LED: Lexicon-Enlightened Dense Retriever for Large-Scale Retrieval

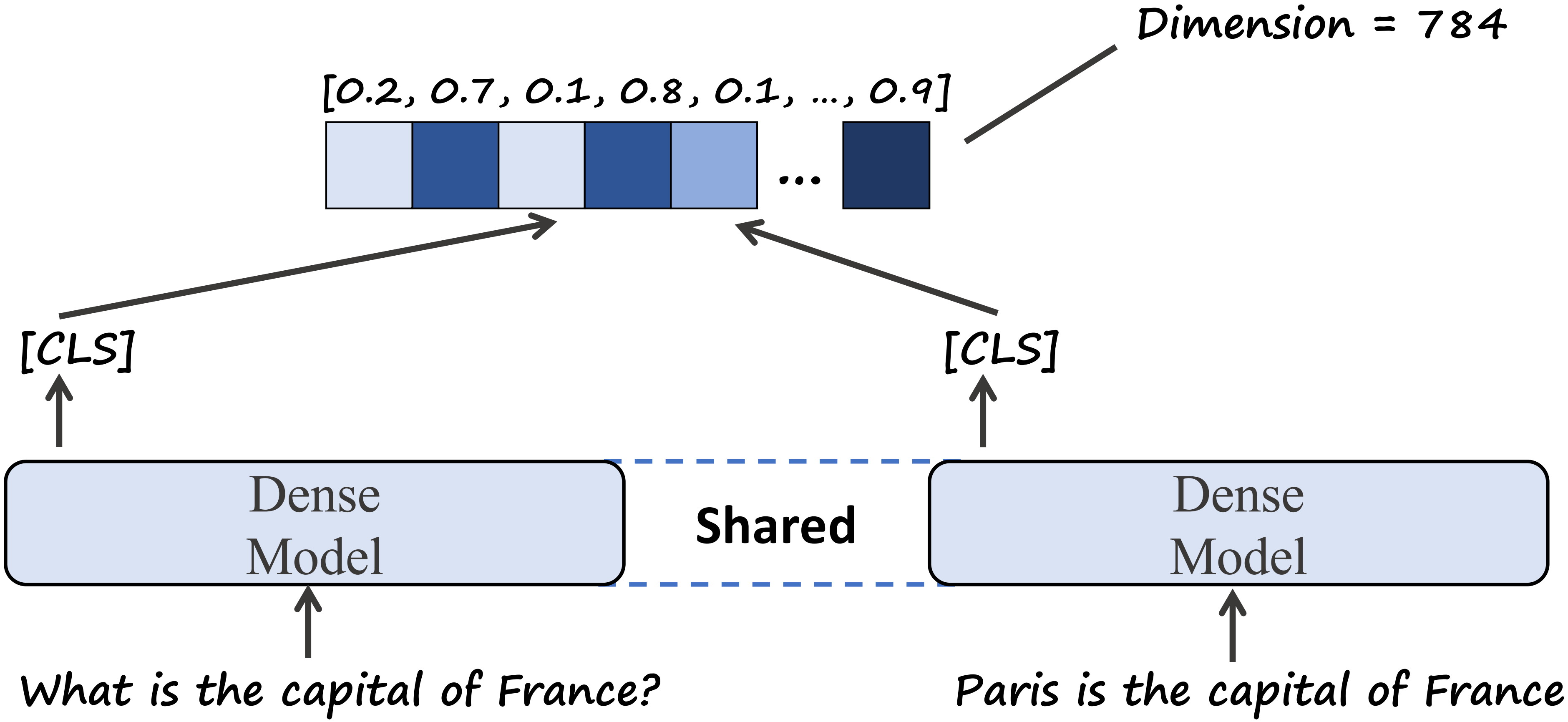
Kai Zhang[1], Chongyang Tao[2], Tao Shen[3], Can Xu[2], Xiubo Geng[2], Binxing Jiao[2], and Daxin Jiang[2]

