B-PROP: Bootstrapped Pre-training with Representative words Prediction

Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Yingyan Li and Xueqi Cheng

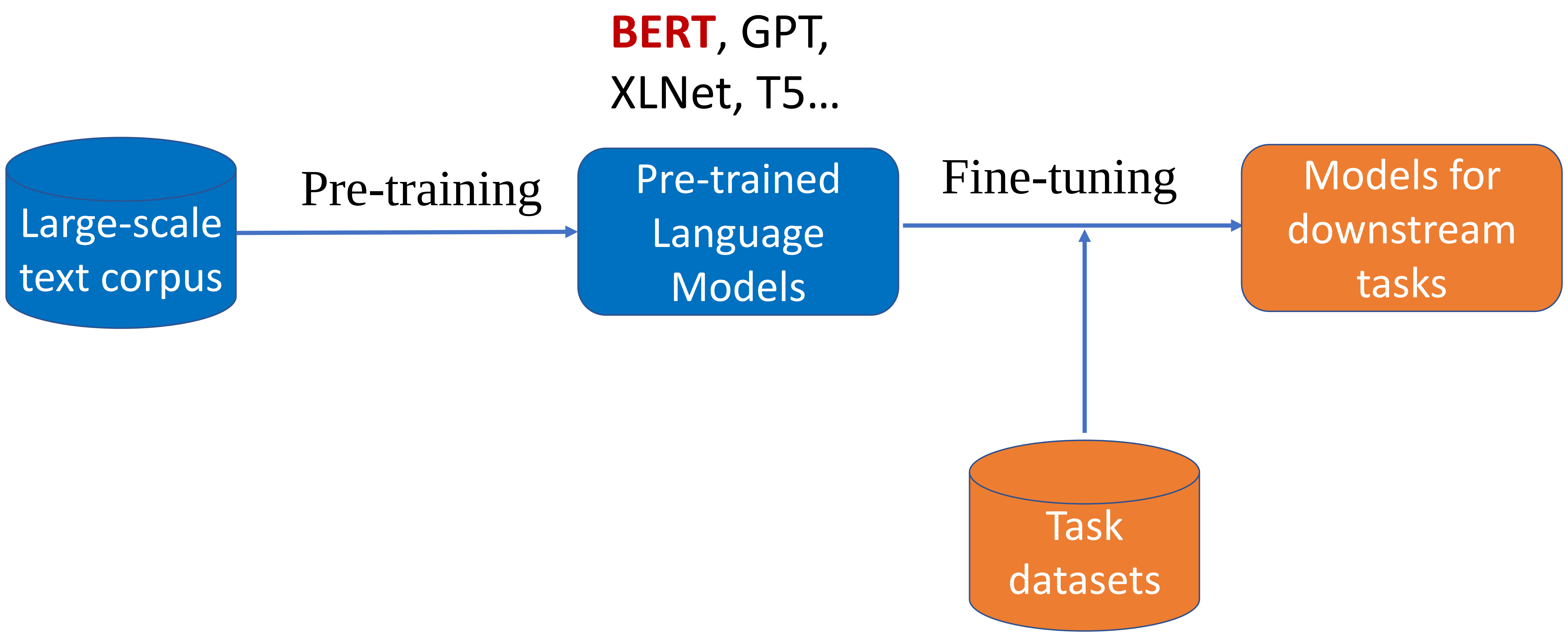
https://arxiv.org/pdf/2104.09791.pdf

1. CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences

2. University of Chinese Academy of Sciences

New Paradigm of NLP

- Pre-training and then fine-tuning paradigm
- Significant benefit for tasks with limited training data

