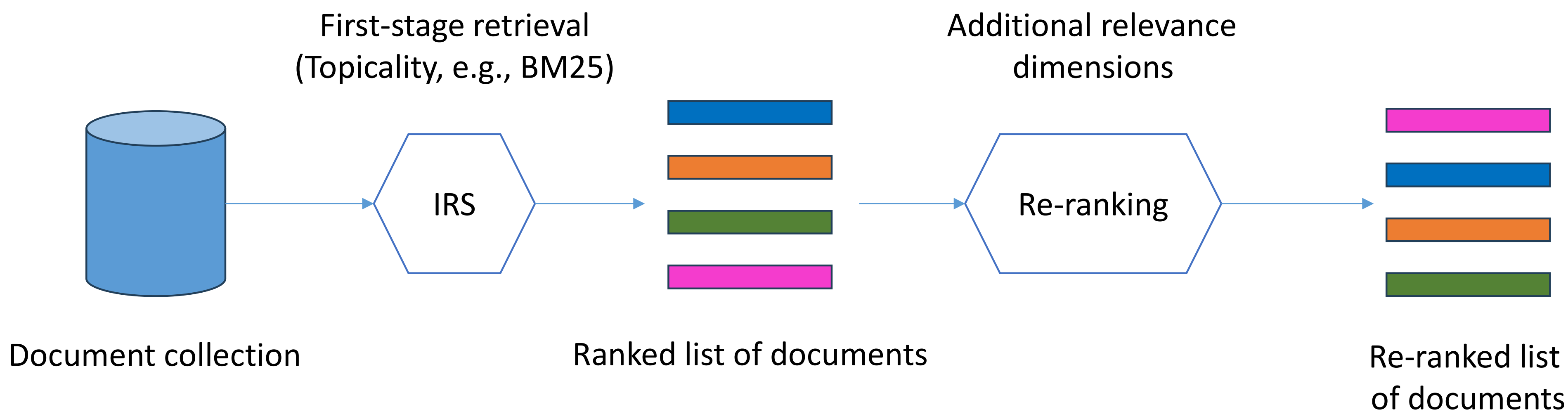
The Context

- Increasing interest in addressing the problem of implementing **effective retrieval models** that consider **multiple dimensions of relevance** across various domains and tasks in the field of Information Retrieval.

→ How to retrieve both topically relevant and “true” documents?

- Today:
  - Document ranking is often achieved by performing a **first-stage retrieval**, usually focused on topical relevance, to efficiently identify a subset of relevant documents from the entire collection;
  - On this subset, a **re-ranking stage** is performed, where topicality and/or additional dimensions of relevance may be considered.

A «Classical» Re-Ranking Architecture



Aggregation and List Fusion

- Prevailing approaches that consider multiple relevance dimensions in re-ranking are based on the **aggregation** of the topicality score with other relevance dimension scores.
  - These methods employ varied techniques to calculate the relevance scores, a common thread among many is their reliance on simple **linear** or **non-linear aggregation**.

- Other approaches leverage **rank fusion** methods, mainly based on Reciprocal Rank Fusion, CombSUM, and Borda count.

Cross-Encoder Re-Ranking (1)

- Approaches that have proven effective for re-ranking are today based on the use of **cross-encoders**.

- A cross-encoder is a type of **neural network architecture** commonly used for re-ranking tasks in Information Retrieval, Question-Answering, and Natural Language Processing.
  - It operates by **jointly** encoding both the query (or input text) and candidate documents (or response options) to determine how well they match or relate to each other.

- So far, they have been used with respect to a **single dimension of relevance**, namely topicality.

Cross-Encoder Re-Ranking (2)

- Two **sequences** – i.e., the *query* q and a *candidate document* d – are concatenated and fed into a **Transformer** model (like BERT).

- Transformer **attention heads** can directly model which elements of one sequence are correlated with elements of the other, allowing a (*topical*) *relevance score* σ to be calculated.

$$\sigma(q, d) = CE([\mathrm{CLS}]\ q\ [\mathrm{SEP}]\ d\ [\mathrm{SEP}]) \cdot W$$

CLS and SEP are special tokens. W is a learned matrix that represents the relationship between the query and document representations.

