A Novel Dense Retrieval Framework for Long Document Retrieval

Jiajia WANG, Weizhong ZHAO, Xinhui TU, Tingting HE

Frontiers of Computer Science, DOI: 10.1007/s11704-022-2041-5

Problems & Ideas

- Problems of dense retrieval models for long document retrieval
  - Existing BERT-based models consider only the local context, while ignore the global context of the whole document, leading to the "topic drift" phenomenon.
- Ideas: we propose a dense retrieval framework (DRSCM) which integrates both local and global information
  - DRSCM uses a pretrained dense retrieval model to learn representations of short segments in an offline pattern, and improves the efficiency of long document retrieval.
  - Both local and global contexts are considered when calculating retrieval scores for short segments to address the issue of topic drift.

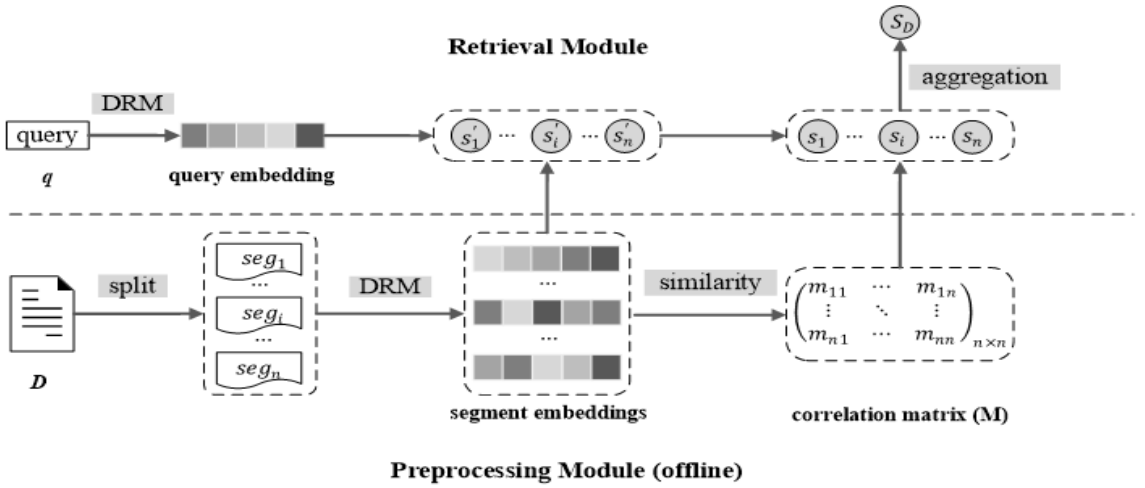

Fig. 1. The schematic representation of DRSCM

Main Contributions

Table 1: The comparison results of DRSCM with two different dense retrieval models and four aggregation strategies on GOV2 and Robust04.

Model	GOV2					Robust04				
	P@10	P@20	NDCG@10	NDCG@20	Latency (ms)	P@10	P@20	NDCG@10	NDCG@20	Latency (ms)
Bag-of-words										
BM25(Anserini)	0.5775	0.5381	0.4856	0.4784	132	0.4382	0.3631	0.4485	0.4240	88
BERT-based Models										
Birch (MS MARCO) [7]	-	-	-	-	-	0.4578	0.3964	0.4645	0.4512	290000
Vinilla_BERT [3]	0.5666	0.5483	0.4714	0.4670	129963	0.4633	0.4050	0.4750	0.4685	58400
CEDR_KNRM [3]	0.5746	0.5437	0.4618	0.4626	475940	0.4936	0.4175	0.5101	0.4832	146000
Aggregation Method "2sum"										
DRSCM(SBERT)	0.6919	**0.6335**	0.5810	0.5637	55	0.5008	0.4098	0.5157	0.4816	34
DRSCM(RepBERT)	**0.6924**	0.6300	0.5832	0.5620	28	0.5008	**0.4237**	0.5192	**0.4941**	11
Aggregation Method "3sum"										
DRSCM(SBERT)	0.6918	0.6284	0.5814	0.5606	55	**0.5064**	0.4205	**0.5238**	0.4919	34
DRSCM(RepBERT)	0.6870	0.6300	0.5828	0.5633	28	0.4972	0.4133	0.5140	0.4836	11
Aggregation Method "Max"										
DRSCM(SBERT)	0.6898	0.6270	**0.5878**	**0.5653**	55	0.5016	0.4088	0.5177	0.4821	34
DRSCM(RepBERT)	**0.6924**	0.6290	0.5840	0.5625	28	0.4980	0.4229	0.5118	0.4897	11
Aggregation Method "Mean"										
DRSCM(SBERT)	0.6233	0.5807	0.5377	0.5211	55	0.4731	0.3932	0.4873	0.4601	34
DRSCM(RepBERT)	0.6313	0.5828	0.5408	0.5228	28	0.4707	0.3974	0.4809	0.4594	11