中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

面向信息检索的预训练和微调方法研究

学生：马新宇
导师：郭嘉丰
单位：中科院网络数据科学与技术重点实验室

目录

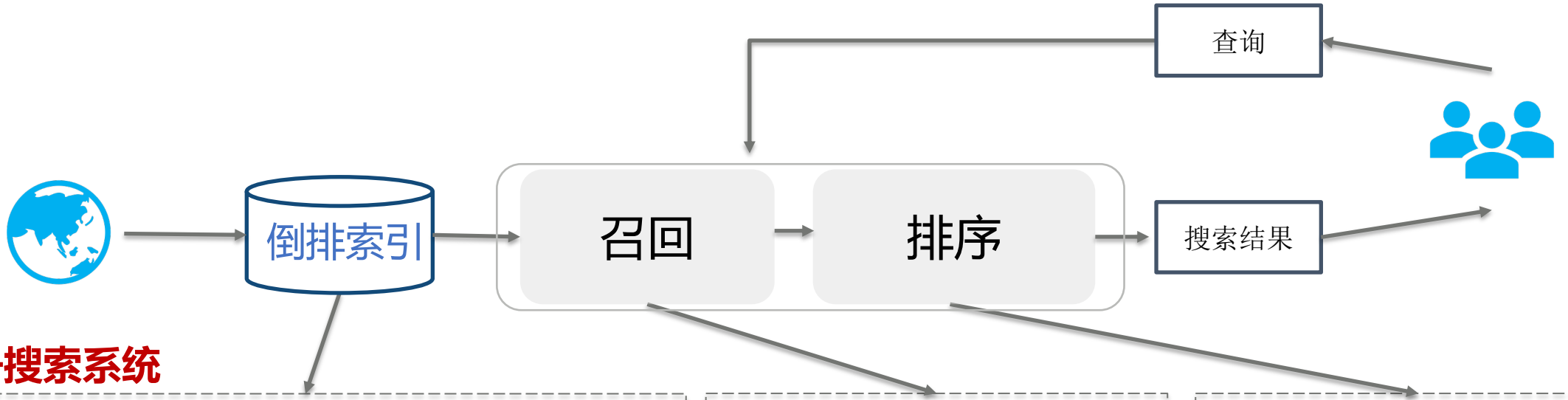
信息检索

- 信息检索是从信息资源集合获得与信息需求**相关**的信息资源的活动。

—维基百科

信息检索的核心问题：查询和文档的相关性匹配!

索引-召回-排序 的多级流程



符号搜索系统

基于词项的方法

- **PIV** (向量空间模型, 1996)
- **QL** (统计语言模型, 1999)
- **BM25** (经典概率模型, 1994)

核心思想：使用精确匹配信号设计相关性评分函数

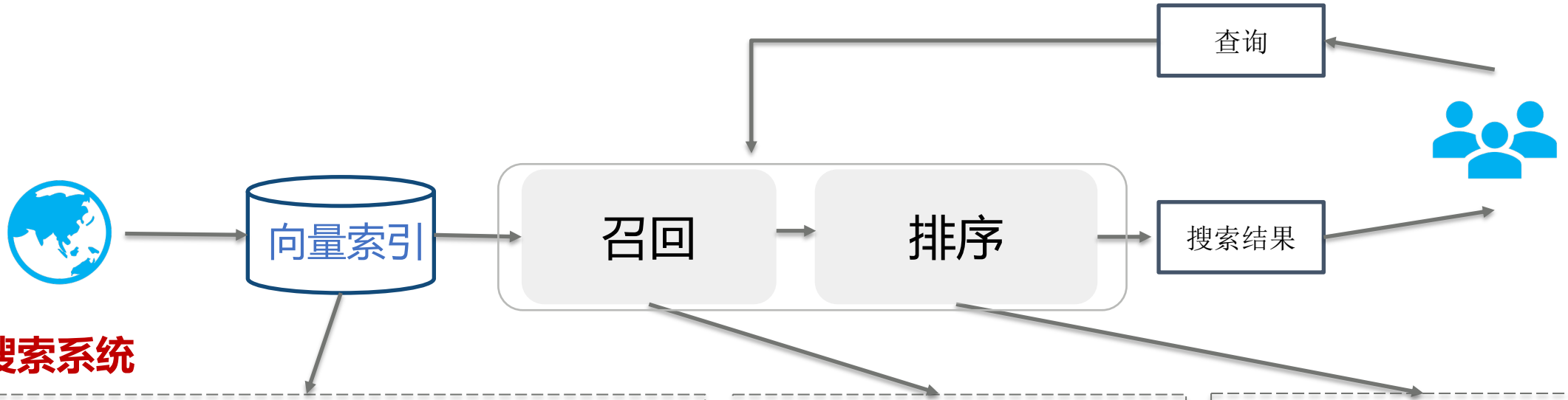
排序学习方法

- **RankNet** (1996)
- **RankSVM** (2001)
- **LambdaMart** (2010)

核心思想：使用基于人工定义特征的监督机器学习方法来排序问题。

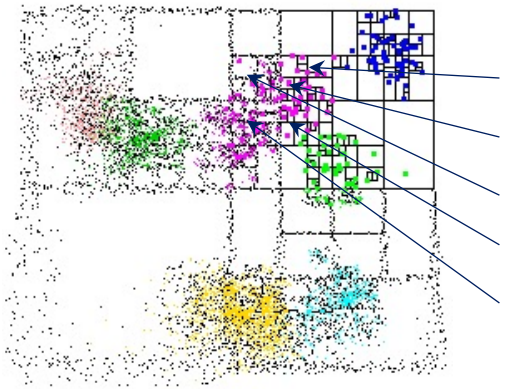
Introduction to Information Retrieval, Cambridge University Press. 2008
Learning to Rank for Information Retrieval, FnTIR 2009

索引-召回-排序 的多级流程



神经搜索系统

clusters



文档

D1: He likes to wink, he likes to drink.

D2: He likes to drink, and drink, and drink.

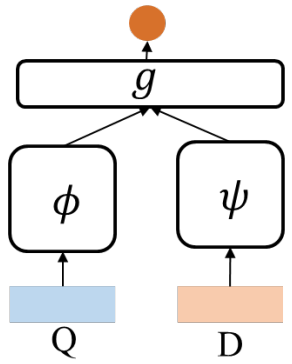
D3: The thing he likes to drink is ink.

D4: The ink he likes to drink is ping.

D5: He drinks to wink and drink pingk ink.

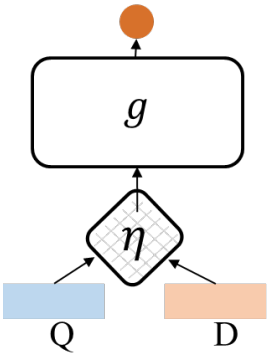
基于表达的结构

$f(q,d) = g(\psi(q), \phi(d))$



基于交互的结构

$f(q,d) = g(\eta(q,d))$



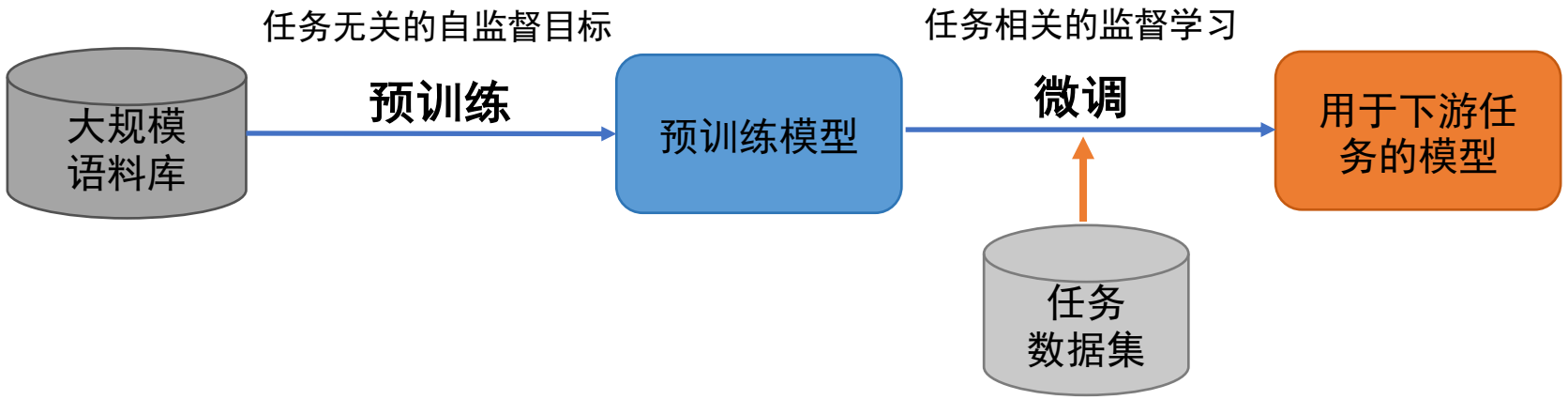
A Deep Look into Neural Ranking Models for Information Retrieval, IPM 2019