[1]The Ohio State University, Columbus, Ohio, USA

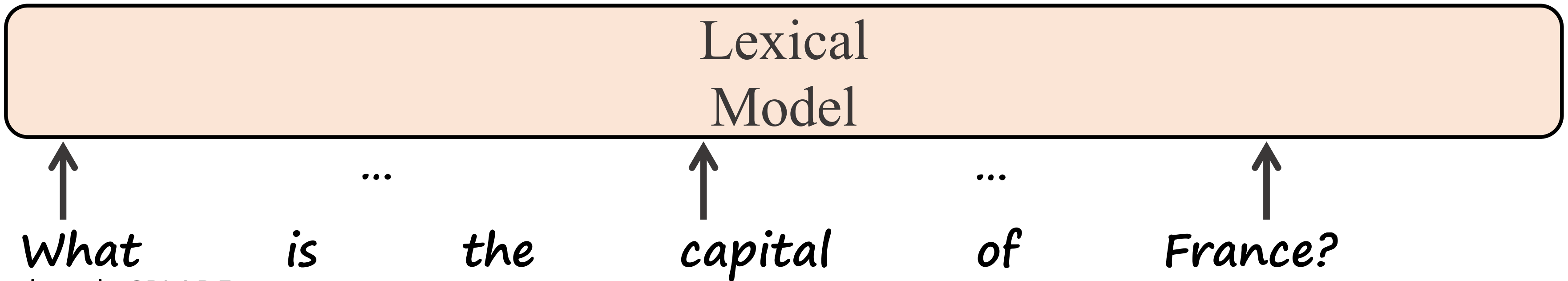
[2]Microsoft Corporation, Beijing, China