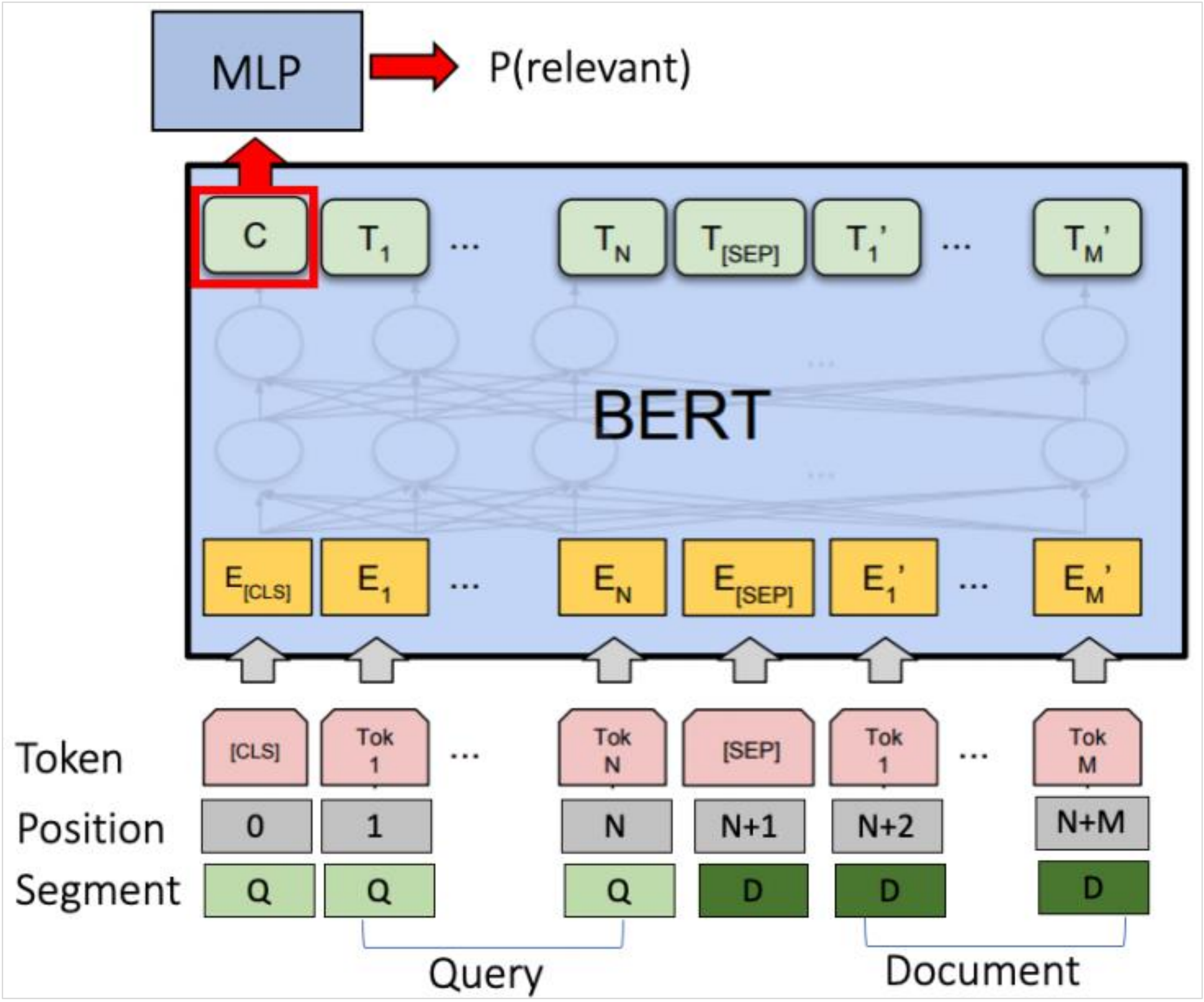


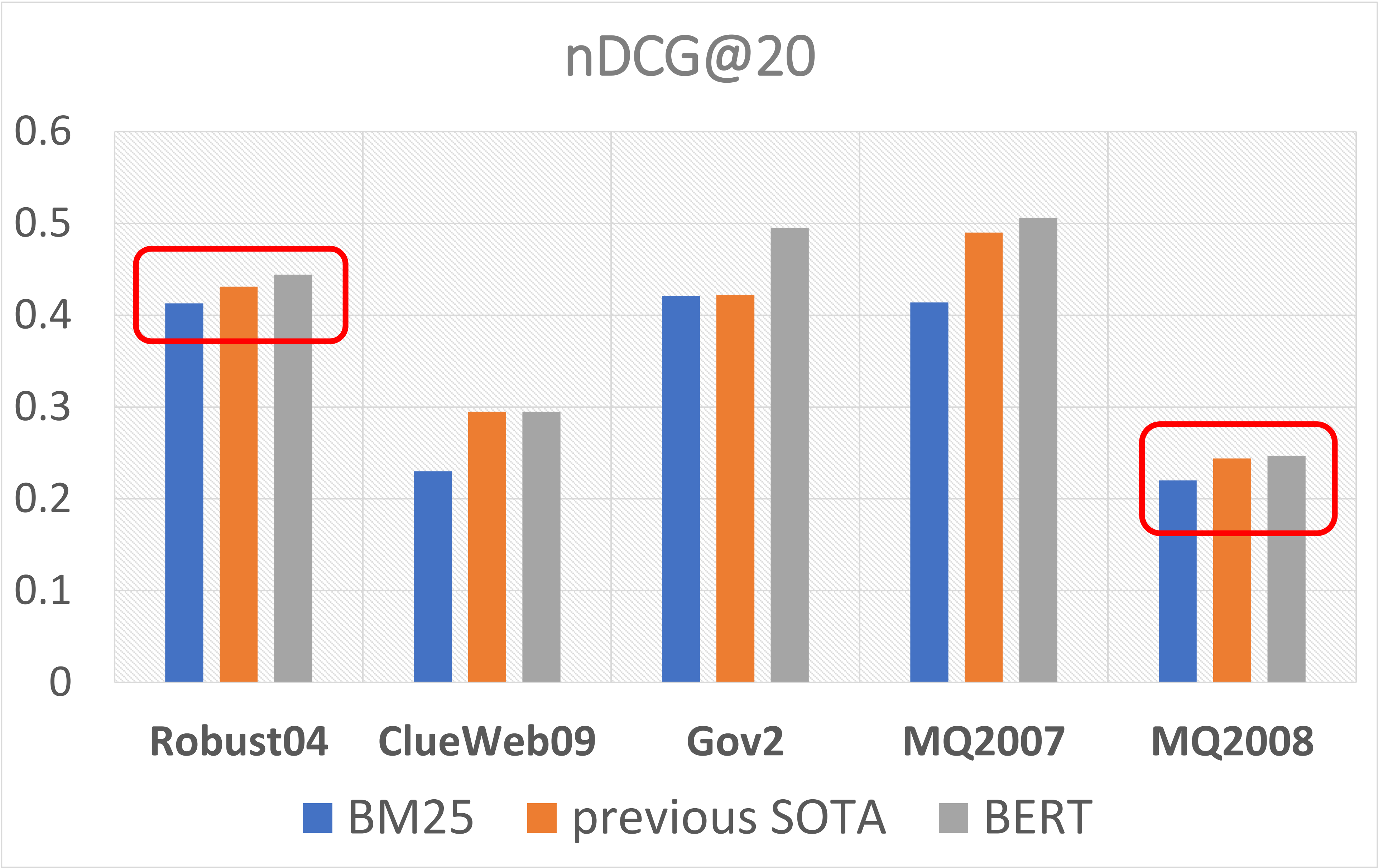
NLP Tasks

- Machine Translation
- Sentiment Analysis
- Question Answering
- Dialogue & Chatbot
- Textual Entailment
- Paraphrasing
- Semantic Parsing
- …

BERT for IR

- Explore BERT in the context of ad-hoc document ranking (reranking)





Pre-trained models do benefit the search tasks, but sometimes the improvement is limited

Deeper text understanding for IR with contextual neural language modeling, SIGIR 2019
Modeling diverse relevance patterns in ad-hoc retrieval, SIGIR 2018

Pretraining Method tailored for IR

- Pre-training for Passage Retrieval in QA: ICT, BFS, WLP (Chang et.al, 2019)
- The underlying belief : **using a pre-training objective that more closely resembles the downstream task could lead to better fine-tuning performance**

- SOTA in ad-hoc retrieval: **P**re-training with **R**epresentative w**O**rds **P**rediction (**PROP**)
  - Key idea: construct the ***representative words prediction (ROP)*** task for pre-training inspired by the query likelihood (QL) model

MS MARCO Document Ranking Leaderboard

Search:

date		description	team	paper	code	type	MRR@100 (Dev)	MRR@100 (Eval)
2021/05/24	🏆	ANCE MaxP + LongP / SEED-Encoder+LongP (ensemble)	Soonhwan Kwon,Minyoung Lee, Samsung SDS AI Research			full ranking	0.487	0.427
2021/04/25	🏆	PROP_step400K base + doc2query top1000(ensemble v0.2)	Yingyan Li, Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Xueqi Cheng - ICT, CAS	[paper]		full ranking	0.479	0.423
2021/04/28		Knowledge Retrieval	HuaweiPoissonLab, RUCIR			full ranking	0.482	0.423
2021/05/10		Knowledge Retrieval	HuaweiPoissonLab, RUCIR			full ranking	0.484	0.423
2021/05/26		Thinking Reranker (single)	Tongyuan - KCAI-Lab			full ranking	0.485	0.422
2021/04/27		ANCE BS+GL	Jiajia Ding*, Chunyu Li* - PingAn			full ranking	0.489	0.421