预训练-微调的新范式

“**预训练和微调**” 的新范式在 NLP 领域取得了重大的成功

① 预训练（任务无关）：在大规模文本语料库上使用自监督任务学习的通用语言模型

② 微调（任务相关）：针对不同的下游任务，使用带标注的监督学习方法，微调预训练模型



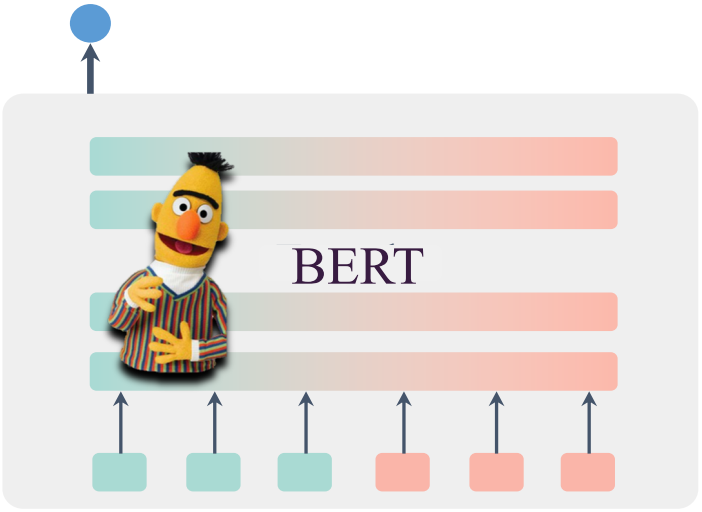
优势

① 在各种下游的NLP任务上取得 **SOTA** 效果

② 一种可以使各种 NLP 任务受益的**通用范式**

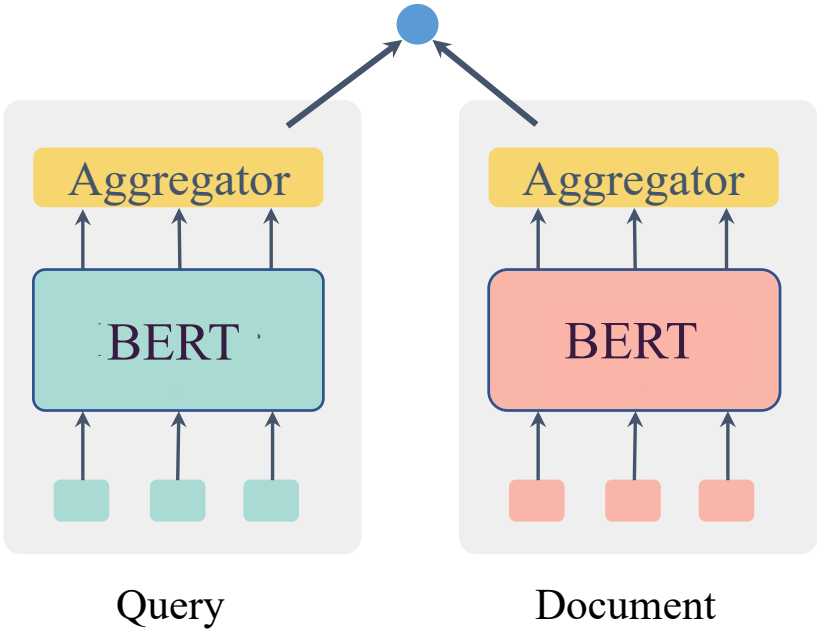
BERT在检索上的直接应用

排序阶段



基于交互的结构

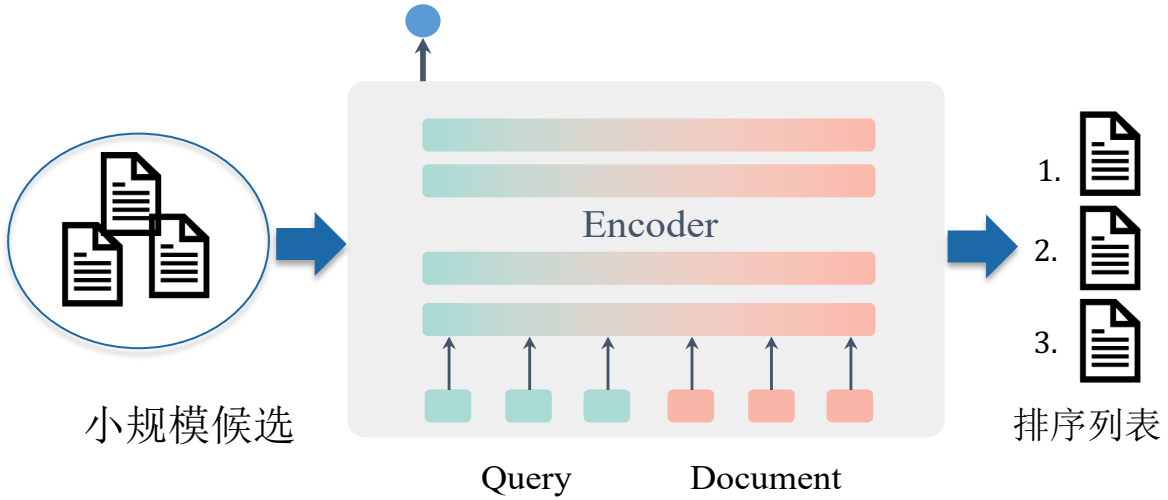
召回阶段



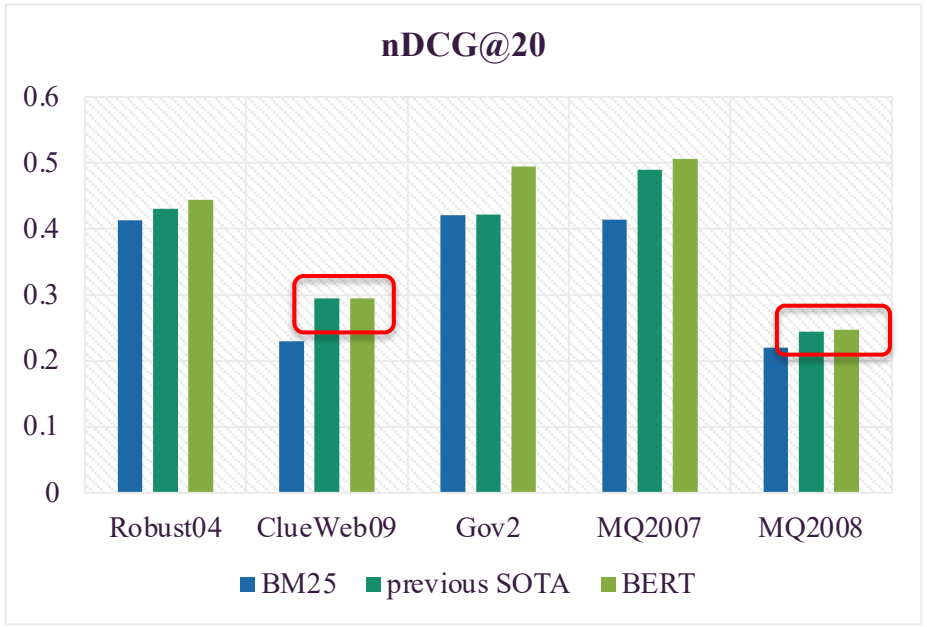
基于表达的结构

Pretraining Methods in Information Retrieval, FnTIR 2022

预训练模型在排序上的效果



基于交互的结构



预训练模型确实有利于搜索任务，但有时改进有限！

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

面向NLP的预训练目标和检索需求的差异

面向NLP的预训练目标 *VS.* **检索的需求**

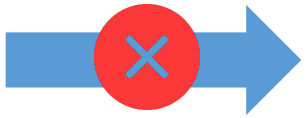
- 序列内任务:
  - 掩码语言模型 (BERT)
  - 排列语言模型 (XLNET)

词粒度的上下文表示

查询和文档的良好的上下文表示

- 序列间任务:
  - 下一句话预测任务(BERT
  - 顺序预测任务(ALBERT)

