SCONER: Scoring Negative Candidates Before Training Neural Re-Ranker For Question Answering

Man Luo, Mihir Parmar, Jayasurya Sevalur Mahendran, Sahit Jain, Chitta Baral

Arizona State University

Problem setting and Motivation

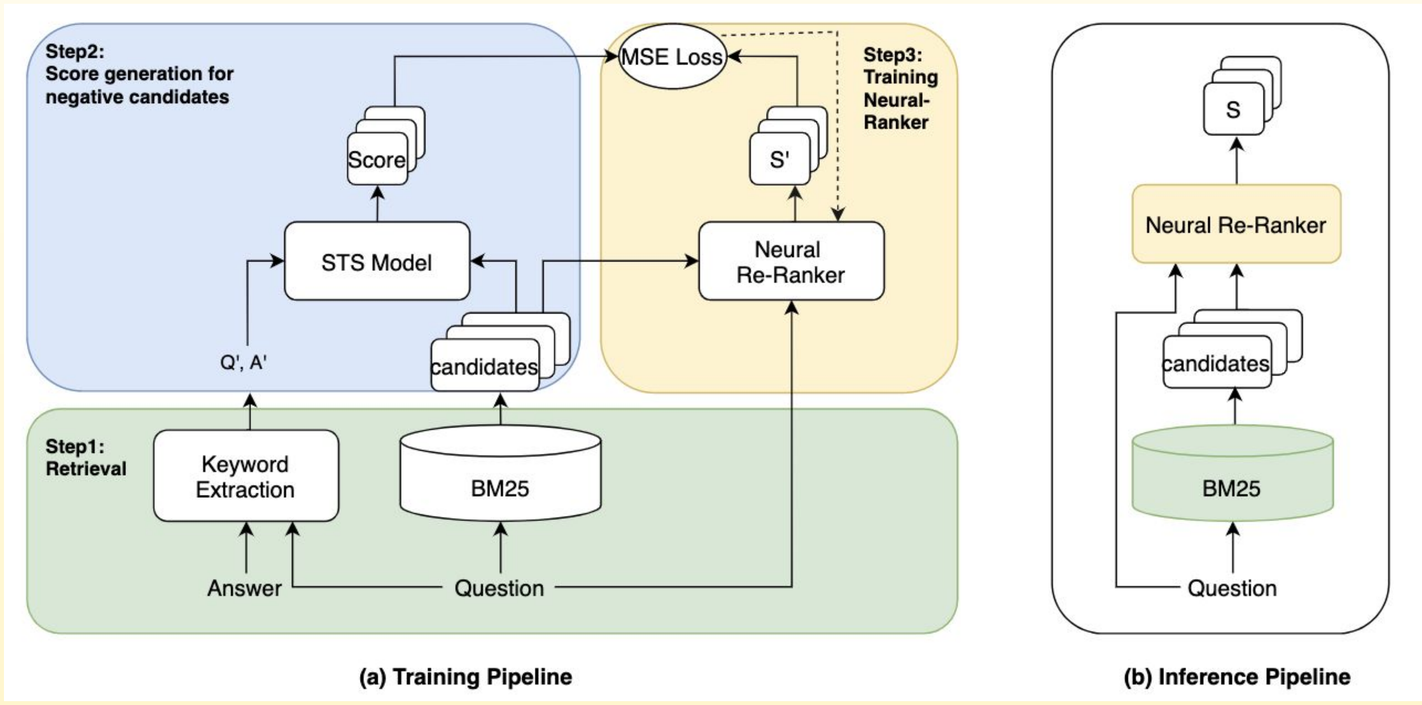
Retrieval-Based QA(ReQA) aims to find the sentence containing the answer span to a given query from a large corpus.

Two Stage Pipeline: (1) retrieve a small set of candidates from a large corpus and (2) re-rank these candidates.

Issues of Neural Reranker (NR): Standard methods train NR with equal weight to all negative candidates.

Research Question: It leads us to ask a question - ``is having different levels of negativeness beneficial for training neural re-rankers?

Proposed Method: SCONER



- Step1: retrieve negative candidates for a question using BM25.
- Step2: use a frozen STS model to generate negativeness scores for a question and candidate pair.
- Step3: train a neural re-ranker using the generated scores given by the STS model.

