Relevance-aware Diverse Query Generation for Out-of-domain Text Ranking

Jia-Huei Ju[1], Chao-Han Huck Yang[2], Szu-Wei Fu[2], Ming-Feng Tsai[3], Chuan-Ju Wang[4]

University of Amsterdam[1], NVIDIA Research[2], National Chengchi University[3], Academia Sinica[4]

Out-of-domain Passage Re-ranking

The challenges

- ✓ Task adaptation: **Heterogenous tasks** lead to different query-document relevancy.
- ✓ Domain adaptation: **Domain-specific** knowledge in the corpus.

Pseudo-relevant query generation

- ✓ Large models with in-context examples (e.g., InPars, PrPG).
- ✓ Small models fine-tuned on MS MARCO (40M ad-hoc search labels).

How can we squeeze more information from MS MARCO ?

- ✓ **Relevance-aware and diverse query generation.**
- ✓ How to use these "document-centric" data (i.e., (d, q^+, q^-))?

Document
(Title) animals environment general health health general weight philosophy ethics. *(Text)* Being vegetarian helps the environment … Modern farming is …

Relevance-aware Queries
1.0 why do you think meat is bad for a planet earth
0.9 what is the philosophy of vegetarian
…
0.7 what is a vegetarian diet?
0.6 what is deforestation in asian countries
….
0.1 what is an asian diet
0.0 what is the difference between food and a burger

Data augmentation: Relevance-aware diverse query generation (ReadQG)

Generalize the query generation process

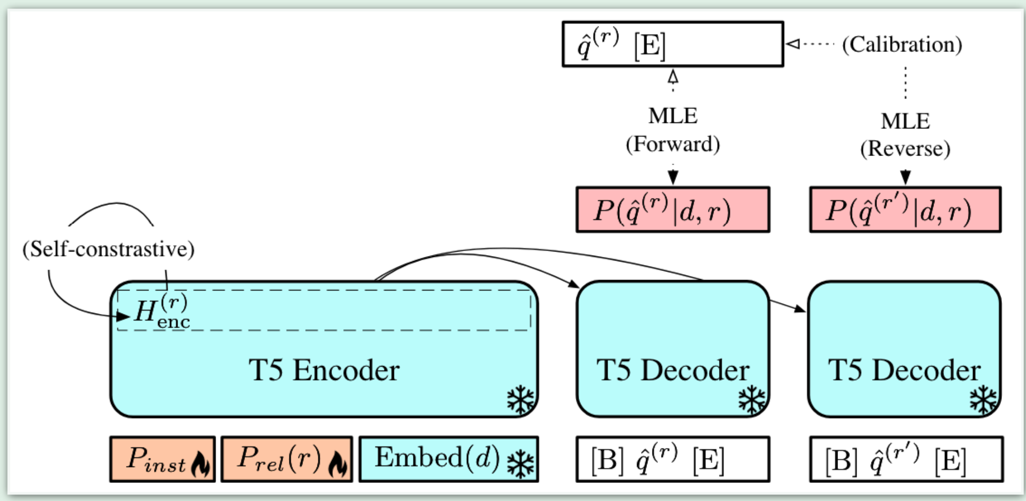
- ✓ Generate query that relevance is r (e.g., $r = 1$ means relevant query generation, see Fig.1).

Soft-prompt tuning serves as relevance-aware signals

- ✓ Instruction soft prompt: P_{inst}
- ✓ **Relevance soft prompt** (function) : $P_{rel}(r)$

Learning for diverse query generation

- ✓ Semi-supervised document-centric pairs: $(d, q^+, \ldots, q^-)$
- ✓ In addition to standard MLE, we consider
  - **Self-contrastive**: maximize distances $\cos(h_{q^{(r=0)}}, h_{q^{(r=1)}})$
  - **Sequence calibration**: diversify MLE



Domain-adaptation: Fine-tune new re-ranking models

Given synthetic document centric pairs $(d, d \rightarrow [\hat{q}^{(r=1)}, \hat{q}^{(r=0)}])$

- ✓ Standard triplet with binary cross-entropy

$$\mathcal{L}_{bce} = -\log P_F(\hat{q}, d) + \log P_F(\hat{q}, d^-)$$

- ✓ Query-based **margin ranking loss**

$$\mathcal{L}_{mr} = \max\big(0, \epsilon - F(\hat{q}, d)\big); \quad \epsilon = F(\hat{q}, \hat{q}^-)$$

Experiments

Experimental setups

- ✓ First-stage retrieval: BM25 top-100
- ✓ Models: FlanT5-base + MiniLM-MSMARCO

Results on BEIR

- ✓ Domain adaptation with limited compute.
- ✓ Negative queries have potential to improve.

	Objectives		Params (M)	nDCG@10					
Retrieval + Re-ranking (synthetic data)	$(\hat{q}, d)$	$(\hat{q}, d, \hat{q}^-)$	Gen./Rank	NFC	FQA	ARG	SCD	SCF	Avg.
BM25	-	-	-	32.5	23.6	41.4	15.8	66.5	36.0
BM25 + MiniLM-MS			- / 0.2M	35.0	34.7	41.7	16.6	68.8	39.4
BM25 + MiniLM-MS (InPars-v2 data)	✓	✗	6B / 0.2M	35.4	**35.2**	42.3	16.6	69.8	39.8
BM25 + MiniLM-MS (ReadQG)	✓	✗	220M / 0.2M	35.4	34.0	42.8	15.7	71.4	39.8
BM25 + MiniLM-MS (ReadQG)	✓	✓	220M / 0.2M	**35.5**	34.4	**49.6**	**16.7**	**71.6**	**41.6**

RepL4NLP 2024 @ The 62nd Annual Meeting of the Association for Computational Linguistics