[3]AAII, FIET, University of Technology Sydney, Sydney, Australia

Dense Retriever - Sequence-level Semantic Matching

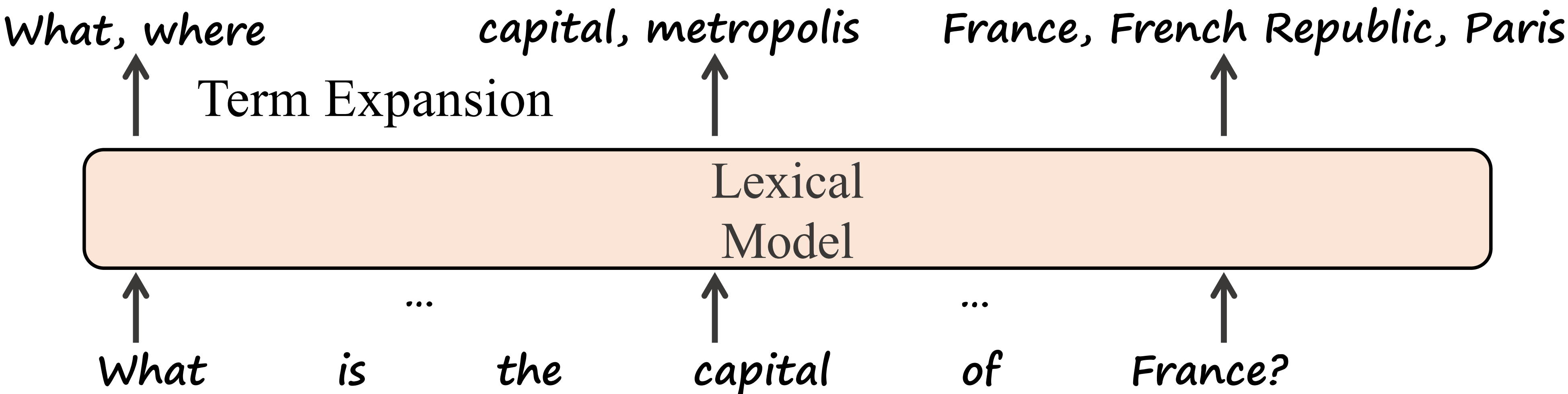


Lexicon-aware Retriever - Term-level Exact Matching



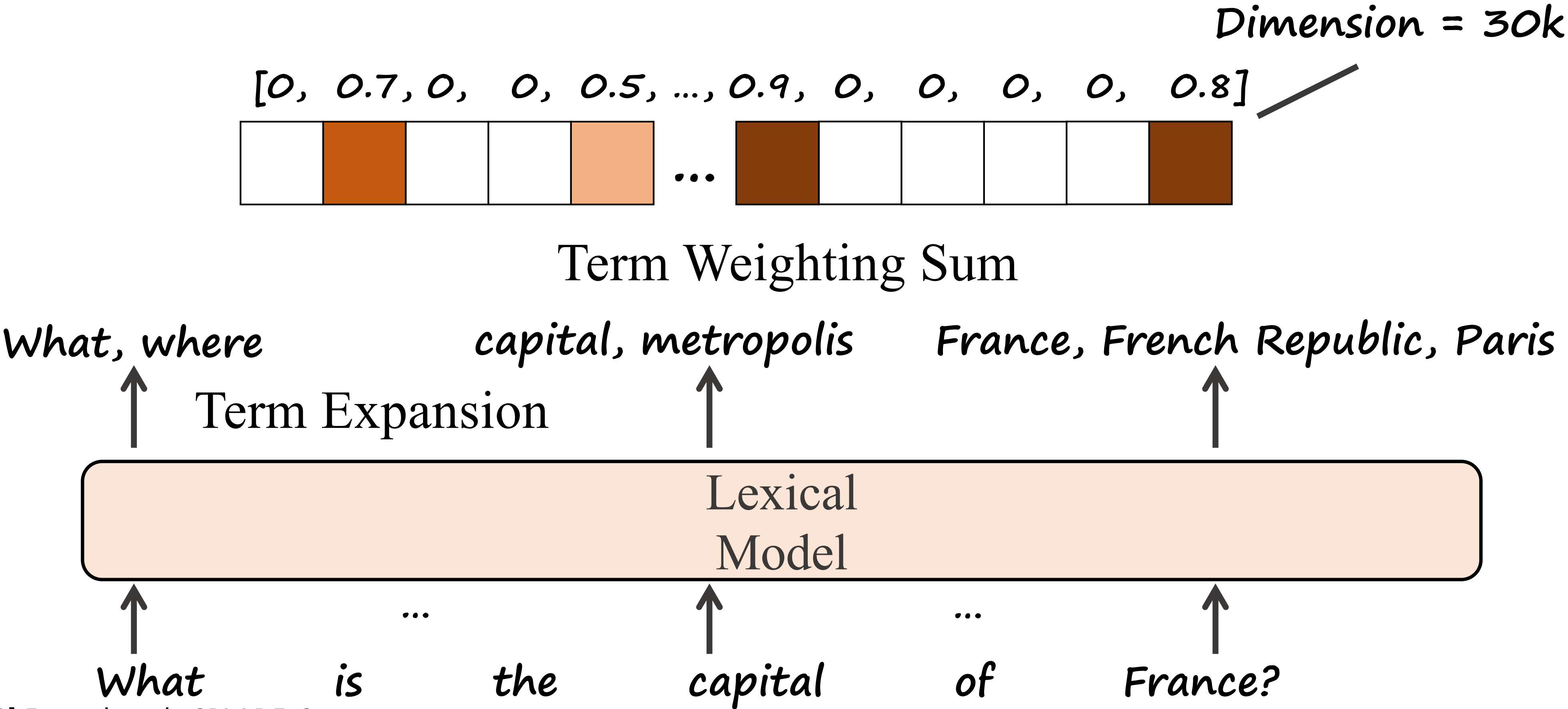
[1] Formal et al., SPLADE.

Lexicon-aware Retriever - Term-level Exact Matching



[1] Formal et al., SPLADE.

Lexicon-aware Retriever - Term-level Exact Matching



[1] Formal et al., SPLADEv2.

Dense and Lexicon-aware Retrieval Systems

- Dense Retriever (38.1 MRR@10 on MS MARCO)
  - Sequence-level Semantic Matching
  - Condensed Embedding Size (e.g., 768)

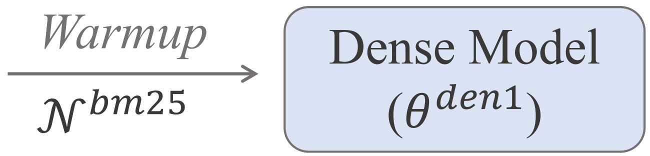
40%
Disagreement

- Lexicon-aware Retriever (38.3 MRR@10 on MS MARCO)
  - Lexicon-level Exact Matching
  - Sparse Embedding Size (e.g., vocab size=30k)

Can one embedding have both retrieval capabilities?