句间的语义连贯性



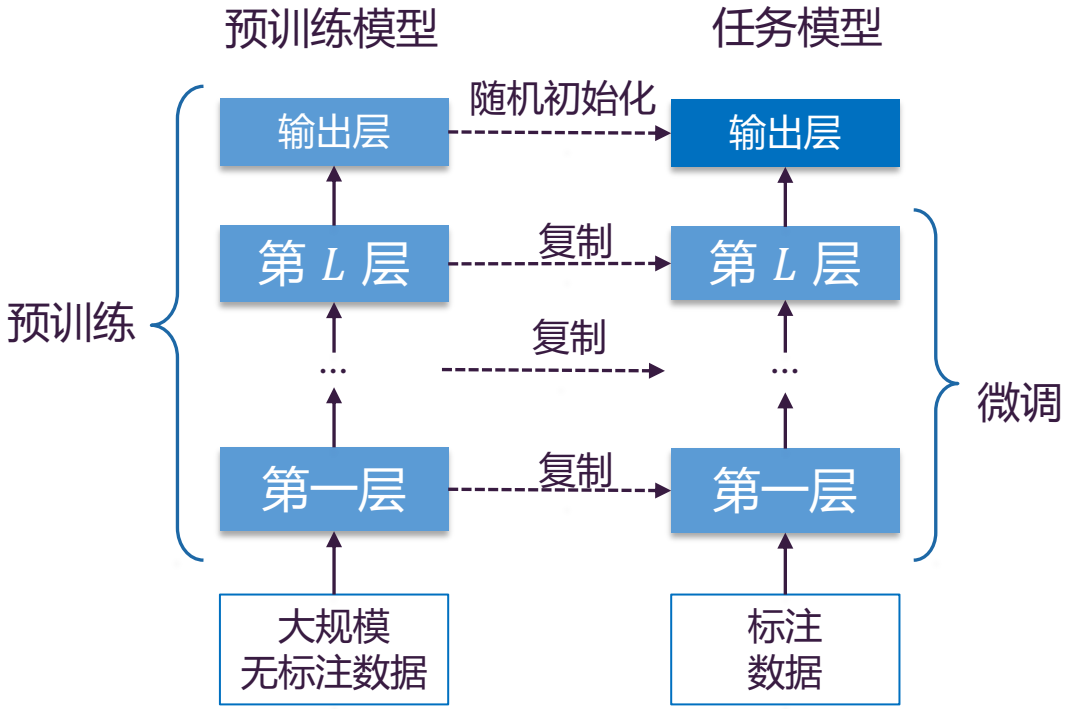
相关性建模

x 句子对 vs. 短查询和长文档

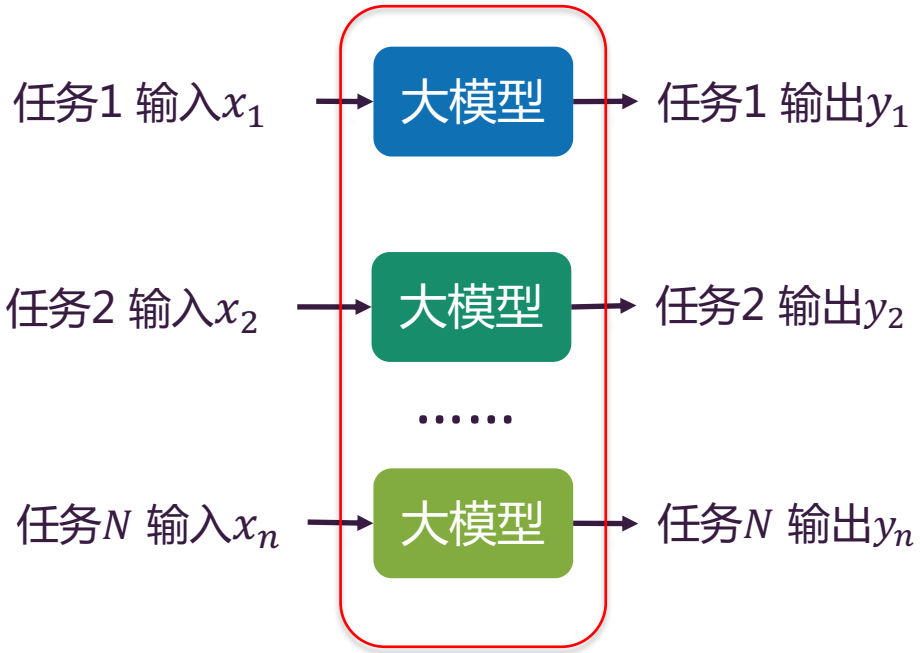
x 连贯性 vs. 相关性

面向NLP的预训练未考虑建模信息检索的核心需求相关性！

主流的微调方法



主流的完全微调方法



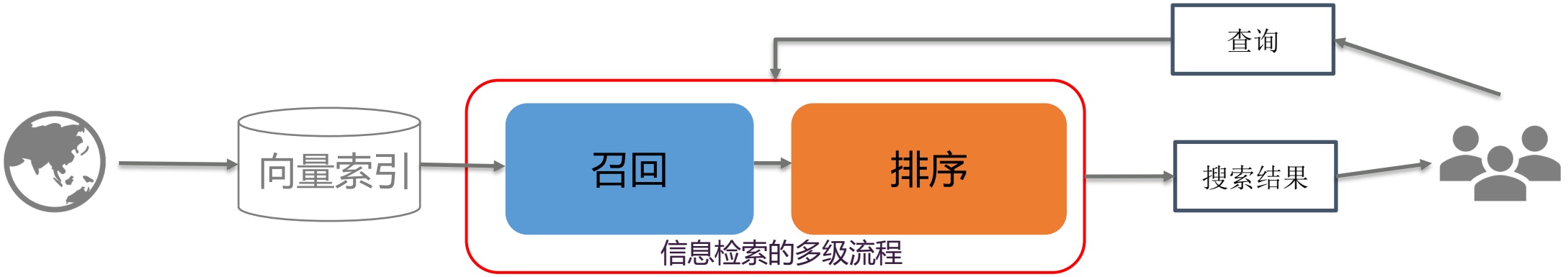
N 个任务需要 ***N*** 个大模型同时在线服务

为多个检索任务完全微调大模型在生产环境中部署代价高、效率低！

已有方法的问题

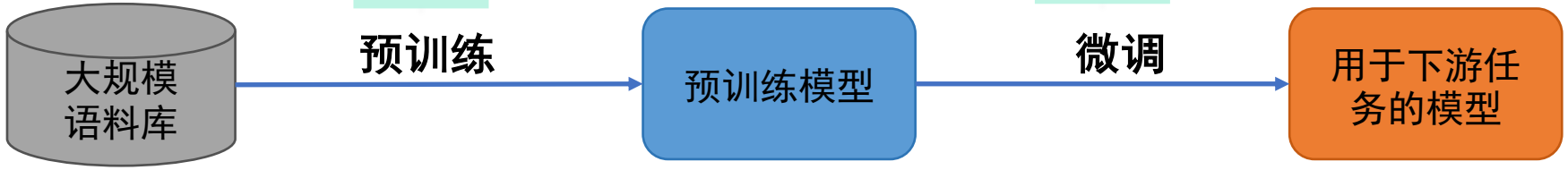
面向**NLP**设计的预训练和微调方法在**IR**上面临着**效果**和**效率**两大问题！

