Robust Information Retrieval

WSDM 2025 tutorial

Yu-An Liu[a,b], Ruqing Zhang[a,b], Jiafeng Guo[a,b] and **Maarten de Rijke**[c]

https://wsdm2025-robust-information-retrieval.github.io/

March 10, 2025

01:30 – 05:00 PM

[a] Institute of Computing Technology, Chinese Academy of Sciences
[b] University of Chinese Academy of Sciences
[c] University of Amsterdam

Section 2: Preliminaries

SCAN
ME

Information retrieval task

Given:

- A query q,
- A document d from corpus D.

Information retrieval task

Given:

- A query q,
- A document d from corpus D.

The goal of an IR system is to employ the ranking function f to generate a score $f(q, d)$ for any query-document pair (q, d), reflecting the relevance degree between them, and produce a relevance permutation $\pi_f(q, D)$ according to the predicted score:

$$f(q, d) = g\left(\psi(q), \phi(d), \eta(q, d)\right),$$

where ψ, ϕ, and η return representations of q, d, and a relevance score

Neural IR model

$$f(q, d) = g\left(\psi(q), \phi(d), \eta(q, d)\right)$$

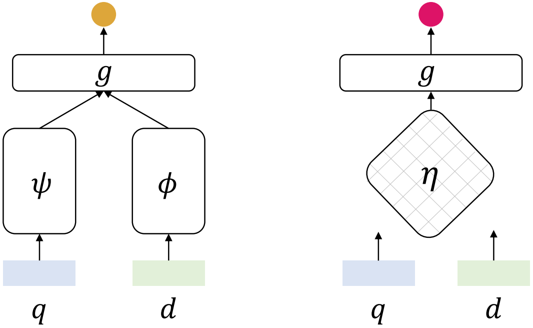
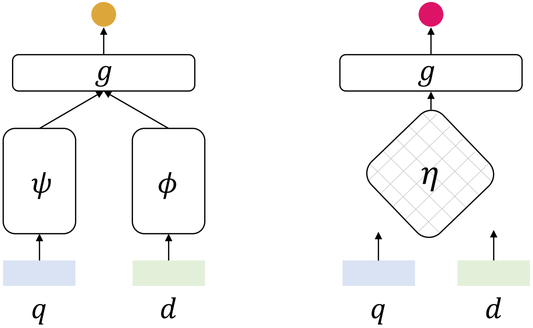


Image source: [Guo et al., 2020]

Neural IR model

$$f(q, d) = g\left(\psi(q), \phi(d), \eta(q, d)\right)$$



Dense retrieval model **efficiently** recalls document candidates with **dual-encoder**

Neural ranking model **effectively** generates the final ranked list with **cross-encoder**

Image source: [Guo et al., 2020]

Evaluation of IR model

In IR, we mainly focus on the top-K ranking result. Given:

- A metric M focus on the top-K ranking results, e.g., NDCG@K and MRR@K;
- A test dataset $\mathcal{D}_{\text{test}}$ with ground truth Y;

Evaluation of IR model

In IR, we mainly focus on the top-K ranking result. Given:

- A metric M focus on the top-K ranking results, e.g., NDCG@K and MRR@K;
- A test dataset $\mathcal{D}_{\text{test}}$ with ground truth Y;

The ranking performance $\mathcal{R}_M$ of the IR model is usually evaluated by

$$\mathcal{R}_M(f; \mathcal{D}_{\text{test}}, K) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(q,D,Y)\in\mathcal{D}_{\text{test}}} M(f; (q, D, Y), K).$$

Evaluation of IR model

In IR, we mainly focus on the top-K ranking result. Given:

- A metric M focus on the top-K ranking results, e.g., NDCG@K and MRR@K;
- A test dataset $\mathcal{D}_{\text{test}}$ with ground truth Y;

The ranking performance $\mathcal{R}_M$ of the IR model is usually evaluated by

$$\mathcal{R}_M(f; \mathcal{D}_{\text{test}}, K) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(q,D,Y) \in \mathcal{D}_{\text{test}}} M(f; (q, D, Y), K).$$

M includes a mapping function h related to ranking and an indicator function $\mathbb{I}\{\cdot\}$:

$$M(f; (q, D, Y), K) = \sum_{(d, y_d) \in (D, Y)} y_d \cdot h(\pi_f(q, d)) \cdot \mathbb{I}\{\pi_f(q, d) \leq K\}.$$

Definition (Top-K robustness in information retrieval)

Let $\delta \geq 0$ denote an acceptable error threshold. Given an IR model $f_{\mathcal{D}_{\text{train}}}$ trained on training dataset $\mathcal{D}_{\text{train}}$ with a corresponding testing dataset $\mathcal{D}_{\text{test}}$, an unseen test dataset $\mathcal{D}^*_{\text{test}}$, for the top-$K$ ranking result, if

$$\left|\mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}}; \mathcal{D}_{\text{test}}, K\right) - \mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}}; \mathcal{D}^*_{\text{test}}, K\right)\right| \leq \delta,$$

we consider the model $f_{\mathcal{D}_{\text{train}}}$ to be Top-K-robust for metric M.

