UNIVERSIDADE DA CORUÑA

LLM-Assisted Pseudo-Relevance Feedback

David Otero, Javier Parapar

IRLab, CITIC, Universidade da Coruña, Spain

Repo Available!

Motivation

- Query expansion through **Pseudo-Relevance Feedback (PRF)** is a technique to alleviate the vocabulary mismatch in Information Retrieval. Traditional methods like **RM3** estimate expanded query models from all first-stage top-k documents
- We explore if LLMs can serve to filter the pseudo-relevance set and approximate the true relevance set

LLM-Filtered RM3

- Traditional RM3 estimates expansion terms from all first-stage top-k documents, D_k:

$$P(t \mid R) \propto \sum_{d \in D_k} P(t \mid d) \cdot \prod_{w \in q} P(w \mid d)$$

- We insert an **LLM-based filtering stage before RM3 estimation**. The LLM acts as a **relevance judge** for documents in D_k, and RM3 is computed only over accepted documents, D_k^{LLM}

LLMF-Prob

- We first apply LLMF to filter relevant documents. At query time, we replace the query-likelihood with the probability that the LLM assigned to the token "true":

$$P(t \mid R) \propto \sum_{d \in D_k^{\text{LLM}}} P(t \mid d) \cdot P_{LLM}(\text{next token} = \texttt{`true'} \mid q, d)$$

- **We weigth the expansion terms with the probability of the token "true"**

Experiments

- We evaluate our approach on multiple IR test collections using standard metrics
- We also include an oracle that takes the true relevant documents from D_k and then applies RM3.
- Our method consistently outperforms QLD baseline, vanilla RM3 and monoT5 reranking across collections and metrics
- **More experiments and results on the paper, see the QR above!**

Method	AP@1000	NDCG@100
QLD	0.2213	0.3774
+ RM3	0.2920	0.4419
+ MonoT5 rerank	0.2474	0.4016
+ MonoT5F + RM3	0.2975	0.4503
+ MonoT5F + RM3 w/prob	**0.3072**	**0.4612**
+ LlaMAF + RM3	0.3002	0.4515
+ LlaMAF + RM3 w/prob	0.2962	0.4493
+ RM3 ORACLE	0.3854	0.5669

AP@1000 and NDCG@100 values on the AP8889 collection for the different evaluated approaches

Prompt ablation with LlaMA

- In the previous experiment, we only used title queries to prompt the LLM
- In this experiment, we investigate the effect of incorporating the *narrative* of the topic to the prompt
- Results show that including the narrative consistently improves the LLM-filtered variants across datasets and metrics

Method	AP@1000	NDCG@100
+ LlaMAF + RM3	0.3338	0.4949
+ LlaMAF + RM3 w/prob	0.3540	0.5125

AP@1000 and NDCG@100 values on the AP8889 collection obtained when including the narrative in the prompt for judging

Consellería de Educación, Ciencia, Universidades e Formación Profesional (Xunta de Galicia - Convenio para o desenvolvemento de accións estratéxicas de I+D+i 2025-2026)