研究目标



信息检索的多级流程

面向信息检索，对预训练和微调方法进行适配和定制，
提升大模型的效果和效率！



预训练-微调的新范式

科学挑战一

- 信息检索中的核心概念**相关性**至今没有统一的明确定义

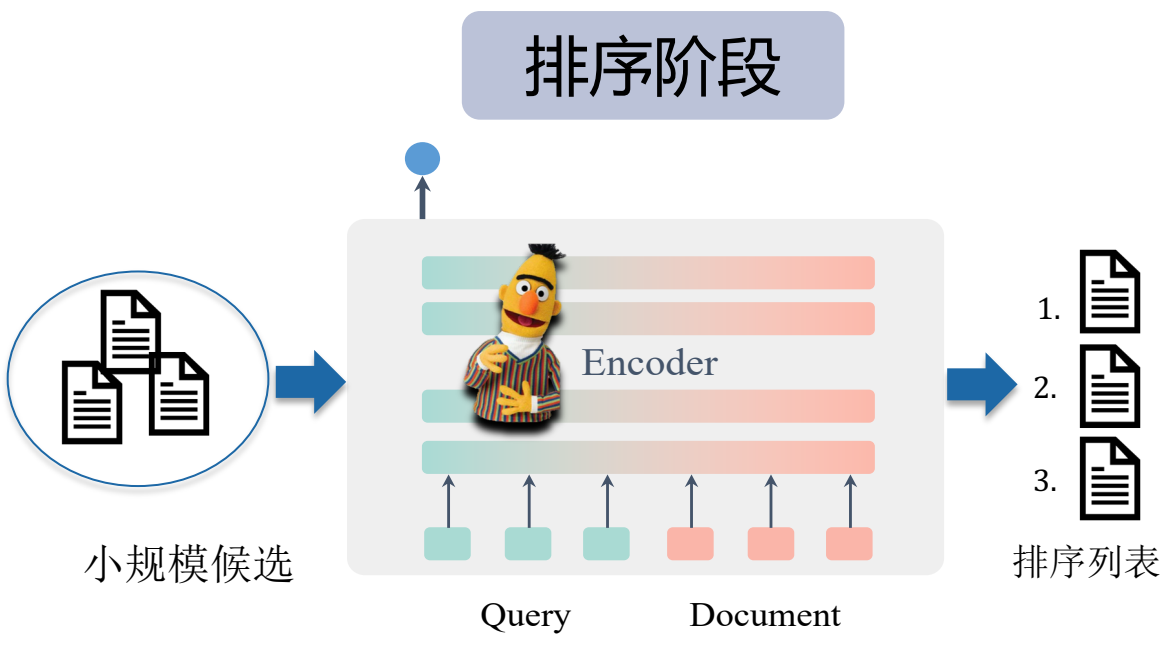
查询：元宇宙

“元宇宙”翻译自英文“Metaverse”一词，由前缀「meta」（意为「元」）和词根「verse」（源于 universe「宇宙」）组成，直译而来便是“元宇宙”。这一概念最早出自于尼尔·斯蒂芬森 1992 年出版的科幻小说《雪崩》(Snow Crash)，指在一个脱离于物理世界，却始终在线的平行数字世界中，人们能够在其中以虚拟人物角色 (avatar) 自由生活。
业界现在对元宇宙目前尚无一个统一的定义，每个人对元宇宙都有不一样的解读和理解，泡咖认为目前比较有共识的元宇宙定义是：元宇宙是具有高度沉浸感的虚拟世界，和现实世界一样，人们可以在里面社交，学习，娱乐，生活，工作。也可以理解为是一个独立又和现实世界交融的平行数字世界。。。

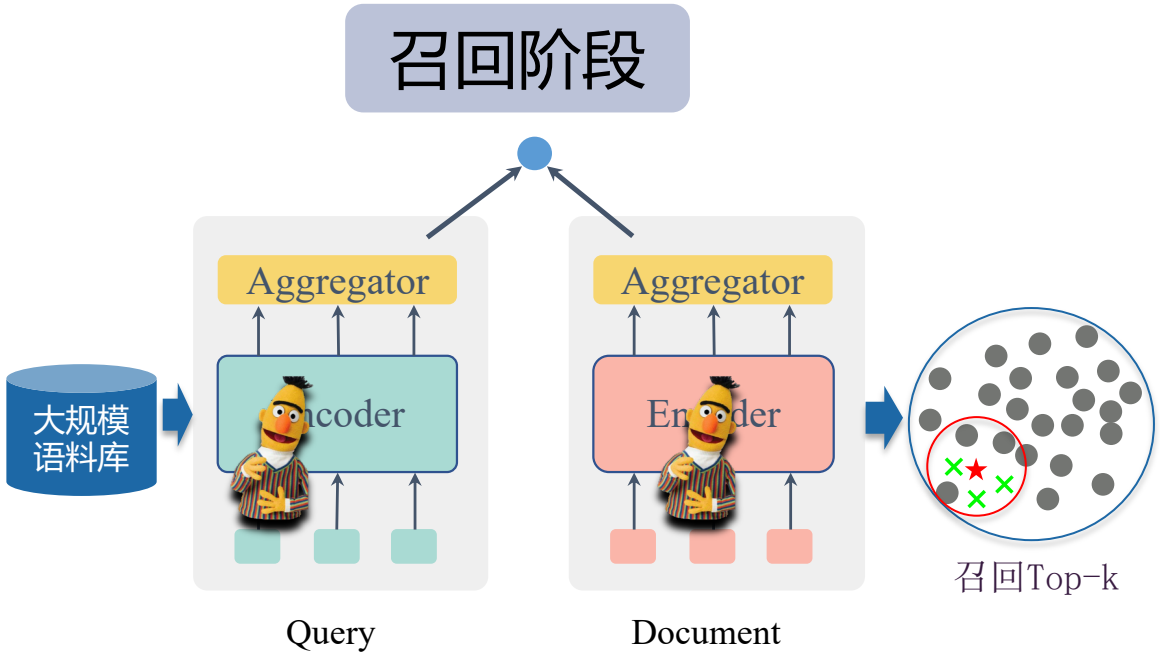
相关性函数： $f(q,d) = x_1 \,?\, x_2 ?\, x_3 \ldots$