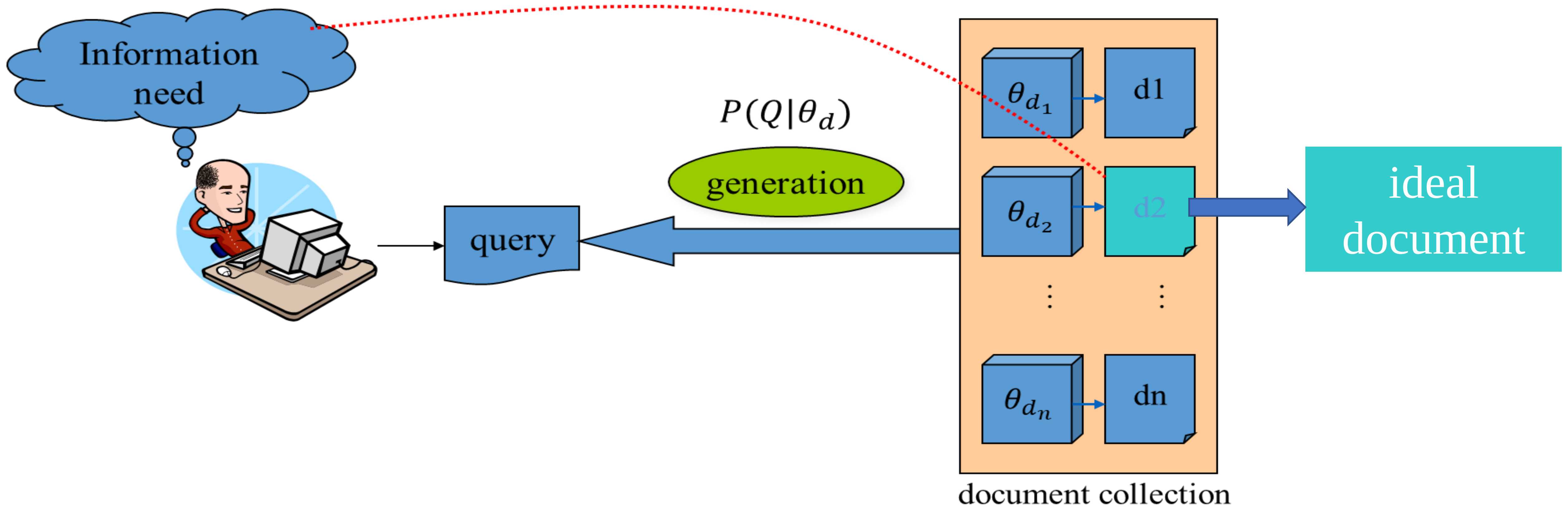
Pre-training Tasks for Embedding-based Large-scale Retrieval, ICLR 2019

PROP: Pre-training with Representative Words Prediction for Ad-hoc Retrieval, WSDM'2021

PROP

- **Inspired by the "Old" Hypothesis in IR**

- A hypothesis underlying *the Query Likelihood model*
  - The user has a reasonable idea of the terms that are likely to appear in the **"ideal" document** that can satisfy his/her information need
  - The query is **generated** as the piece of text representative of the "ideal" document



- **Rank documents** based on the probability that they "**generate**" the query

$$score(Q, D) = P(Q|\theta_D)$$

Document language model

$$= \prod_{w \in V} P(w|\theta_D)^{c(w,Q)}$$

Multinomial unigram language model

A language modeling approach to information retrieval, SIGIR 1998

The ROP Pre-training Task for Ad-hoc Retrieval

- **Representative words prediction** (**ROP**) task: Pre-train the Transformer model to predict the pairwise preference between word sets with respect to their representativeness to the document
  ① **Paired Sampling**: Sample pairs of word sets according to *document language model*
    - Sample words: $S_1 = (w_1, w_2, \dots, w_x)$, $w_i \sim P(w_i|\theta_D)$
  ② **Preference Learning**: The word set with higher likelihood is deemed as more "representative" of the document
    - Compute likelihood: $P(S_1|\theta_D) = \prod_{w \in V} P(w|\theta_D)^{c(w,Q)}$

Unigram language model with Dirichlet prior smoothing

Information retrieval (IR) is the process of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing. Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for the metadata that describes data, and for databases of texts, images or sounds. Automated information retrieval systems are used to reduce what has been called information ……

Document

Pre-trained LM

IR searchers > used to reduce

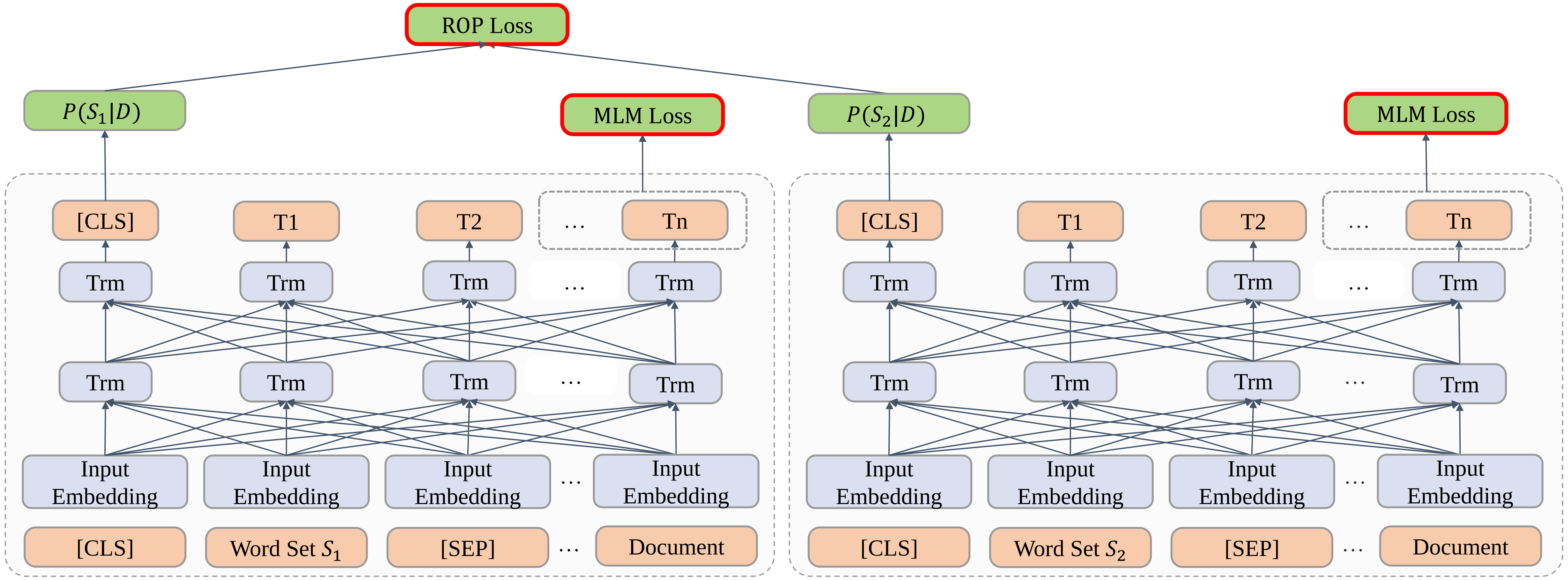
a collection < Web systems

Representativeness Preference

PROP: Pre-training with Representative Words Prediction for Ad-hoc Retrieval, WSDM'2021

PROP: Pre-training with Representative wOrds Prediction

- Pre-train the Transformer towards the following two objectives
  - **The ROP objective**
    - Sample a pair of word sets, suppose set S_1 has a higher likelihood score than $S2$
    - Pairwise Loss: $\mathcal{L}_{ROP} = \max(0, 1 - P(S_1|D) + P(S_2|D))$
  - **The MLM objective**
    - Masks out some tokens from the input
    - Cross-entropy Loss: $\mathcal{L}_{MLM} = -\sum_{\hat{x} \in X} \log p(\hat{x}|X_{\backslash \hat{x}})$