Experiment Setup

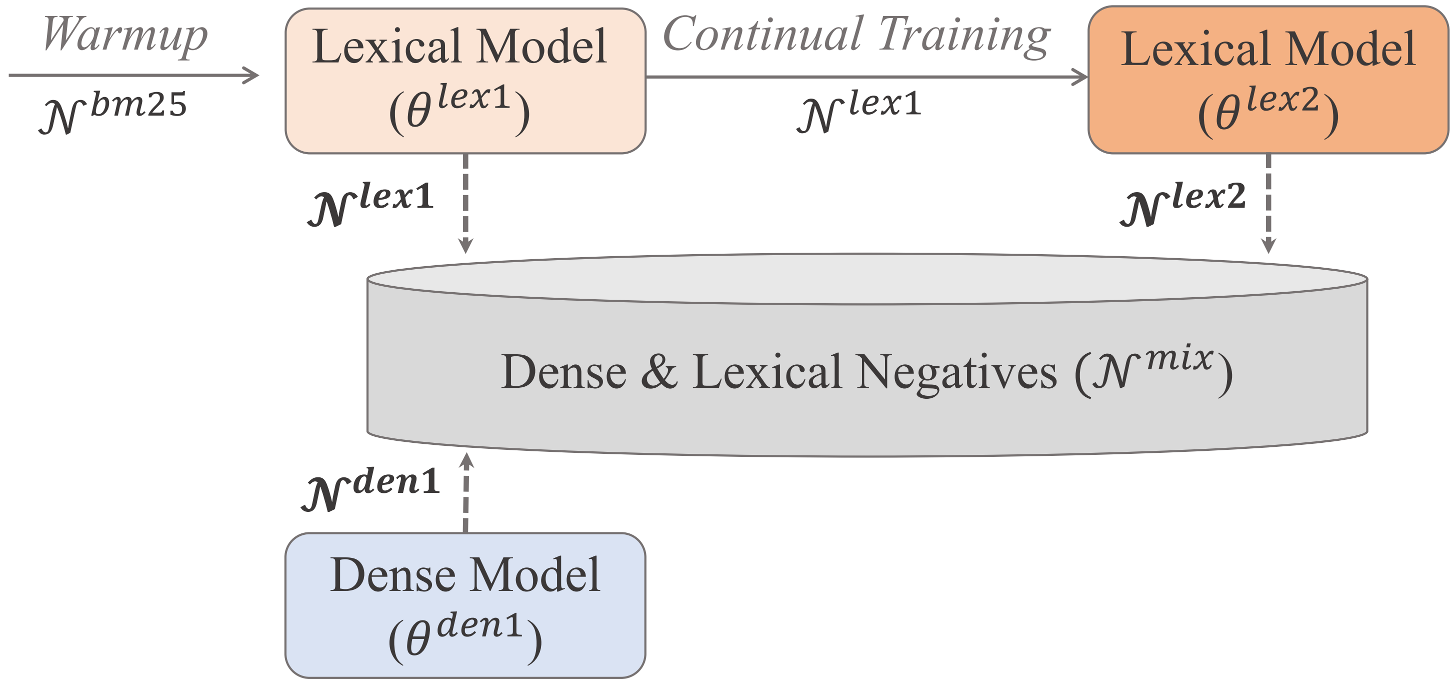
- Dense Retriever: coCondenser (110M)



- Lexicon-aware Retriever: DistilBERT (66M)

$\mathcal{N}^{S}$: hard negatives for method S (High-ranked false passages by S)
S Could be BM25, Lexical Retriever, and Dense Retrievers