（1）如何在预训练阶段为信息检索自监督的建模相关性？

科学挑战二



基于交互的结构：效果

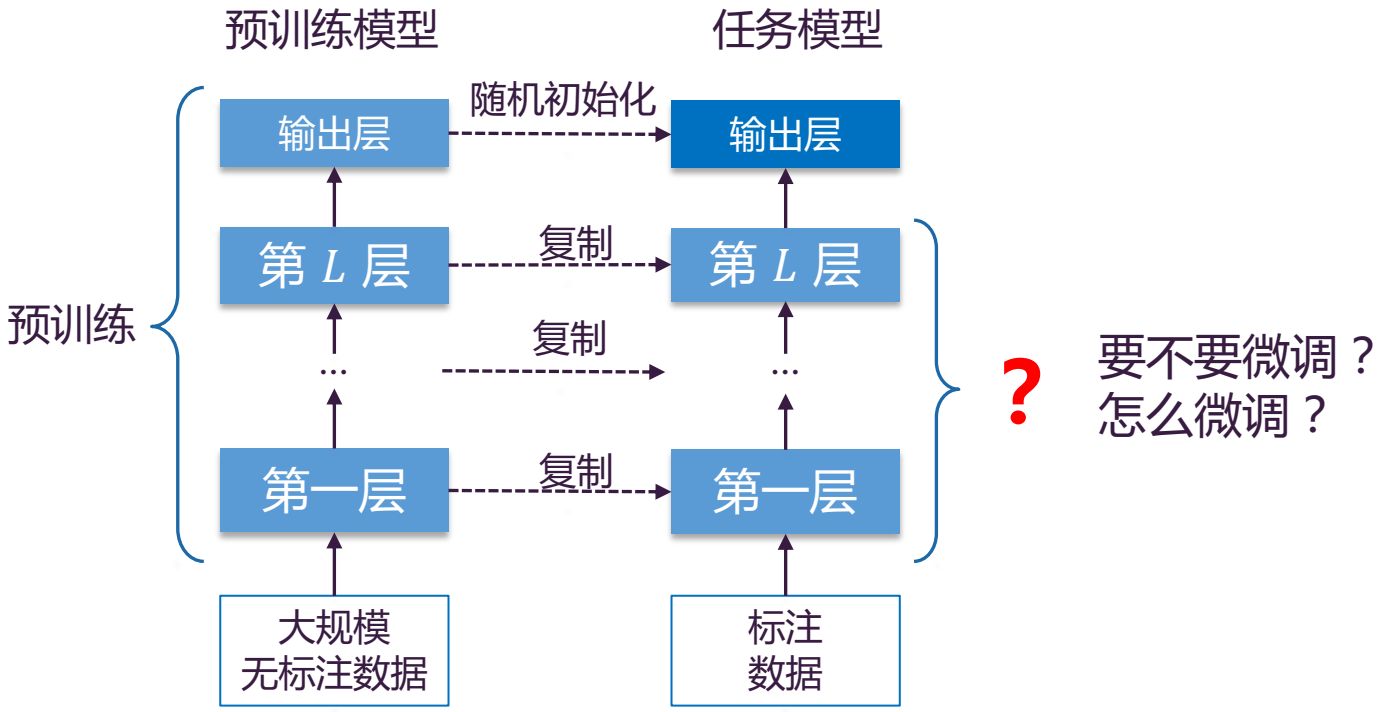


基于表达的结构：效率

（2）排序和召回的功能不同、网络架构不同，这是否会对预训练要求不同？

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, EMNLP 2019

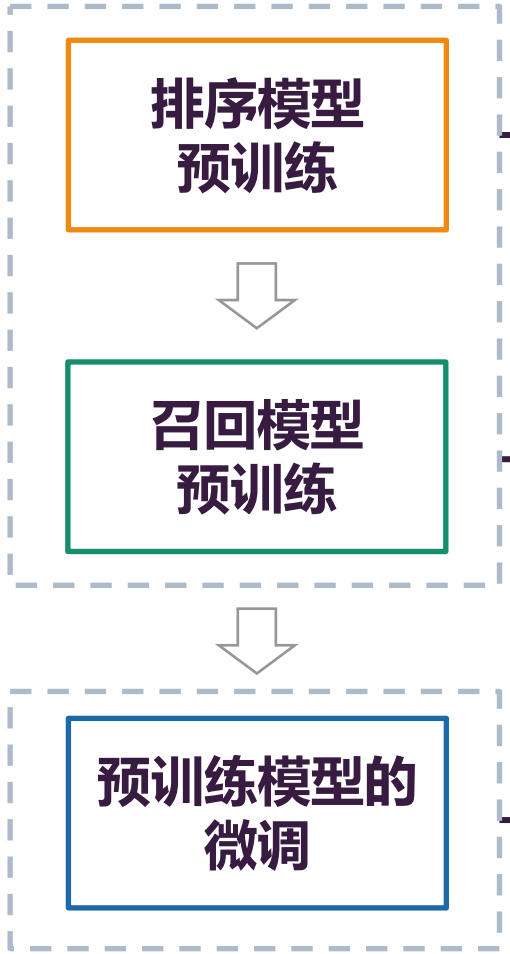
科学挑战三



主流的完全微调方法

（3）如何在保持效果的同时，兼顾微调检索模型的效率？

三个科学挑战



如何在预训练阶段为信息检索自监督的建模相关性？

- 信息检索中的核心需求相关性没有统一的明确定义

排序模型和召回模型对预训练有什么不同的要求？

- 功能的侧重点不同：召回-效率 *vs.* 排序-效果
- 网络结构不同：双塔 *vs.* 单塔

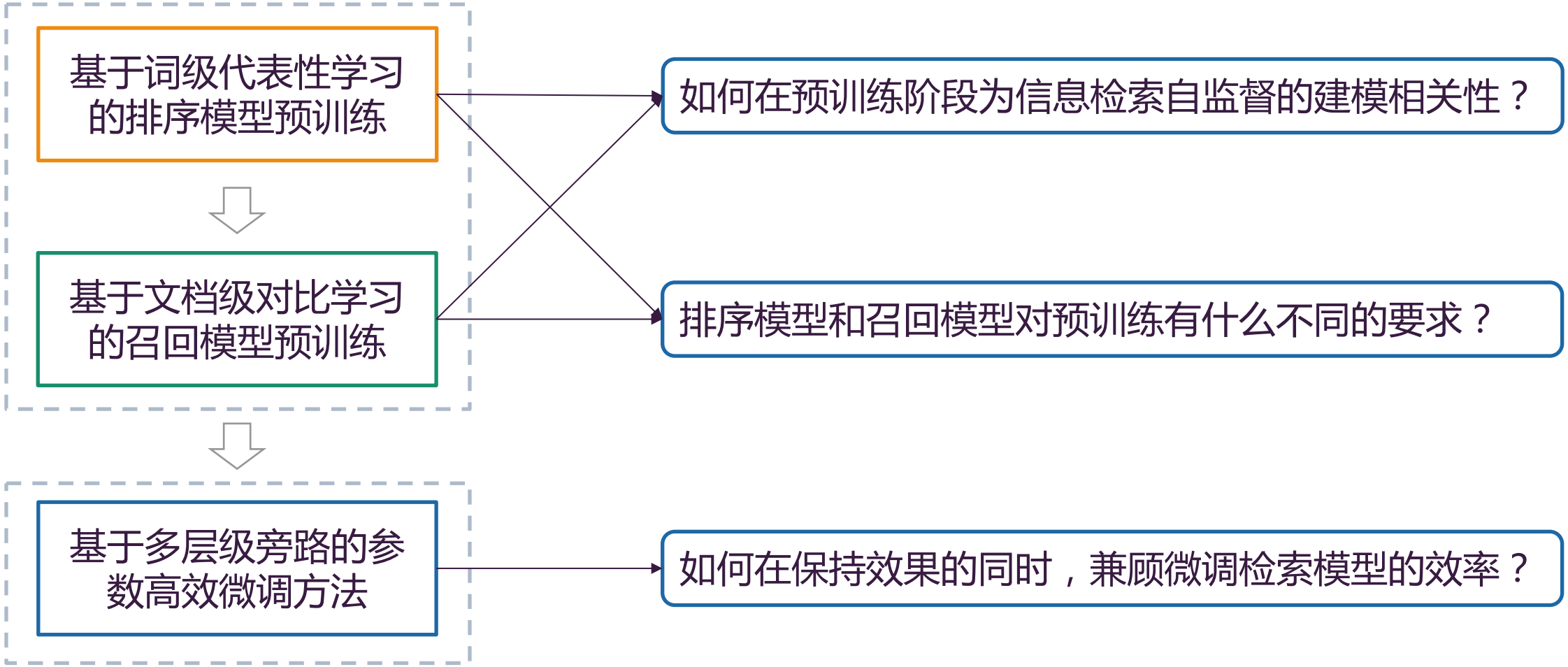
如何在保持效果的同时，兼顾微调检索模型的效率？

- 计算资源有限 *vs.* 多个大模型的总体参数量大
- 模型换代快 *vs.* 同时部署多个大模型操作复杂、成本高

三个环节

三个科学挑战

研究工作

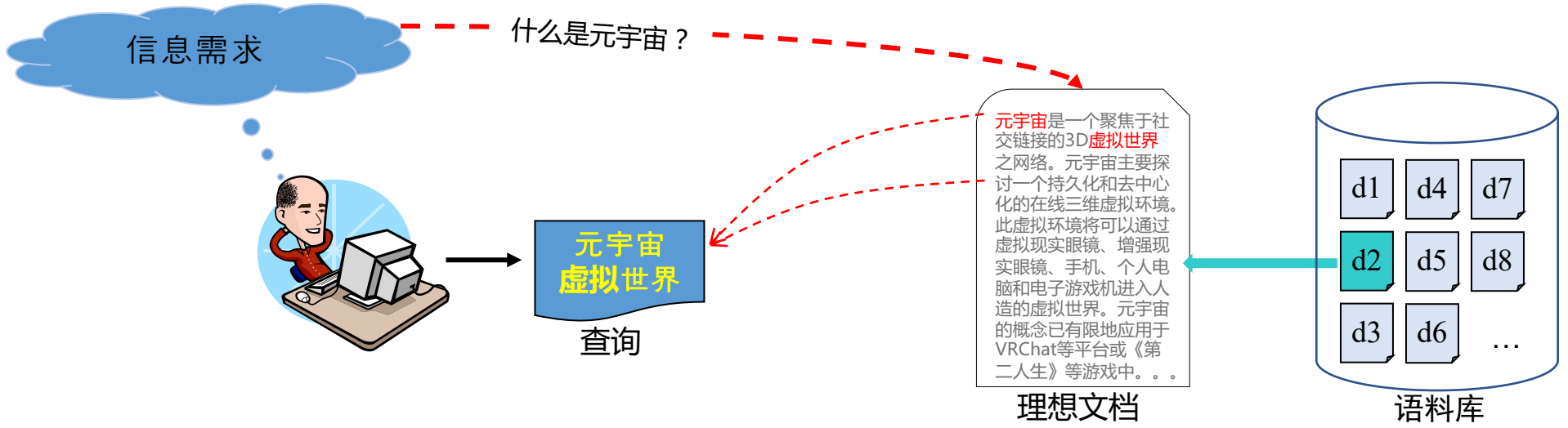

基于词级代表性学习的排序模型预训练
基于文档级对比学习的召回模型预训练
基于多层级旁路的参数高效微调方法
如何在预训练阶段为信息检索自监督的建模相关性？
排序模型和召回模型对预训练有什么不同的要求？
如何在保持效果的同时，兼顾微调检索模型的效率？
研究工作
科学挑战

主要研究内容与成果

- 基于词级代表性学习的排序模型预训练
  - ① PROP: 使用代表词预测任务预训练模型 (WSDM'2021)
  - ② B-PROP：自举式预训练代表词预测任务模型(SIGIR'2021)
- 基于文档级对比学习的召回模型预训练
  - ① 基于文档词分布的对比学习方法 (CIKM'2021)
  - ② 基于文档片段的组级别对比学习方法 (SIGIR'2022)
- 基于多层级旁路的参数高效微调方法(CIKM'2022)
  - ① 稳训练的参数高效方法