Back to Sampling Process

- The success of PROP heavily relies on the quality of the sampled "representative" words
- But the sampling process of PROP is according to a **unigram language model $\boldsymbol{\theta_D}$**
  - **Term independent assumption:** $\theta_D = \prod_{i=1}^{|D|} P(w_i) = P(w_1)P(w_2)P(w_3)$ …
  - Difficult to fully capture the document semantic by **ignoring the correlation between words**

pulmonary fibrosis synonyms interstitial pulmonary fibrosis a chest x-ray demonstrating pulmonary fibrosis believed to be due to amiodarone. specialty pulmonology pulmonary fibrosis (literally ""scarring of the lungs "") is a respiratory disease in which scars are formed in the lung tissues, leading to serious breathing problems. scar formation, the accumulation of excess fibrous connective tissue (the process called fibrosis), leads to thickening of the walls, and causes reduced oxygen supply in the blood. as a consequence patients suffer from perpetual shortness of breath. [1]in some patients the…
Unigram Language Model: fibrosis, uip, interstitial, idiopathic, pulmonary, fibrous, inspiratory, auscultation, pulmonology, amiodarone, crackl,

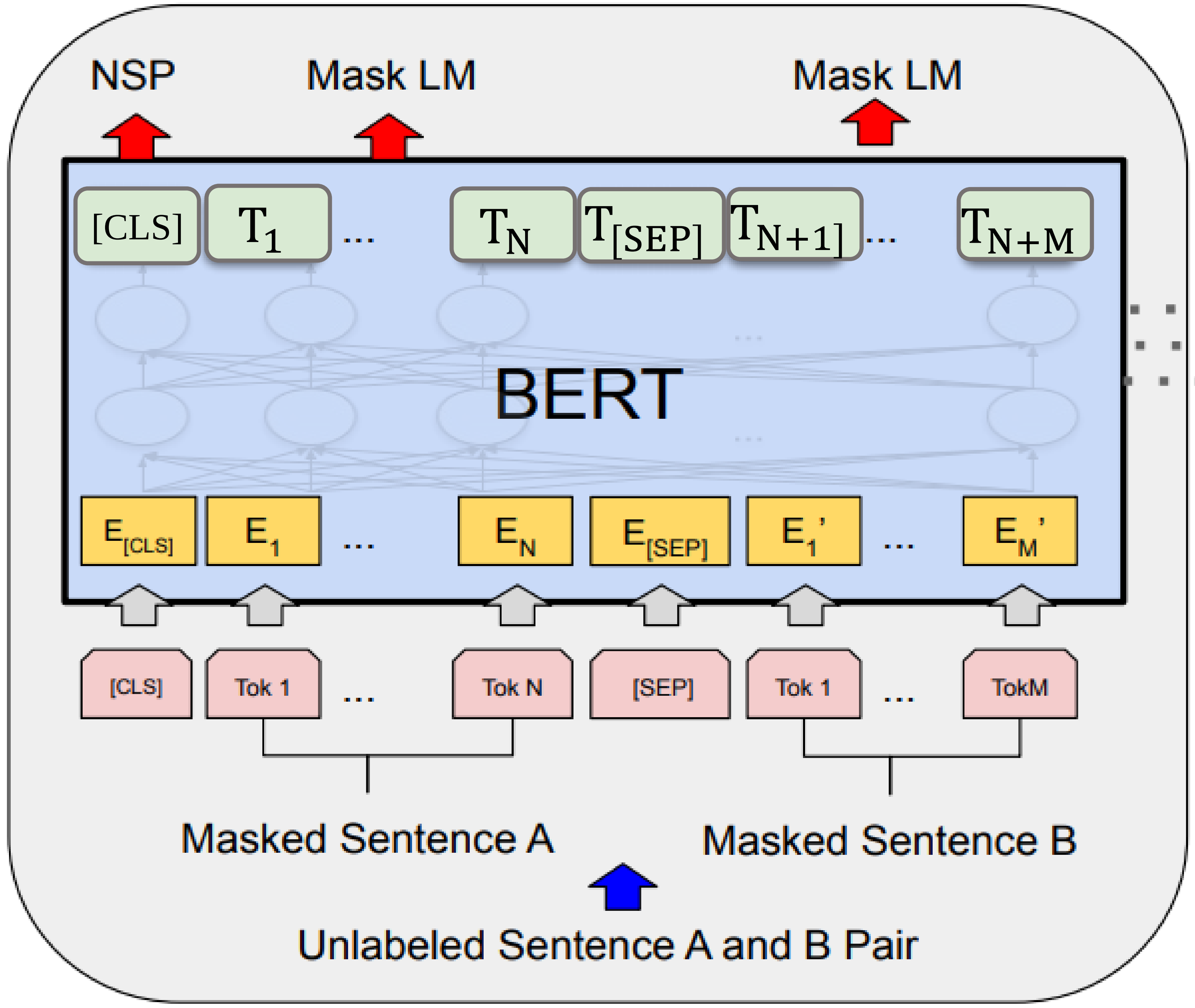
- Tend to **favor rare words** in a document which may not be representative to the document semantic

Can we leverage a better document language model for higher quality of representative words sampling?

PROP: Pre-training with Representative Words Prediction for Ad-hoc Retrieval, WSDM'2021

Contextual Language Model: BERT

- Contextual language encoding
  - Each token in BERT accumulates the information from both left and right context to enrich its representation
- Success in language semantic tasks (sentence, sentence pair, document)
  - Semantic textual similarity: STS, MRPC
  - Text classification: AGNews, DBpedia
  - Document sentiment: IMDB, Yelp

