Collective Relevance Labeling for Passage Retrieval

SEOUL

NATIONAL

UNIVERSITY

2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics Seung-won Hwang

Hybrid: Seattle, Washington + Online

July 10–15, 2022

Summary

✤ We study passage retrieval task using MSMARCO and TREC-DL'19/20 datasets.

Minsoo Kim

✤ We tackle the challenge of incomplete relevance annotation.

✤ We adopt knowledge distillation, and introduce collective bi-encoder teacher.

Task: Information Retrieval (IR)

IR aims to retrieve relevant passages to the given query.

IR systems can be used not only for search engines but also for retrieval-augmented approaches.

Challenge: Incomplete Relevance Annotation

Q. Do we have enough labels?

keyphrase retrieve 👹 yoast.com > what-is-keyphrase-density-and-why-is-it-i What is **keyphrase** density and why is it important? b www.yourdictionary.com > keyphrase Keyphrase Meaning | Best 2 Definitions of Keyphrase What does keyphrase mean? (cryptography) A phrase used in end



NATIONAL

UNIVERSITY

YONSEI UNIVERSITY

Jihyuk Kim

A. No, we don't. Since there are too many passages (e.g., 8.8M) to be annotated, it is challenging to annotate all relevant passages.

Dataset	# of passages	# of queries	# of relevant passages per query	
MSMARO Train set	8.8M	500k	1.1→ Only tiny portions are annotated!	+ What if new relevant passages
TREC-DL 2019 Test set		43	58.2	

Approach: Knowledge Distillation (KD)											
KD introduces high-capacity teachers,				Content	Relevance Prediction						
to provide pseudo-relevance labels. <i>pseudo-relevance label</i>			Q/P		Individual	Collective					
passage P (Q, P) = 0.7			Q	Is caffeine an analgesic?							
			Тор-1 Р 🤡	caffeine, and codeine treat headaches.							
Teacher Capacity Efficiency		Тор-2 Р 🤡	butalbital, and caffeine								
Baseline	cross-encoder	0	×	Ton-3 P 🖌	Caffeine is considered a safe ingredient						
	bi-encoder	×									
Ours	collective bi-			Collective knowledge	treat tension headaches						

Experiment: Passage Retrieval

Dataset

 \succ We use public benchmark datasets:

MSMARCO and TREC-DL test collections.

Comparison

➤ w/o KD (-)

- ➢ KD from a cross-encoder (CE)
- > KD from an ensemble of CE (CE Ensemble)
- > KD from collective bi-encoder (**Ours**)

Evaluation Results

- \geq 1) **Teachers** provide more informed labels than incomplete labels.
- 2) Our collective bi-encoder teacher is more effective than cross-encoder teachers.