Enhancing Cross-Encoder Re-Ranking

- The $CE_{BM25CAT}$ model (2023) has been proposed to improve the effectiveness of BERT-based re-rankers by **injecting the topicality score** obtained by a first-stage BM25 model **as a token** (BM25) into the input of the cross-encoder.
    - Askari, A., Abolghasemi, A., Pasi, G., Kraaij, W., & Verberne, S. (2023, March). Injecting the BM25 score as text improves BERT-based re-rankers. In European Conference on Information Retrieval (pp. 66-83). Cham: Springer Nature Switzerland.

$$\sigma(q, d) = CE([\mathrm{CLS}]\ q\ [\mathrm{SEP}]\ \mathrm{BM25}\ [\mathrm{SEP}]\ d\ [\mathrm{SEP}]) \cdot W$$

Cross-Encoders and Relevance Dimensions

- The $CE_{BM25CAT}$ model **does not account** for **additional relevance dimensions** to be used for re-ranking.

- In this work, we aim to explore the impact of **incorporating other dimensions of relevance into a cross-encoder** for document re-ranking.
  - E.g., Novelty, readability, **credibility**.
  - Upadhyay, R., Askari, A., Pasi, G., & Viviani, M. (2024, March). Beyond Topicality: Including Multidimensional Relevance in Cross-encoder Re-ranking: The Health Misinformation Case Study. In European Conference on Information Retrieval (pp. 262-277). Cham: Springer Nature Switzerland.

The Proposed Solution (1)

- We **DO NOT manipulate the input sequence** of the cross-encoder with an additional relevance score for an additional relevance dimension.

- We **integrate** a so-called **relevance statement** into each document.
  - This statement is constituted by a text related to the relevance dimension under consideration and its associated relevance score.

- This **"enhanced" document** is provided, along with the query, as **input of a cross-encoder** to obtain the **overall** relevance score.

The Proposed Solution (2)

- The **cross-encoder-based model** proposed in this work to perform re-ranking is named $CE_{rel.stat}$.

- It is based on performing **four steps**:
  1. An **initial retrieval phase** to compute topicality scores.
  2. **Computation of a relevance score** for an additional relevance dimension → credibility;
  3. **Enhancement of the (retrieved) documents** with a text related to the additional relevance dimension in the form of a relevance statement;
  4. **Actual re-ranking** that occurs by feeding the cross-encoder with the query and the related enhanced documents.

Detailed Steps (1)

- 1. Ranking → BM25 → **Topicality score**.

- 2. Additional relevance dimension → **Computing credibility** (in the health domain).
  - This approach involves comparing the content of retrieved documents, given a query, with scientific articles, which are considered reliable sources of evidence for the same query.
  - Both the documents and scientific articles are represented using BioBERT.
  - How to compute it? → **Next slide**.

Detailed Steps (2)

- **Credibility score**: $cred(d,q)$ → A linear combination of the cosine similarity scores between d and the top-k scientific articles j_is that were deemed relevant to the same query for which d was retrieved.