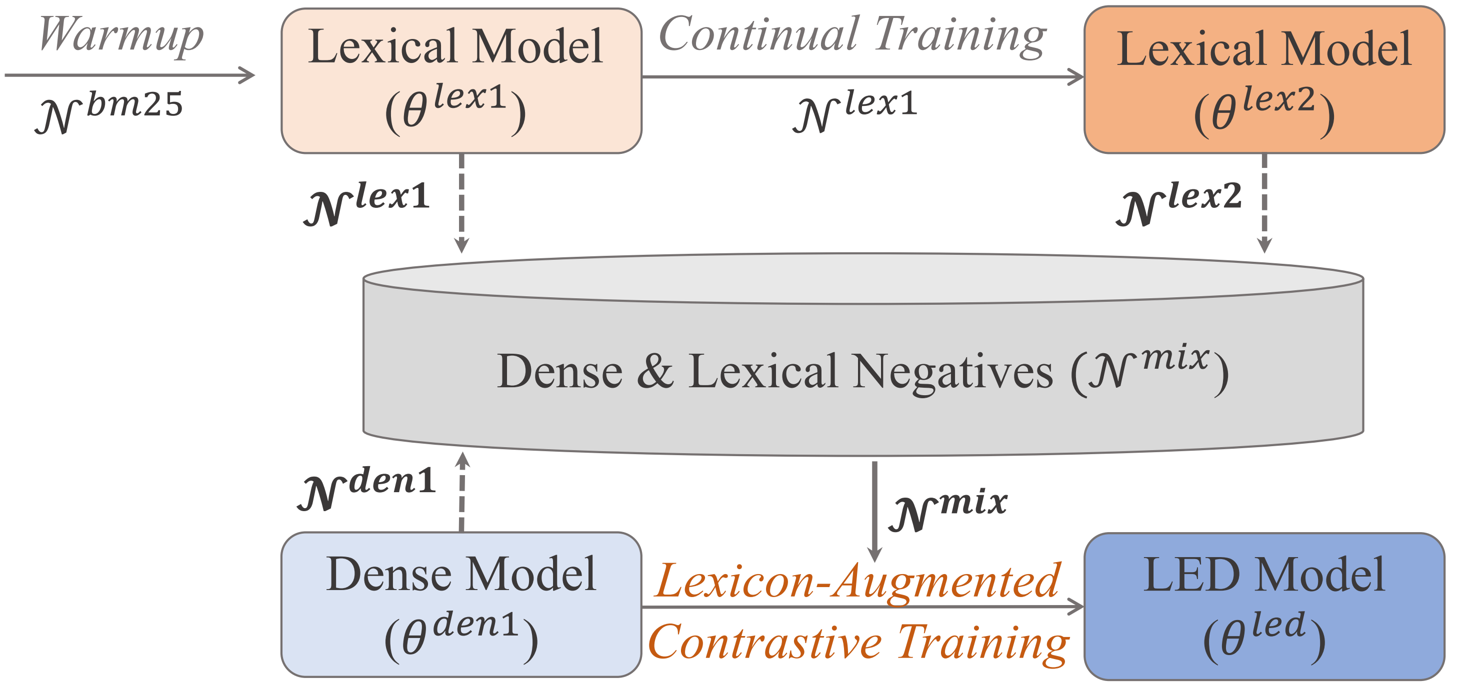
Strategy 1 - Lexicon-Augmented Contrastive Training



$\mathcal{N}^{S}$: hard negatives for method S (High-ranked false passages by S)