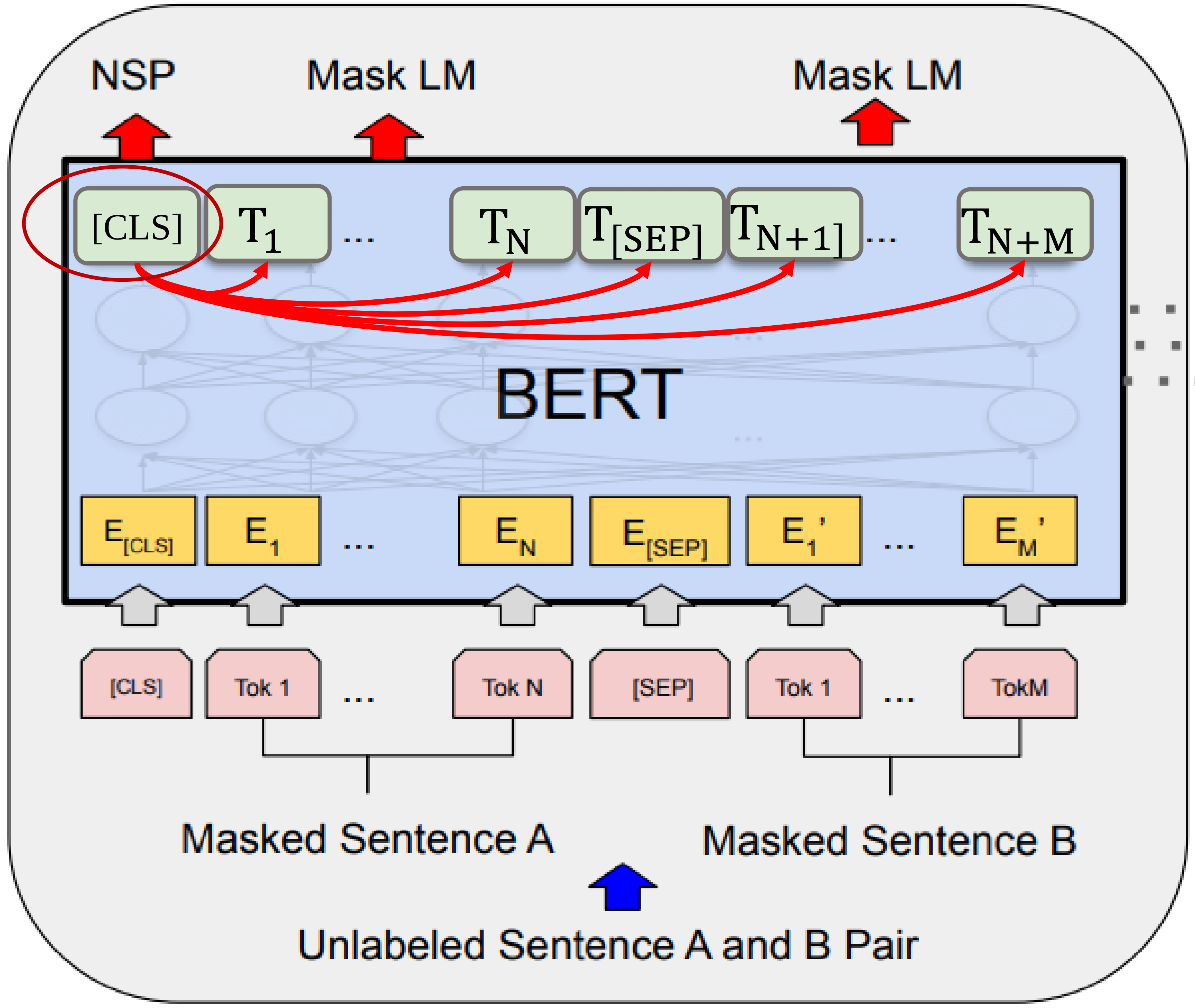


BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL'2019

B-PROP: Bootstrapped Pre-training with Representative wOrds Prediction

- **Key idea:** Leverage **BERT** to replace the classical unigram language model for the ROP task construction, and re-train BERT itself towards the tailored objective for IR

- **An intuitive solution**
  - **The special classification token [CLS]** is an aggregate of the entire sequence representation
  - **[CLS]-Token attention** indicates how much meaningful information a particular token contributes to the entire sequence



- Directly sample representative words according to BERT's [CLS]-Token attention
  - Summing up and re-normalizing the vanilla attention weights over distinct terms

ROP construction with BERT

- We found that the Vanilla [CLS]-Token attention-based term distribution can **generate representative words, but also favor common words**

pulmonary fibrosis synonyms interstitial pulmonary fibrosis a chest x-ray demonstrating pulmonary fibrosis believed to be due to amiodarone. specialty pulmonology pulmonary fibrosis (literally ""scarring of the lungs "") is a respiratory disease in which scars are formed in the lung tissues, leading to serious breathing problems. scar formation, the accumulation of excess fibrous connective tissue (the process called fibrosis), leads to thickening of the walls, and causes reduced oxygen supply in the blood. as a consequence patients suffer from perpetual shortness of breath. [1]in some patients the specific cause of the disease can be diagnosed, but in others the probable cause cannot be determined, a condition called idiopathic pulmonary fibrosis. there is …
Vanilla Attention-based Term Distribution: pulmonary, fibrosis, in, interstitial, the, of, disease, can, and, lung, chest, is, cause, to,

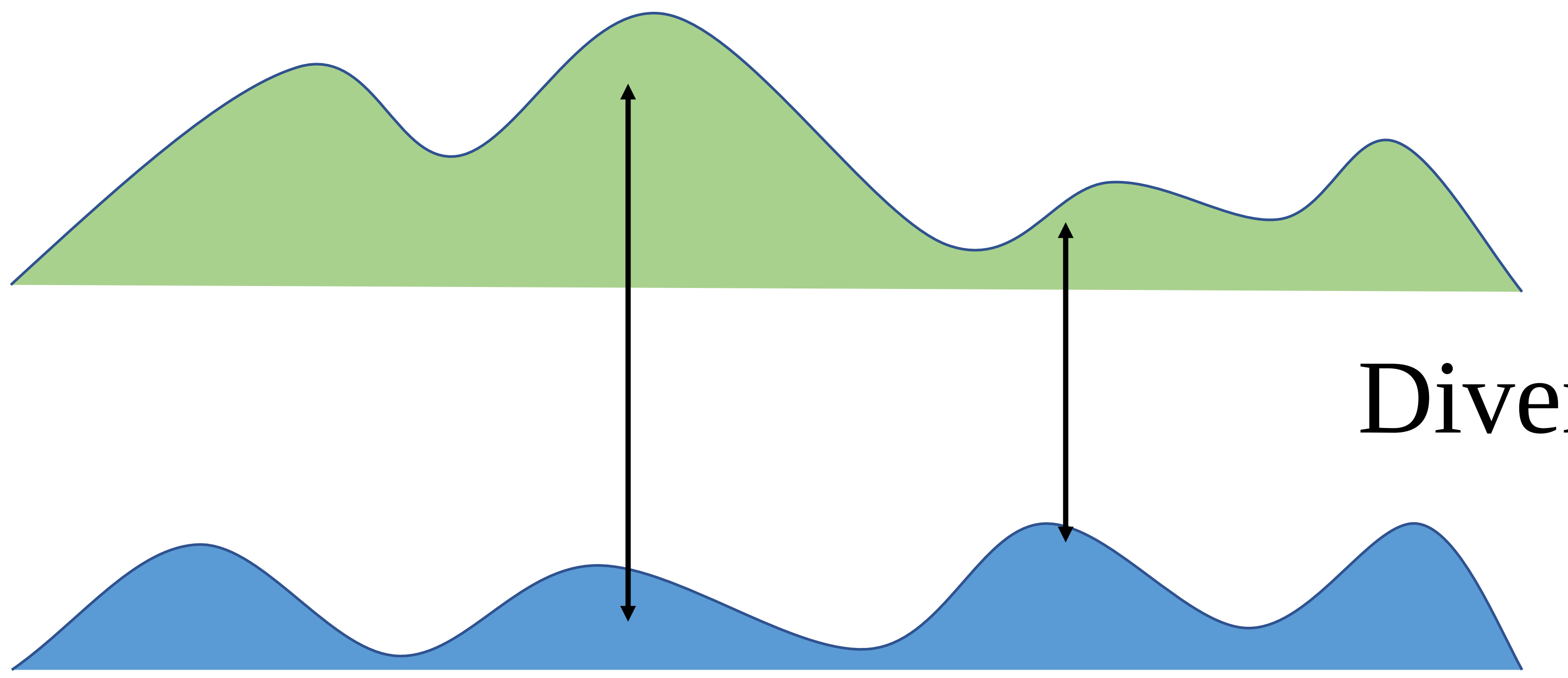
- The underlying reason
  - BERT focuses on encoding as much semantic information in a document as possible
  - The term distribution obtained from its vanilla attention is a **semantic distribution**, but **not necessarily a representative/informative distribution**

Divergence from Randomness

- The **informativeness/representativeness** of a term could be computed by **measuring the divergence** between a term distribution produced by **a random process** and **the actual term distribution** in a document (Amati and Rijsbergen, 2002)