信息检索中一个“古老”的假设

- 经典统计语言模型的假设

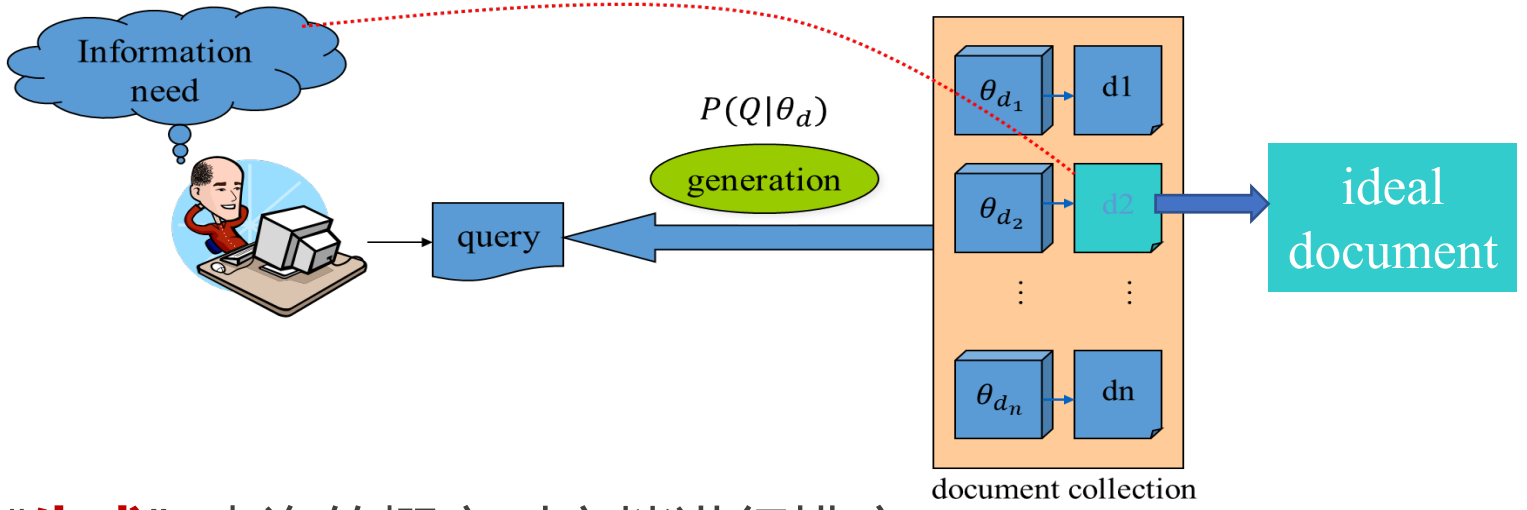


- 用户对满足其信息需求的“**理想**”**文档**中可能出现的术语有合理的想象
- 查询是从“理想”文档**生成的一段代表性文本**

A language modeling approach to information retrieval, SIGIR 1998

基于此假设的查询似然模型

- 查询是从“理想”文档**生成的一段代表性文本**



- 根据文档“**生成**”查询的概率对文档进行排序

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

文档语言模型

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

$$P(D|Q) \propto P(Q|\theta_D)P(D) \propto P(Q|\theta_D)$$

Document language model

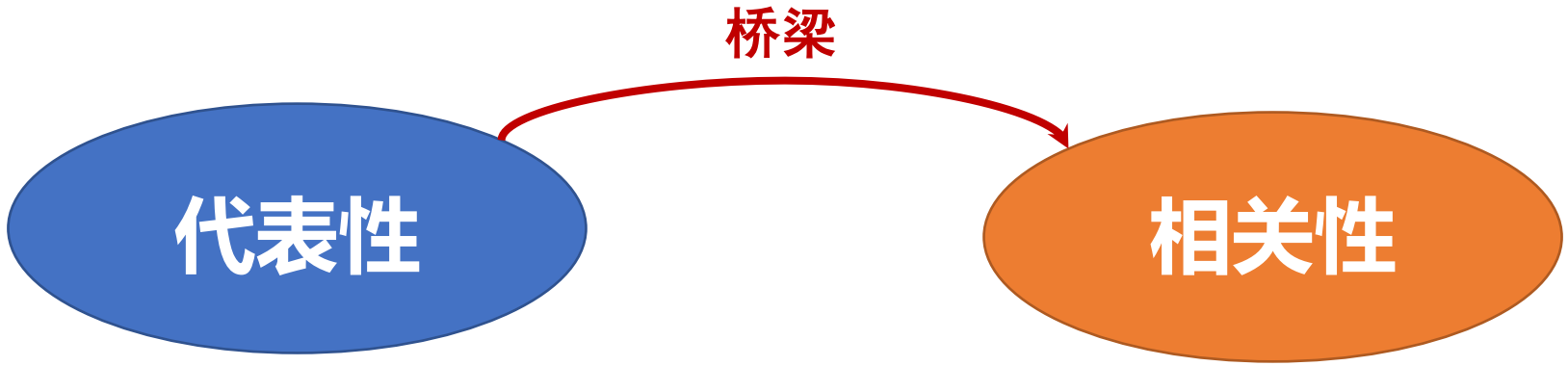
Query generation probability　Uniform distribution

多项式一元语言模型

A language modeling approach to information retrieval, SIGIR 1998

“古老”假设的启发

- 查询是从“理想”文档**生成的一段代表性文本**



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

连贯性 VS. 代表性

自然语言预训练目标

语义连贯性

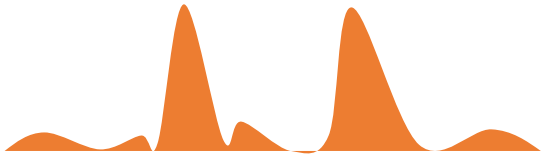
..... 狭义的信息检索仅指信息查询 .....

缓和的、平滑的

VS

信息检索预训练目标

语义代表性



..... 狭义的 信息检索仅指信息查询 .....

尖锐的、陡峭的

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

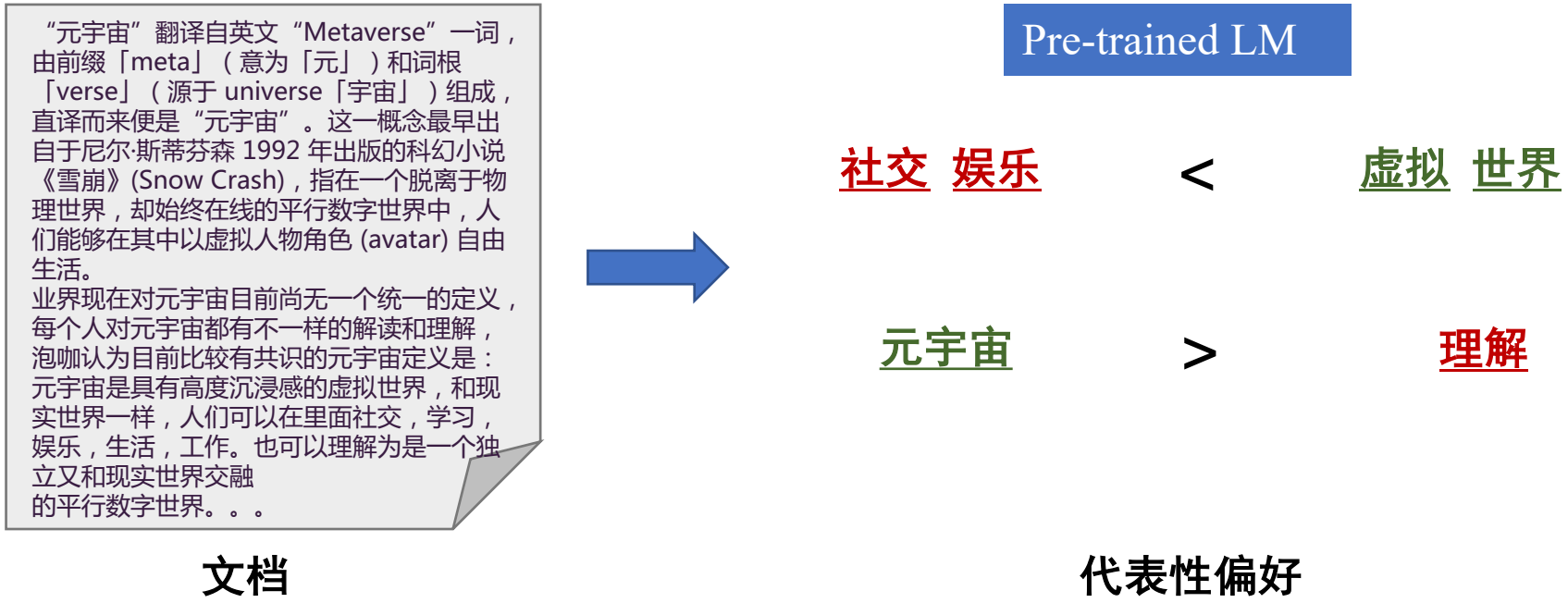
面向排序阶段的预训练假设

如果语言模型能够更好地识别文档中的代表词，它将更好地捕捉查询和文档之间的相关性匹配关系

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

面向排序阶段的代表词预测任务

- **代表词预测任务（ Representative words prediction，ROP）**：
  - 给定某个文档，预训练 Transformer 模型能够预测词集之间关于代表性的成对偏好