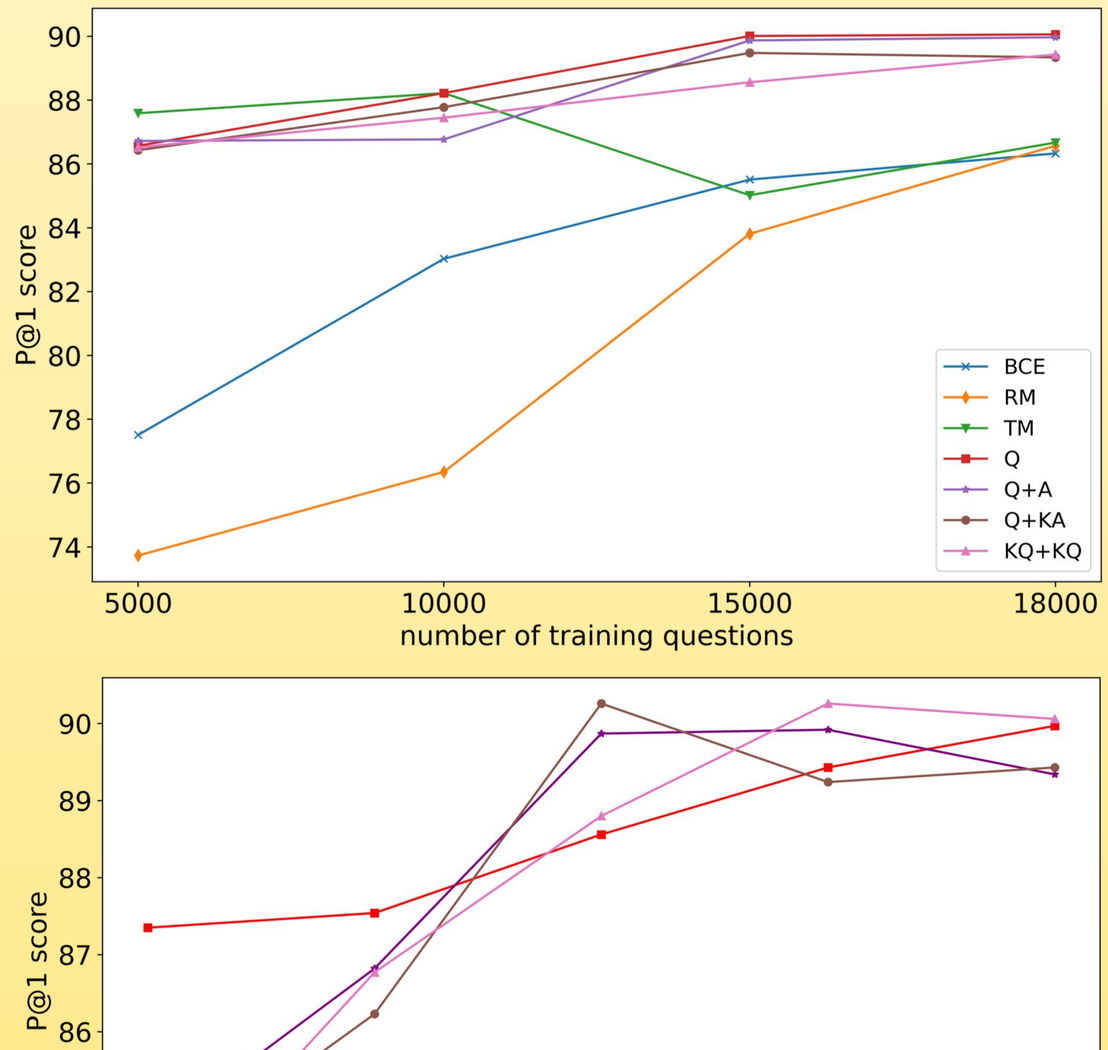
Experiment Results

Metric	Model	MultiReQA					
		NQ	SQuAD	HQA	SQA	TQA	Avg.
	Existing Approach (without re-ranking)						
	BM25	25.54	69.37	28.33	37.39	42.97	40.72
	USE-QA	38.00	66.83	31.71	31.45	32.58	40.11
	BERT	36.22	55.13	32.05	30.20	29.11	36.54
	Baselines						
	BCM	46.07	83.71	76.60	65.48	62.05	66.78
	RM	44.76	85.36	70.61	69.79	60.41	66.19
P@1	TM	50.33	85.65	70.00	73.03	65.43	68.89
	SCONER (Ours)						
	Q	48.64	89.09	64.76	68.64	62.20	66.67
	Q+A	49.97	89.14	**79.80**	70.27	64.73	70.78
	Q+KA	50.87	**89.48**	71.71	**78.26**	65.16	71.10
	KQ+KA	**52.80**	88.37	76.28	75.64	**65.45**	**71.71**
	Existing Approach (without re-ranking)						
	BM25	37.66	75.95	49.99	55.62	55.19	54.88
	USE-QA	52.27	75.86	43.77	50.70	42.39	53.00
	BERT	52.02	64.74	46.21	47.08	41.34	50.28
	Baselines						
	BCM	58.03	89.72	84.73	73.94	71.97	75.68
	RM	57.02	90.58	80.45	78.81	70.67	75.51
MRR	TM	60.87	90.27	81.00	82.22	**75.30**	77.93
	SCONER (Ours)						
	Q	58.46	92.51	70.73	76.64	68.94	73.46
	Q+A	60.14	92.36	**85.88**	78.62	72.48	77.90
	Q+KA	60.16	**92.71**	80.08	84.72	72.51	78.04
	KQ+KA	**61.50**	91.92	82.87	**83.02**	72.54	**78.37**

Augmenting the Input when generate negative score:

- Q: without augmentation, only use the question.
- Q+A: using the question and gold answer.
- Q+KA: using question and the key words in answers.
- KQ+KA: using keywords in question and answers

Take-Away



- SCONER outperforms three baselines by up to 13% absolute improvement on the SearchQA dataset and 5.5% on average across all datasets in terms of P@1.
- SCONER uses of a different negativeness score achieves better performance than the same score even when fewer negative candidates are used.
- SCONER has a significant advantage in a low resource setting