Actual Term Distribution $P(w|\theta_d)$

The Target Document LM



Divergence

Random Term Distribution $P(w|\theta_{random})$

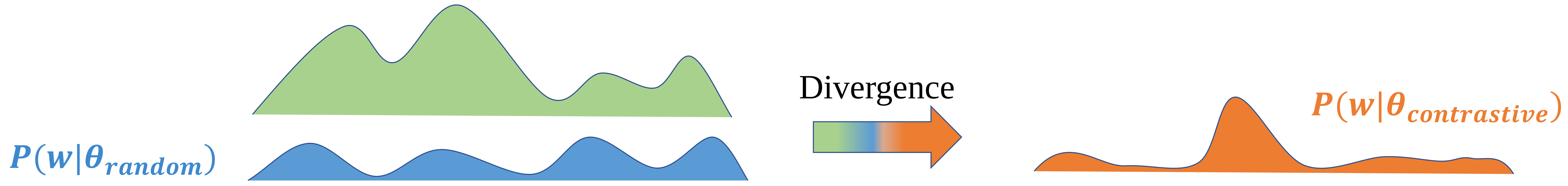
A Random Document LM

Contrastive Sampling for ROP with BERT

- We introduce a novel **contrastive** method to leverage BERT's attention mechanism to sample representative words from a document
- **Contrastive Term Distribution:**
  - Compute the **cross-entropy (i.e., the divergence)** between the document term distribution $P(w_k|\theta_d)$ and the random term distribution $P(w_k|\theta_{random})$

$$\gamma_{w_k} = CE(\theta_d|\theta_{random}) = -P(w_k|\theta_d)log_2 P(w_k|\theta_{random}))$$

$$P(w_k|\theta_{contrastive}) = \frac{\exp(\gamma_{w_k})}{\sum_{w_k \in V} \exp(\gamma_{w_k})}$$



Contrastive Sampling for ROP with BERT

- **Document Term Distribution**
  - Average multi-head attention weight, and sum up the weight of the same term over different positions

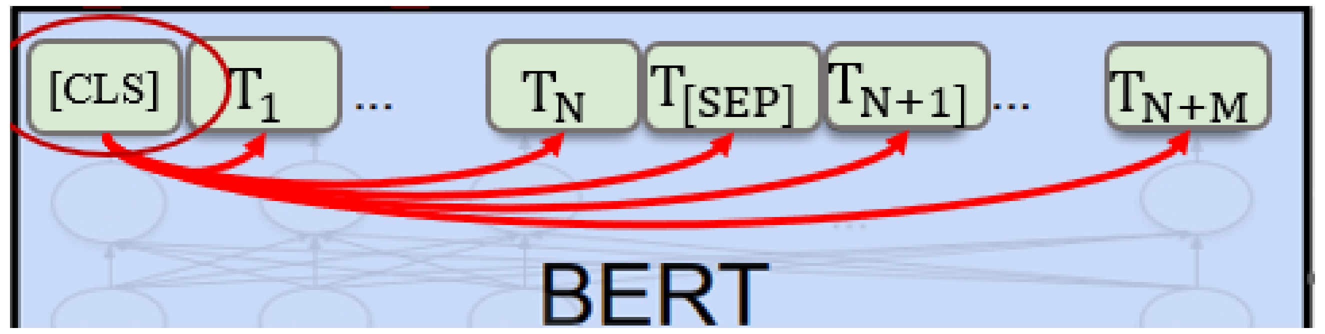
$$a^t = \frac{1}{h}\sum_{i=1}^{h} a_i^t\,, \qquad where\ a_i^t = softmax(\frac{Q_i^{[CLS]} * K_i^{x_t}}{\sqrt{d/h}})$$

$$\beta_{w_k} = \sum_{x_t = w_k} a^t\,, x_t \in d, \text{i.e., word } x \text{ in postion } t$$

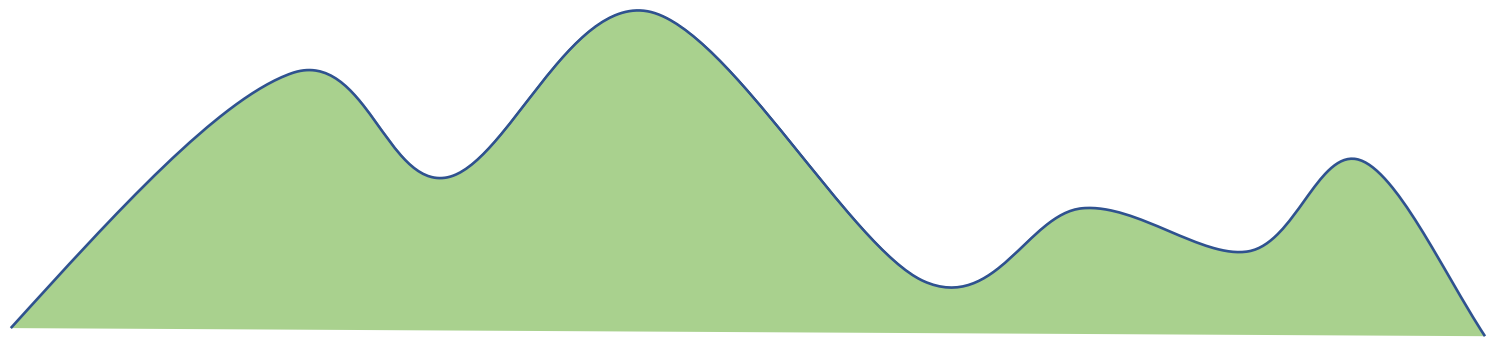
  - Saturate and re-normalize the vanilla [CLS]-Token attention weights over distinct terms

$$P(w_k|\theta_d) = softmax\left(\frac{\beta_{w_k}}{b + \beta_{w_k}}\right), \text{where } b \text{ is a hypeparameter}$$

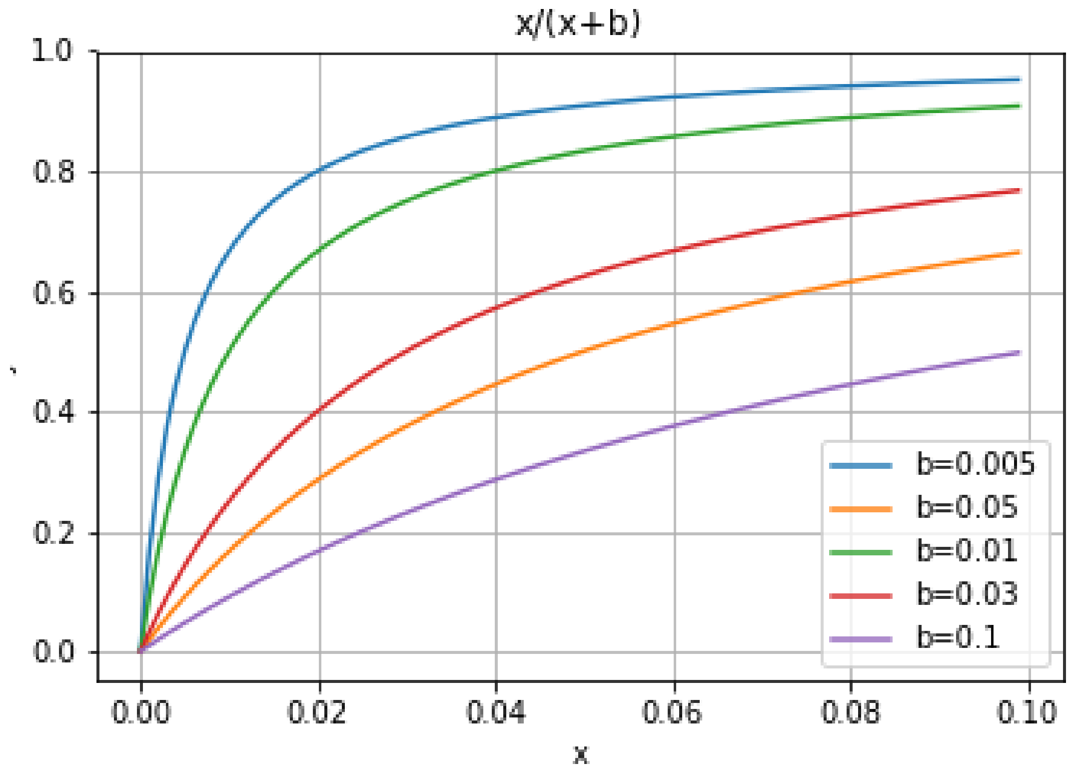
Document: pulmonary, fibrosis, synonyms, interstitial, ……



