Pseudo-Relevance for Enhancing Document Representation

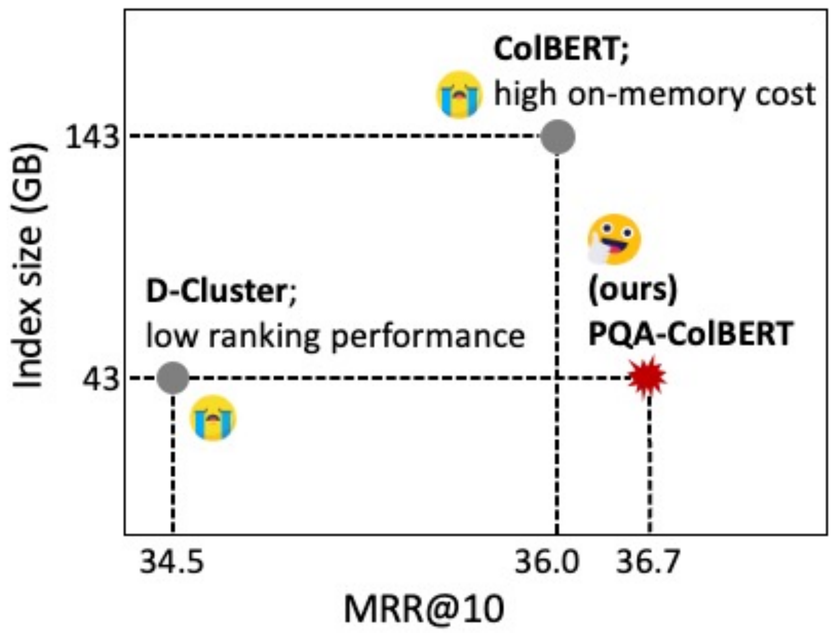
EMNLP 2022

Jihyuk Kim, Seung-won Hwang, Seoho Song, Hyeseon Ko, Young-In Song

YONSEI UNIVERSITY · SEOUL NATIONAL UNIVERSITY · NAVER

Summary

- ❖ We study **passage/document retrieval tasks** using MSMARCO and real-world search queries**.**
- ❖ We aim to improve **memory efficiency** without compromising the ranking effectiveness.
- ❖ Our proposed solution decreases the latency and the memory footprint, up to **8- and 3-fold**.



Task: passage/document retrieval

Given a query, a retriever is tasked to retrieve relevant documents. Aiming for scalable retrievers, we adopt **bi-encoder** design where *documents are indexed before queries are given*.

Two goals for representing doc: 1. achieving memory-efficiency, 2. fully preserving document semantics

➢ DPR/ANCE using single vector

[CLS] This is a document [SEP]

➢ ColBERT using all token vectors

[CLS] This is a document [SEP]

Goal

1. Efficiency: Is the index size scalable to Web-scale corpus?
2. Effectiveness: Are the document semantics fully preserved?

	1. Efficiency	2. Effectivenes
DPR/ANCE	✓	✗
ColBERT	✗	✓
Our target 🤔	Can we improve memory efficiency, w/o compromising the effectiveness?	

Our hypothesis:
A few query-relevant terms may be enough to match the queries, and the others can be pruned out, to decrease index overhead.

q	**How long** is the **flight** from **Chicago** to **Cairo**?
d	… total **flight duration** from **Chicago**, **IL** to **Cairo**, **Egypt** is 12 hours, 47 minutes. This assumes an average flight speed for a commercial airliner of …

Approach: Pseudo-Query-Aware ColBERT, or PQA-ColBERT

➢ Ours using only selected tokens , i.e., **pseudo-query (PQ) terms**

[CLS] This is a document [SEP]

Given q at test time, only relevance to PQ terms are considered, i.e., $\mathrm{rel}(q,d) \approx \mathrm{rel}(q,\bar{q}^d)$ where $\bar{q}^d$ denotes extracted pseudo-query terms from d.

🤔 **RQ 1. How can we obtain supervision for training a pseudo-query extractor?**

💡 ColBERT produces q**-matched terms in** d!

: $\mathrm{rel}(q,d) = \sum_{j\in[1,|q|]} \max_{i\in[1,|d|]} \mathbf{d}_i^\top \mathbf{q}_j$.

E.g., "How long" in q may max-pool "**flight duration**" in d.

🤔 **RQ 2. Would pseudo-query extractor preserve relevance?**

💡 We design our training objective as the degree of preserving $\mathrm{rel}(q,d)$ by using $\mathrm{rel}(q,\bar{q}^d)$ where $\bar{q}^d$ denotes extracted pseudo-queries from d.

Our proposed supervision y^* for extractor

y^* 1 1 … 0 0

… **flight duration** … commercial airliner

Our training objective for PQA-ColBERT

$\mathrm{rel}(q,d)$ w/ all tokens in d

d …

$\mathcal{L} = D_{\mathrm{KL}}\left(p_{\mathrm{rel}(q,d)} \,\|\, p_{\mathrm{rel}(q,\bar{q}^d)}\right)$

q

$\mathrm{rel}(q,\bar{q}^d)$ w/ only PQ tokens ($\bar{q}^d$)

$\bar{q}^d$ …

Experiment: passage/document retrieval

Model comparison

- ➢ Single-vector: **ANCE**
- ➢ Multi-vector/Cross-encoder (high-cost): **ColBERT**/**IDCM**
- ➢ Multi-vector (low-cost): **ME-BERT**, **D-Cluster**, **Ours**

query-agnostic term selection	query-aware term selection
ME-BERT, D-Cluster; *Limited capacity*	IDCM at query-time*; High-cost* **Ours at indexing-time; low-cost**