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

面向排序阶段的代表词预测任务

- **代表词预测任务 (ROP)** ：成对偏好预测任务

① **成对采样**: 根据**文档语言模型**成对的采样词集

- 长度选择: $P(x) = \frac{\lambda^x e^{-\lambda}}{x}, x = 1, 2, 3, \ldots$
- 词集采样: $S_1 = (w_1, w_2, \ldots, w_x),\ w_i \sim P(w_i|\theta_D)$

面向排序阶段的代表词预测任务

- **代表词预测 (ROP) 任务**：成对偏好预测任务
  - ① 成对采样: 根据文档语言模型成对的采样词集
  - ② **偏好学习**：具有更高似然概率的词集被认为更能“代表”文档

$$\text{Likelihood: } P(S_1|\theta_D) = \prod_{w \in V} P(w|\theta_D)^{c(w,Q)}$$

“元宇宙”翻译自英文“Metaverse”一词，由前缀「meta」（意为「元」）和词根「verse」（源于 universe「宇宙」）组成，直译而来便是“元宇宙”。这一概念最早出自于尼尔·斯蒂芬森 1992 年出版的科幻小说《雪崩》(Snow Crash)，指在一个脱离于物理世界，却始终在线的平行数字世界中，人们能够在其中以虚拟人物角色 (avatar) 自由生活。
业界现在对元宇宙目前尚无一个统一的定义，每个人对元宇宙都有不一样的解读和理解，泡咖认为目前比较有共识的元宇宙定义是：元宇宙是具有高度沉浸感的虚拟世界，和现实世界一样，人们可以在里面社交，学习，娱乐，生活，工作。也可以理解为是一个独立又和现实世界交融
的平行数字世界。。。

文档

成对采样

虚拟 世界 > 社交 娱乐

平行 数字 世界 > 不一样 解读 理解

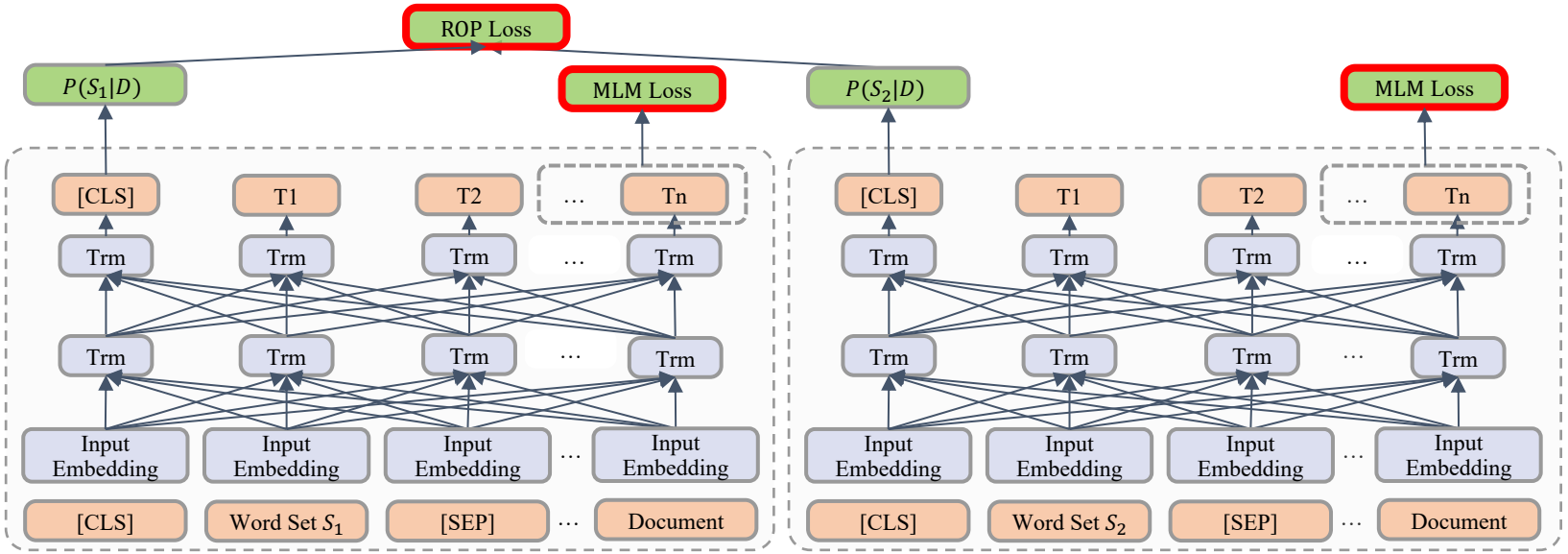
元宇宙 > 英文

……

似然比较

面向排序阶段，建模文档内语义代表性的交互预训练模型PROP

- 使用以下两个目标对 Transformer 进行预训练
  - **代表词预测目标**
    - 成对的采样词集 (S_1, S_2)，假设s_1的代表性概率高于s_2
    - 偏好学习: $\mathcal{L}_{ROP} = \max(0, 1 - P(S_1|D) + P(S_2|D))$
  - **掩码语言模型目标**
    - 基于上下文，预测随机掩盖的一些词
    - 交叉熵: $\mathcal{L}_{MLM} = -\sum_{\hat{x} \in X} \log p(\hat{x}|X_{\setminus \hat{x}})$



实验结果

Table 2: Comparisons between PROP and the baselines. $*$, † and ‡ indicate statistically significance with $p-value \leq 0.05$ over BM25, BERT and Transformer$_{ICT}$, respectively.

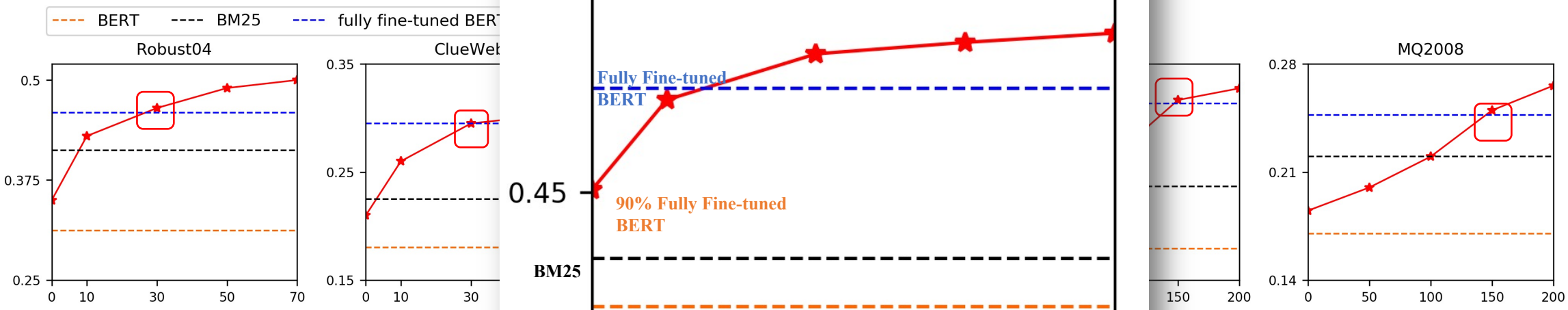
Model	Robust04		ClueWeb09-B		Gov2		MQ2007		MQ2008	
	nDCG@20	P@20	nDCG@20	P@20	nDCG@20	P@20	nDCG@10	P@10	nDCG@10	P@10
QL	0.413	0.367	0.225	0.326	0.409	0.510	0.423	0.371	0.223	0.241
BM25	0.412	0.363	0.230	0.334	0.421	0.523	0.414	0.366	0.220	0.245
Previous SOTA	**0.538**▲	**0.467**▲	0.296	-	0.422	0.524	0.490	0.418	0.244	0.255
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$_{Wikipedia}$	**0.502***†‡	**0.421***†‡	0.316*†‡	0.384*†‡	0.519*†‡	0.593*†‡	**0.523***†‡	**0.432***†‡	0.262*†‡	0.267*†‡
PROP$_{MSMARCO}$	0.484*†‡	0.408*†‡	**0.329***†‡	**0.391***†‡	**0.525***†‡	**0.594***†‡	0.522*†‡	0.430*†‡	**0.266***†‡	**0.269***†‡