[CLS]-Token attention weights: 0.2, 0.21, 0.04, 0.07,0.09, …



Document Term Distribution $\boldsymbol{P(w|\theta_d)}$

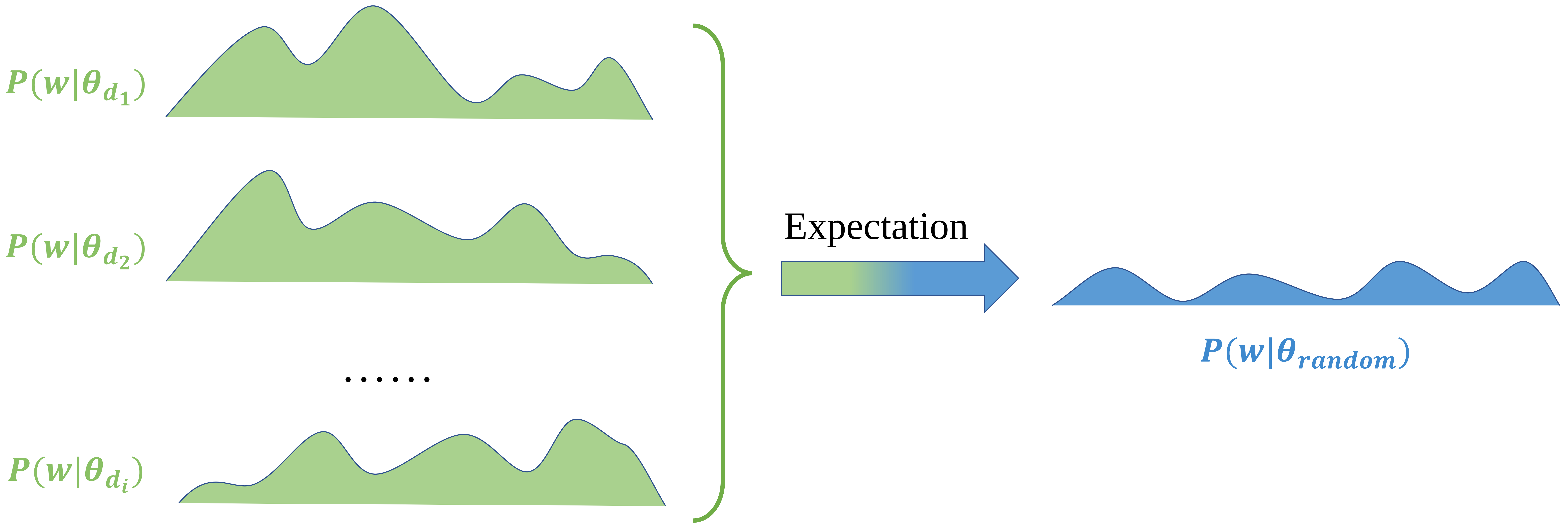


The term saturation function is used to alleviate that the document is dominated by terms with large attention weights

Contrastive Sampling for ROP with BERT

- **Random Term Distribution**:
  - Take an expectation over all the term distributions in the document collection

$$P(w_k|\theta_{random}) = \mathbb{E}(w_k|\mathcal{D}) = \frac{1}{|\mathcal{D}|} = \sum_{d \in \mathcal{D}} P(w_k|\theta_d)$$



Contrastive Sampling for ROP with BERT

pulmonary fibrosis synonyms interstitial pulmonary fibrosis a chest x-ray demonstrating pulmonary fibrosis believed to be due to amiodarone. specialty pulmonology pulmonary fibrosis (literally ""scarring of the lungs "") is a respiratory disease in which scars are formed in the lung tissues, leading to serious breathing problems. scar formation, the accumulation of excess fibrous connective tissue (the process called fibrosis), leads to thickening of the walls, and causes reduced oxygen supply in the blood. as a consequence patients suffer from perpetual shortness of breath. [1]in some patients the specific cause of the disease can be diagnosed, but in others the probable cause cannot be determined, a condition called idiopathic pulmonary fibrosis. there is …	
Unigram Language Model:	fibrosis, uip, interstitial, idiopathic, pulmonary, fibrous, inspiratory, auscultation, pulmonology, amiodarone, crackl,
Vanilla Attention-based Term Distribution:	pulmonary, fibrosis, in, interstitial, the, of, disease, can, and, lung, chest, is, cause, to,
Contrastive Term Distribution:	fibrosis, pulmonary, interstitial, idiopathic, lung, chest, disease, diseases, cause, patients, x-ray, scars,

- Contrastive term distribution can now obtain representative words for a document
  - By eliminating the impact of the common words, i.e., stop words, using the contrastive method
- Contrastive term distribution is better than unigram language model in terms of representativeness
  - (lung, chest,…) ***vs.*** (inspiratory, auscultation,…)

B-PROP Learning Objective

- **Re-train BERT** towards the tailored objective for IR

$$\mathcal{L}_{total} = \mathcal{L}_{ROP} + \mathcal{L}_{MLM}$$

- **ROP**: pairwise preference prediction objective
  - Sample a pair of word sets, suppose set S_1 has a higher likelihood score than $S2$
  - Pairwise Loss: $\mathcal{L}_{ROP} = \max(0, 1 - P(S_1|D) + P(S_2|D))$
- **MLM**: masked tokens prediction objective
  - Masks out some tokens from the input
  - Cross-entropy Loss: $\mathcal{L}_{MLM} = -\sum_{\hat{x}\in X} \log p(\hat{x}|X_{\backslash \hat{x}})$

Experiment Setting

- Pretraining datasets：
  - Wikipedia, over 10 million documents
  - MS MARCO, about 3.4 million documents
- 5 downstream ad-hoc retrieval tasks：
  - 5 small datasets: Robust04, ClueWeb09-B, Gov2, MQ2007, MQ2008
  - 2 large-scale datasets: MS MARCO Document ranking and TREC DL Document ranking
- Baseline models:
  - Traditional retrieval models: BM25, QL
  - Neural-IR models: DRMM, Conv-KNRM
  - Other pretraining method: PROP, BERT, $\text{Transformer}_{\text{ICT}}$

Dataset	#genre	#queries	#docs
Robust04	News	250	0.5M
ClueWeb09-B	web pages	150	50M
Gov2	.gov pages	150	25M
MQ2007	.gov pages	1692	25M
MQ2008	.gov pages	784	25M
MS MARCO	web pages	**0.37M**	3.2M
TREC DL	web pages	**0.37M**	3.2M

Main Results on Small datasets

Table 3: Performance Comparisons between B-PROP and the baselines on 5 small datasets. Two-tailed t-tests demonstrate the improvements of B-PROP to the best baseline PROP are statistically significant ($*$ indicates $p \leq 0.05$).