Adversarial robustness in IR: Definition

To avoid the vulnerabilities of neural IR models being exploited by black hat SEO, we study adversarial robustness.

Adversarial robustness in IR: Definition

To avoid the vulnerabilities of neural IR models being exploited by black hat SEO, we study adversarial robustness.

Definition (Adversarial robustness in information retrieval)

Given an IR model $f_{\mathcal{D}_{\text{train}}}$ trained on training dataset $\mathcal{D}_{\text{train}}$ with a corresponding testing dataset $\mathcal{D}_{\text{test}}$, a new document set D_{adv} containing adversarial examples, and an acceptable error threshold δ, for the top-K ranking result, if

$$\left|\mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}}; \mathcal{D}_{\text{test}}, K\right) - \mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}}; \mathcal{D}'_{\text{test}}, K\right)\right| \leq \delta \text{ such that } \mathcal{D}'_{\text{test}} \leftarrow \mathcal{D}_{\text{test}} \cup D_{\text{adv}},$$

where $\mathcal{D}_{\text{test}} \cup D_{\text{adv}}$ denotes injecting the set of all generated adversarial examples D_{adv} into the original test dataset, and then model f is considered δ-robust against adversarial examples for metric M.

Out-of-distribution robustness: Definition

OOD generalizability stands as a pivotal requirement for contemporary IR systems, given the dynamic nature of user needs and evolving data landscapes.

Out-of-distribution robustness: Definition

OOD generalizability stands as a pivotal requirement for contemporary IR systems, given the dynamic nature of user needs and evolving data landscapes.

Definition (Out-of-distribution robustness of information retrieval)

Given an IR model $f_{\mathcal{D}_{\text{train}}}$, an original dataset with training and test data, $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, drawn from the original distribution $\mathcal{G}$, along with a new test dataset $\tilde{\mathcal{D}}_{\text{test}}$ drawn from the new distribution $\tilde{\mathcal{G}}$, and an acceptable error threshold δ, for the top-K ranking result, if

$$\left|\mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}};\mathcal{D}_{\text{test}},K\right)-\mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}};\tilde{\mathcal{D}}_{\text{test}},K\right)\right|\leq\delta \text{ where } \mathcal{D}_{\text{train}},\mathcal{D}_{\text{test}}\sim\mathcal{G},\tilde{\mathcal{D}}_{\text{test}}\sim\tilde{\mathcal{G}},$$

the model f is considered δ-robust against out-of-distribution data for metric M.

Performance variance: Definition

A robust neural IR model should not only have good performance over the entire query set, but also ensure that the performance on individual queries is not too bad.

Performance variance: Definition

A robust neural IR model should not only have good performance over the entire query set, but also ensure that the performance on individual queries is not too bad.

Definition (Performance variance of information retrieval)

Given an IR model $f_{\mathcal{D}_{\text{train}}}$ trained on training dataset $\mathcal{D}_{\text{train}}$ with a corresponding testing dataset $\mathcal{D}_{\text{test}}$, and an acceptable error threshold δ, for the top-K ranking result, if

$$\text{Var}\left(\left\{M\left(f_{\mathcal{D}_{\text{train}}}; (q, D, Y), K\right) \mid (q, D, Y) \in \mathcal{D}_{\text{test}}\right\}\right) \leq \delta,$$

where $\mathrm{Var}(\cdot)$ is the variance of the ranking performance of the IR model $f_{\mathcal{D}_{\text{train}}}$ on $\mathcal{D}_{\text{test}}$, then the model f is considered δ-robust in terms of performance variance for metric M.

Robustness in IR: Taxonomy

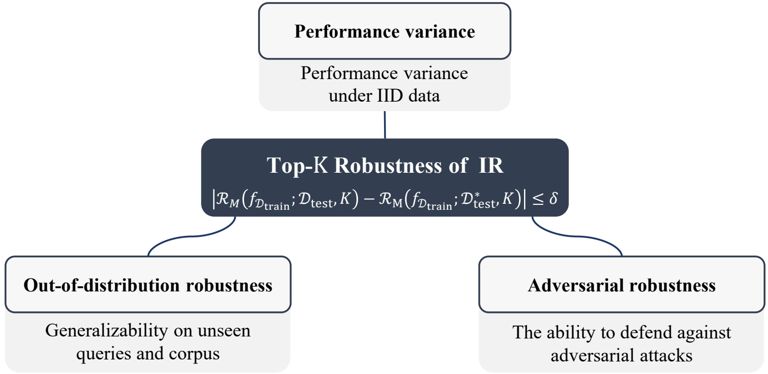


Image source: [Liu et al., 2024]

We will address adversarial robustness in Section 3 and OOD robustness in Section 4!

References

References i

J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, and X. Cheng. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6): 102067, 2020.

Y.-A. Liu, R. Zhang, J. Guo, M. de Rijke, Y. Fan, and X. Cheng. Robust neural information retrieval: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2407.06992*, 2024.