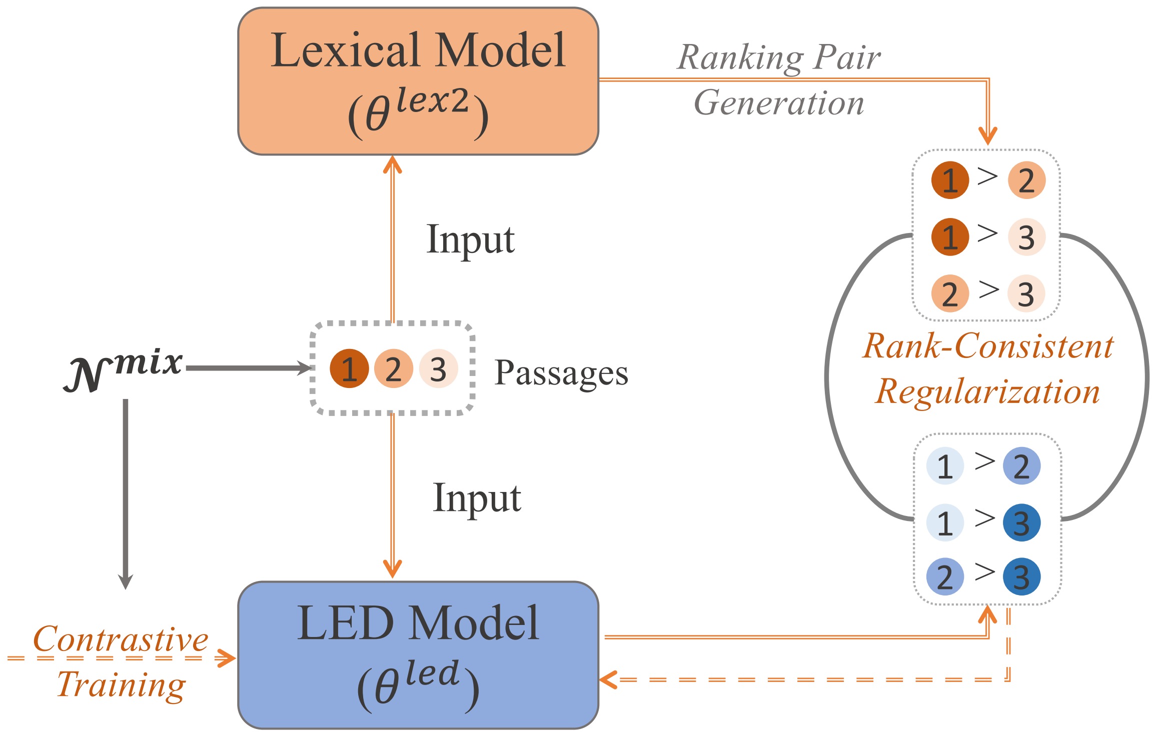
S Could be BM25, Lexical Retriever, and Dense Retrievers

Strategy 1 - Lexicon-Augmented Contrastive Training



$\mathcal{N}^{S}$: hard negatives for method S (High-ranked false passages by S)
S Could be BM25, Lexical Retriever, and Dense Retrievers

Strategy 2 - Rank Consistent Regularization



1. Pair-wise ranking supervision
2. No margin requirement for weak supervision
3. No training for lexical model (teacher)

Table 1: Experimental results on MS MARCO, TREC DL 2019 (DL'19), and TREC DL 2020 (DL'20) datasets (%). We mark the best results in bold and the second-best underlined. Numbers marked with '*' mean that the improvement is statistically significant compared with the baseline (t-test with p-value < 0.05).

Methods	PLM	Ranker Taught	Multi Vector	MS MARCO Dev			DL'19	DL'20
				MRR@10	R@50	R@1k	NDCG@10	NDCG@10
Lexicon-Aware Retriever								
BM25 [40]	-			18.7	59.2	85.7	50.6	48.0
DeepCT [7]	$\text{BERT}_{\text{base}}$			24.3	69.0	91.0	55.1	55.6
COIL-full [14]	$\text{BERT}_{\text{base}}$			35.5	-	96.3	70.4	-
UniCOIL [26]	$\text{BERT}_{\text{base}}$			35.2	80.7	95.8	-	-
SPLADE-max [10]	DistilBERT			34.0	-	96.5	68.4	-
DistilSPLADE-max [10]	DistilBERT	✓		36.8	-	97.9	72.9	-
UniCOIL Λ [4]	$\text{BERT}_{\text{base}}$			34.1	82.1	97.0	-	-
Dense Retriever								
ANCE [49]	$\text{RoBERTa}_{\text{base}}$			33.0	-	95.9	64.5	64.6
ADORE [52]	$\text{RoBERTa}_{\text{base}}$			34.7	-	-	68.3	66.6
TAS-B [17]	DistilBERT	✓		34.7	-	97.8	71.7	68.5
TAS-B + CL-DRD [51]	DistilBERT	✓		38.2	-	-	72.5	68.7
TCT-ColBERT [28]	$\text{BERT}_{\text{base}}$	✓		35.9	-	97.0	71.9	-
ColBERTv1 [21]	$\text{BERT}_{\text{base}}$		✓	36.0	82.9	96.8	67.0	66.8
ColBERTv2 [42]	$\text{BERT}_{\text{base}}$	✓	✓	39.7	86.8	**98.4**	72.0	62.1
coCondenser [13]	$\text{BERT}_{\text{base}}$			38.2	-	**98.4**	-	-
PAIR [38]	$\text{ERNIE}_{\text{base}}$	✓		37.9	86.4	98.2	-	-
RocketQAv2 [39]	$\text{ERNIE}_{\text{base}}$	✓		38.8	86.2	98.1	-	-
AR2-G [53]	$\text{BERT}_{\text{base}}$	✓		39.5	-	-	-	-
Our Models								
LEX (Warm-up)	DistilBERT			36.1	84.2	97.5	67.4	66.4
LEX (Continue)	DistilBERT			38.3	85.9	98.0	72.8	67.7
DEN (Warm-up)	$\text{BERT}_{\text{base}}$			36.1	83.5	97.7	64.7	65.9
DEN (Continue)	$\text{BERT}_{\text{base}}$			38.1	86.3	**98.4**	69.1	67.8
DEN (w/ RT)	$\text{BERT}_{\text{base}}$	✓		39.6	86.7	**98.4**	71.8	69.7
LED	$\text{BERT}_{\text{base}}$			39.6	86.6	98.3	70.5	67.9
LED (w/ RT)	$\text{BERT}_{\text{base}}$	✓		**40.2***	**87.6***	**98.4**	**74.4***	**70.2***