Model Type	Model Name	Robust04		ClueWeb09-B		Gov2		MQ2007		MQ2008	
		nDCG@20	P@20	nDCG@20	P@20	nDCG@20	P@20	nDCG@10	P@10	nDCG@10	P@10
Traditional Retrieval Models	QL	0.413	0.367	0.225	0.326	0.409	0.510	0.423	0.371	0.223	0.241
	BM25	0.412	0.363	0.230	0.334	0.421	0.523	0.414	0.366	0.220	0.245
Neural IR Models	DRMM	0.425	0.371	0.246	0.349	0.457	0.545	0.441	0.382	0.221	0.248
	Conv-KNRM	0.414	0.360	0.238	0.336	0.462	0.552	0.431	0.377	0.215	0.239
Pre-trained Models	BERT	0.459	0.389	0.295	0.367	0.495	0.586	0.506	0.419	0.247	0.256
	Transformer_{ICT}	0.460	0.388	0.298	0.369	0.499	0.587	0.508	0.420	0.245	0.256
	PROP_{Wiki}	0.502	0.421	0.316	0.384	0.519	0.593	0.523	0.432	0.262	0.267
	PROP_{MARCO}	0.484	0.408	0.329	0.391	0.525	0.594	0.522	0.430	0.266	0.269
Our Approach	B-PROP_{Wiki}	**0.519***	**0.430***	0.331	0.393	0.534*	0.599*	**0.529***	0.436*	0.271*	0.273
	B-PROP_{MARCO}	0.510*	0.429*	**0.353***	**0.407***	**0.552***	**0.606***	**0.529***	**0.439***	**0.273***	**0.275***

- B-PROP perform better than PROP and BERT on small datasets by a large margin
- Pre-training on a similar domain leads to better fine-tuning performance

Main Results on Large datasets

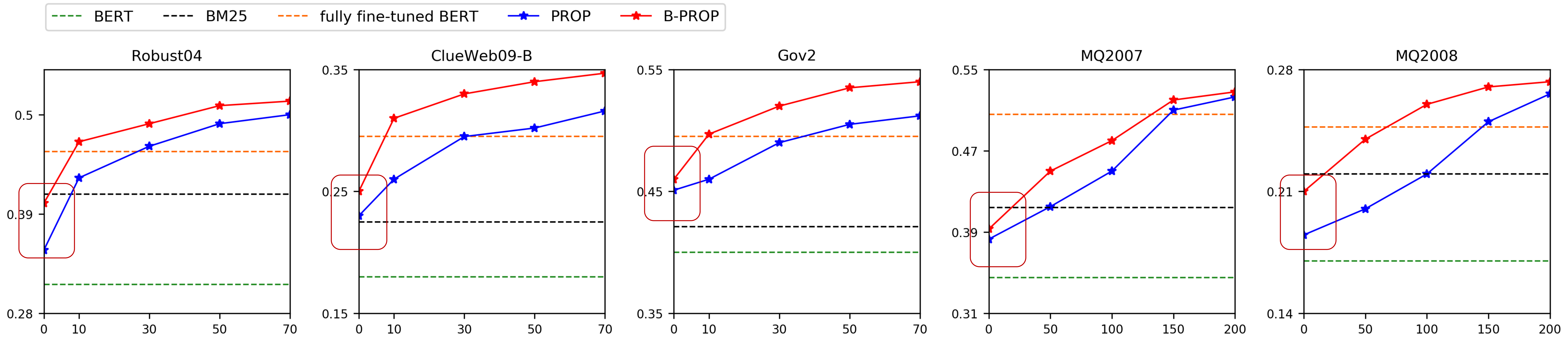
Table 4: Comparisons between B-PROP and the baselines on 2 large-scale datasets. Two-tailed t-tests demonstrate the improvements of B-PROP to the best baseline PROP are statistically significant ($*$ indicates $p \leq 0.05$).

Model Type	Model Name	MS MARCO				TREC DL			
		rerank		fullrank		rerank		fullrank	
		MRR@10	MRR@100	MRR@10	MRR@100	nDCG@10	nDCG@100	nDCG@10	nDCG@100
Traditional Retrieval Models	QL	-	-	0.287	0.300	-	-	0.600	0.559
	BM25	-	-	0.315	0.326	-	-	0.592	0.552
Neural IR Models	DRMM	0.137	0.152	0.164	0.197	0.249	0.390	0.301	0.422
	Conv-KNRM	0.155	0.179	0.183	0.225	0.311	0.476	0.360	0.456
Pre-trained Models	BERT	0.391	0.397	0.410	0.418	0.642	0.519	0.657	0.567
	Transformer_{ICT}	0.394	0.399	0.411	0.423	0.639	0.521	0.658	0.569
	PROP_{Wiki}	0.401	0.405	0.419	0.427	0.654	0.533	0.662	0.572
	PROP_{MARCO}	0.410	0.415	0.426	0.435	0.668	0.547	0.676	0.573
Our Approach	B-PROP_{Wiki}	0.415*	0.419*	0.428	0.439*	0.670	0.552*	0.679	0.581*
	B-PROP_{MARCO}	**0.419***	**0.423***	**0.437***	**0.441***	**0.675***	**0.558***	**0.694***	**0.590***

- B-PROP perform better than PROP and BERT on large datasets
- The performance trend of $\text{B}-\text{PROP}_{Wiki}$ and $\text{B}-\text{PROP}_{MARCO}$ on small datasets and large-scale datasets is consistent

Zero- and Low-Resource Setting

- **Black dashed line: BM25**
- **Green dashed line: BERT**
- **Orange dashed line: Fully fine-tuned BERT**
- **Blue solid line: PROP**
- **Red solid curve: B-PROP**



- B-PROP outperforms PROP significantly on all the datasets using the same number of limited supervised data
- B-PROP fine-tuned on limited supervised data can achieve comparable/better performance against BERT fine-tuned on full supervised data
- Under the zero-resource setting, B-PROP could outperform PROP on all the datasets

Conclusion

- B-PROP leverages the powerful contextual language model BERT to replace the unigram language model for the ROP task construction.
- We introduced a contrastive method to obtain the representativeness distribution.
- B-PROP can achieve significant improvements over PROP and BERT, and further push forward the SOTA on a variety of ad-hoc retrieval tasks.

Code and the pre-training models are released at:
https://github.com/Albert-Ma/PROP

An awesome paper list about pretraining for IR :
https://github.com/Albert-Ma/awesome-pretrained-models-for-information-retrieval

Thanks!

Xinyu Ma

maxinyu17g@ict.ac.cn