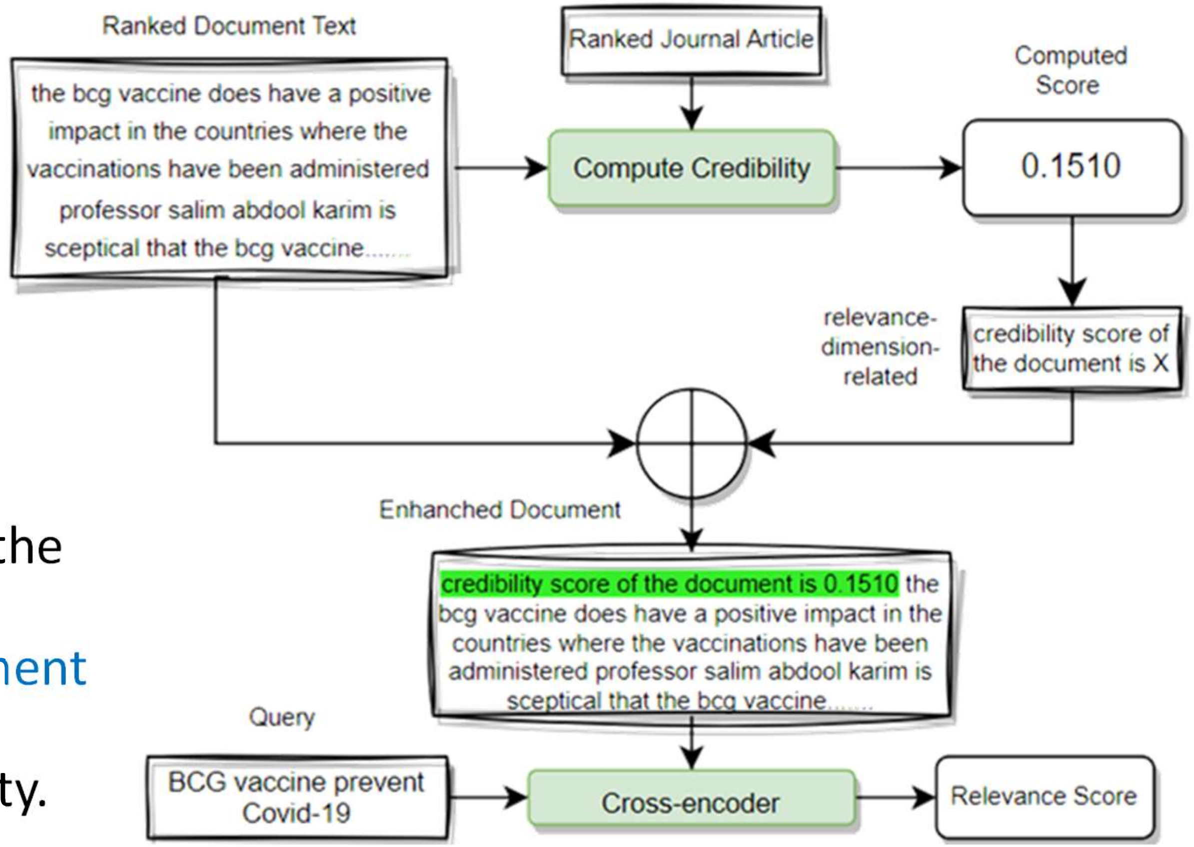
$$cred(d,q) = \omega_1 \cdot \cos(d, j_1) + \omega_2 \cdot \cos(d, j_2) + \cdots + \omega_k \cdot \cos(d, j_k)$$

where $\omega_1, \omega_2, \ldots, \omega_k \mid \sum \omega_i = 1$, and $\omega_i \geq \omega_{i+1}$ $(1 \leq i \leq k-1)$.

- These weights allow assigning greater emphasis to the similarity scores according to the rank of the retrieved articles j_is.

Detailed Steps (3)

- 3. We **enhance each document retrieved** in the first-stage phase with a relevance statement related to the additional relevance dimension(s) considered.
  - When considering credibility, the form of the statement is: "credibility score of the document is X", where X is the relevance score associated with credibility.



Detailed Steps (4)

- 4. **Cross-Encoder re-ranking**: we replace the original document in the CE input sequence with the enhanced document representation.
  - I.e., we replace d with $\tilde{d}$ (i.e., the enhanced document).

$$\sigma(q, d) = CE\big([\mathrm{CLS}]\ q\ [\mathrm{SEP}]\ \tilde{d}\ [\mathrm{SEP}]\big) \cdot W$$

Task and Considered Datasets

- We focused on the **ad-hoc retrieval** task in **Consumer Health Search**.