1. LED **signifincatly benefits** from lexical knowledge, even outdoing its teacher.
2. Lexical knowledge distillation is **comparable** with reranker distillation.
3. Lexical knowledge distillation is **compatible** with reranker distillation. Combining them together could reach SoTA.

Teaching strategies comparison

Table 2: Evaluation results of different teaching strategies on MS MARCO Dev (%). '*' refers to statistical significance.

Methods	MRR@10	R@1k
No Distillation	38.1	**98.4**
Filter [35]	38.4	**98.4**
Margin-MSE [16]	38.5	98.3
ListNet [48]	38.7	98.2
KL-Divergence [53]	39.0	**98.4**
Ours	**39.6***	98.3

1. Any lexical teaching strategies could improve dense retriever.
2. Weak supervision is the key.

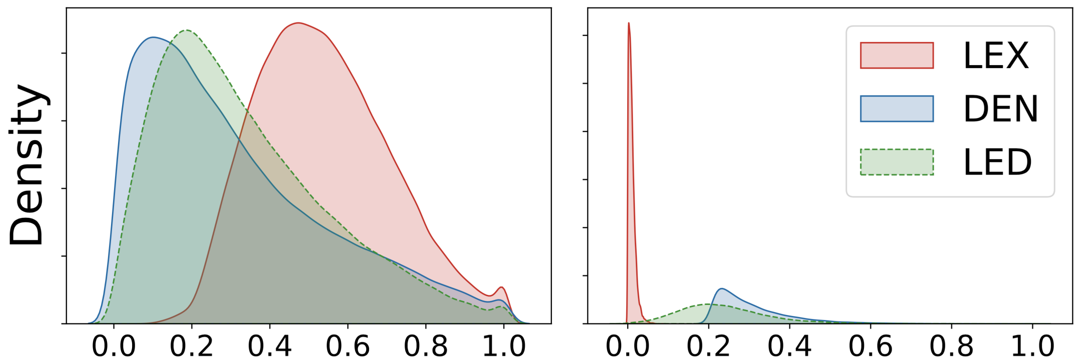
Ablation Study

Table 4: Ablation Study on MS MARCO Dev (%). Negs is short for negatives. '*' indicates statistical significance.

Retrievers	MRR@10	R@1k
LED	**39.6****	98.3
w/o Rank Regularization	37.9	**98.5**
w/o LEX Continue Negs ($\mathcal{N}^{\text{lex2}}$)	39.4	98.3
w/o LEX Warm-up Negs ($\mathcal{N}^{\text{lex1}}$)	39.4	98.3
w/o LEX Mixed Negs ($\mathcal{N}^{\text{lex1}} \cap \mathcal{N}^{\text{lex2}}$)	39.2	98.4

1. Removing lexical examples doesn't change the performance but removing rank regularization leads worse performance than simple dense continual training (38.3).

Visualization



(a) Top-ranked samples (b) Bottom-ranked samples

1. Comparing to dense (DEN) model, LED model's retrieval passages are more aligned with lexical model (LEX).
2. Thanks to weak supervision, the alignment is not too strong. LED keeps most dense properties.

Thanks!
Q & A

Code

Paper