- Datasets:
  - **TREC-2020 Health Misinformation** Track Dataset.
  - **CLEF-2020 eHealth** Track Dataset.
  - A subset of 1 million documents from each track was used, with the TREC-2020 Track covering 46 topics related to Coronavirus and the CLEF-2020 Track covering 50 medical condition topics.

Baselines

- BM25: the BM25 retrieval model as implemented by PyTerrier;
- WAM: a current state-of-the-art aggregation-based multidimensional relevance model;
- CE: the original cross-encoder model for re-ranking;
- $CE_{BM25CAT}$: the cross-encoder re-ranker where the BM25 score is injected into the input sequence of the cross-encoder;
- $CE_{CredCAT}$: a cross-encoder re-ranker, where a credibility score is injected into the input sequence instead of the BM25 score;
- $CE_{BM25CredCAT}$: a cross-encoder re-ranker, where both BM25 and credibility scores are injected into the input sequence.

Implementation Details

- We employed **PyTerrier** for indexing and implementing the BM25 model.
  - We created **two indexes**, one for TREC-2020 and another for CLEF-2020.
- As the considered document set is health-related, we used **BioBERT** along with the base version of the BERT model for cross-encoder re-ranking training and inference.
- We **trained the CE on 80% of the queries-documents** from one dataset (e.g. TREC-2020) and used the other query set (e.g., CLEF-2020) as the **test set** and vice versa.

Results (TREC)

TREC 2020

Represent.	Model	NDCG@10	P@10	MRR@10	MAP
		TREC 2020			
Lexical	BM25	0.4166	0.4177	0.5107	0.2142
	WAM	0.5065	0.4976	0.5546	0.2453
BERT	$CE_{rel.stat}$	**0.6157**	**0.5977**	**0.7101**	**0.3208**
	$CE_{BM25CredCAT}$	0.5784	0.5671	0.6823	0.2875
	$CE_{CredCAT}$	0.5587	0.5581	0.6622	0.2652
	$CE_{BM25CAT}$	0.5374	0.5398	0.6341	0.2499
	CE	0.5589	0.5501	0.6619	0.2664
BioBERT	$CE_{rel.stat}$	**0.6704**	**0.6622**	**0.7961**	**0.3865**
	$CE_{BM25CredCAT}$	0.6219	0.6245	0.7512	0.3324
	$CE_{CredCAT}$	0.6111	0.6001	0.7061	0.3015
	$CE_{BM25CAT}$	0.5875	0.5812	0.6801	0.2765
	CE	0.6055	0.6059	0.6997	0.2986

Results (CLEF)

CLEF 2020

Represent.	Model	NDCG@10	P@10	MRR@10	MAP
		CLEF 2020			
Lexical	BM25	0.1054	0.1081	0.1578	0.1064
	WAM	0.0865	0.1002	0.1232	0.1102
BERT	$CE_{rel.stat}$	**0.3327**	**0.3401**	**0.5403**	**0.1601**
	$CE_{BM25CredCAT}$	0.3098	0.3141	0.5173	0.1356
	$CE_{CredCAT}$	0.2633	0.2703	0.4543	0.1198
	$CE_{BM25CAT}$	0.2288	0.2301	0.4147	0.0964
	CE	0.2579	0.2601	0.4456	0.1165
BioBERT	$CE_{rel.stat}$	**0.3762**	**0.3669**	**0.6187**	**0.1964**
	$CE_{BM25CredCAT}$	0.3221	0.3221	0.5731	0.1642
	$CE_{CredCAT}$	0.2805	0.2824	0.4812	0.1437
	$CE_{BM25CAT}$	0.2414	0.2522	0.4702	0.1274
	CE	0.2743	0.2811	0.4801	0.1474

Further Research

- **Use of LLMs** for both credibility assessment and relevance statement generation.

- Extension of the study to **other dimensions of relevance**.
  - Need for labeled datasets.

- **Explainability** of the results obtained (w.r.t. each relevance dimension).
  - Need for evaluation metrics addressing distinct relevance dimensions independently.

- Open to **further discussion**.

Thank you for your attention!

Grazie per l’attenzione!

Questions?