▲ bert+neuIR ensemble

1. PROP在4/5个数据集上超过以前的SOTA
2. PROP 显著性超过BERT 和 $Transformer_{ICT}$
3. 相关语料上预训练效果更好

低资源场景

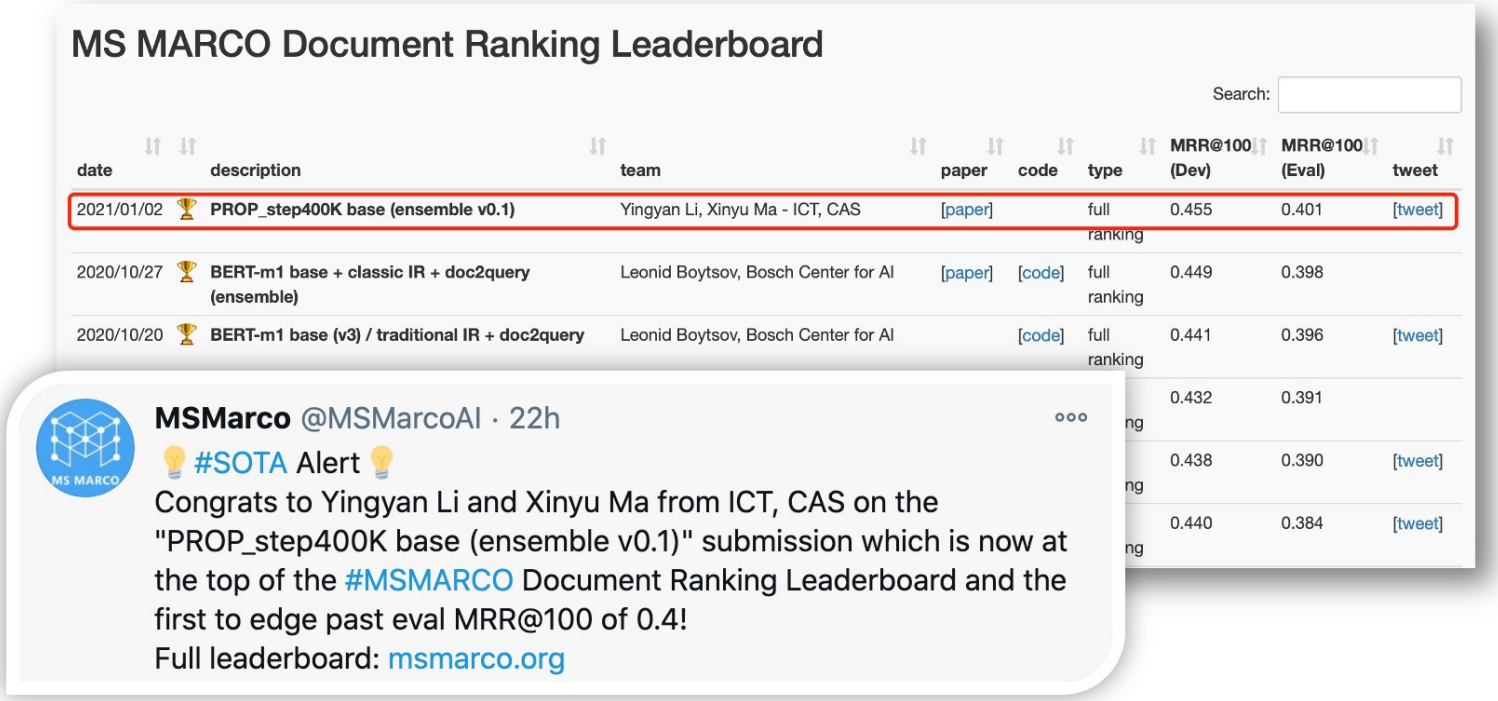
- **黑色虚线: BM25**
- **蓝色虚线: 全量数据微调的BERT**
- **橙色虚线: BERT**



1. 使用**有限监督数据微调**的 PR，可以达到相当/更好的性能
2. 在零样本的情况下, PROP 也取得了出色的效果
   - 在Gov2 数据集上，PROP 超过 BM25
   - 达到了90% 的完全微调的BERT的性能

PROP登顶国际最大排序榜单

- 超越清华-微软、斯坦福、卡内基梅隆等团队，两次登顶MS MARCO（2021/1/2，2021/4/25）
- **第一个在排序指标 MRR@100超过0.4的队伍！**



MS MARCO Document Ranking Leaderboard

Search:

date	description	team	paper	code	type	MRR@100 (Dev)	MRR@100 (Eval)	tweet
2021/01/02	**PROP_step400K base (ensemble v0.1)**	Yingyan Li, Xinyu Ma - ICT, CAS	[paper]		full ranking	0.455	0.401	[tweet]
2020/10/27	**BERT-m1 base + classic IR + doc2query (ensemble)**	Leonid Boytsov, Bosch Center for AI	[paper]	[code]	full ranking	0.449	0.398	
2020/10/20	**BERT-m1 base (v3) / traditional IR + doc2query**	Leonid Boytsov, Bosch Center for AI		[code]	full ranking	0.441	0.396	[tweet]
					ng	0.432	0.391	
					ng	0.438	0.390	[tweet]
					ng	0.440	0.384	[tweet]

MSMarco @MSMarcoAI · 22h

#SOTA Alert

Congrats to Yingyan Li and Xinyu Ma from ICT, CAS on the "PROP_step400K base (ensemble v0.1)" submission which is now at the top of the #MSMARCO Document Ranking Leaderboard and the first to edge past eval MRR@100 of 0.4!

Full leaderboard: msmarco.org

PROP中采样策略的影响

- 基于统计语言模型 成对采样 vs. 随机 成对采样

Table 5: Impact of Different Sampling Strategies. Two-tailed t-tests demonstrate the improvements of document language model-based sampling to the random sampling strategy are statistically significant († indicates p-value < 0.05).

	nDCG@20			nDCG@10	
	Robust04	ClueWeb09-B	Gov2	MQ2007	MQ2008
Random	0.471	0.304	0.505	0.513	0.252
docLM-based	0.493†	0.317†	0.517†	0.516†	0.257†

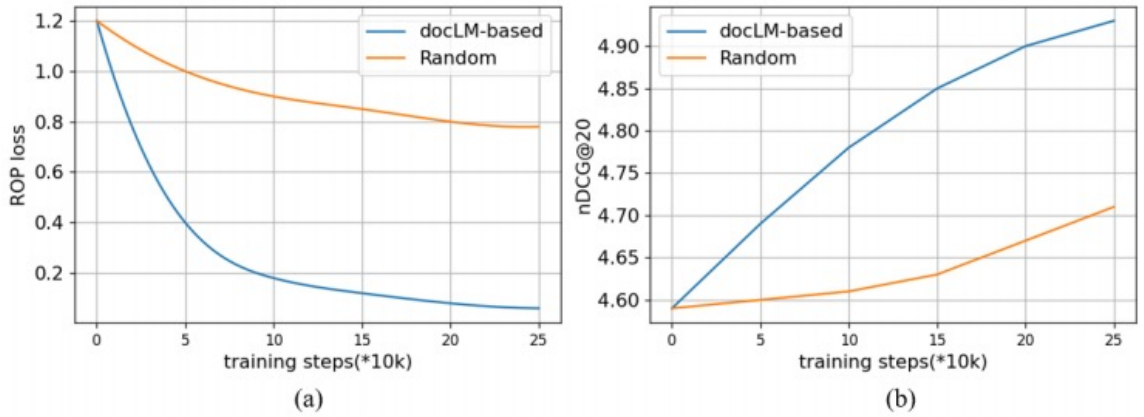


Figure 1: (a) ROP learning curve on Wikipedia over the pre-training steps. (b) The test performance curve on Robust04 in terms of nDCG@20 over pre-training steps.

基于 统计语言模型 的采样会取得**更好的性能**，伴随着**更快的收敛速度**

PROP 的成功**很大程度上依赖于采样的“代表词” 的质量**

PROP 中基于 统计语言模型 采样的问题

- PROP 中的代表词采样基于 统计语言模型 θ_D
  - θ_D 是多项式一元语言模型+狄利克雷平滑
  - **词独立性假设:** $\theta_D = \prod_{i=1}^{|D|} P(w_i) = P(w_1)P(w_2)P(w_3)$ ...
  - **忽略了词与词之间的关系，无法捕捉更高层次的语义**

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…

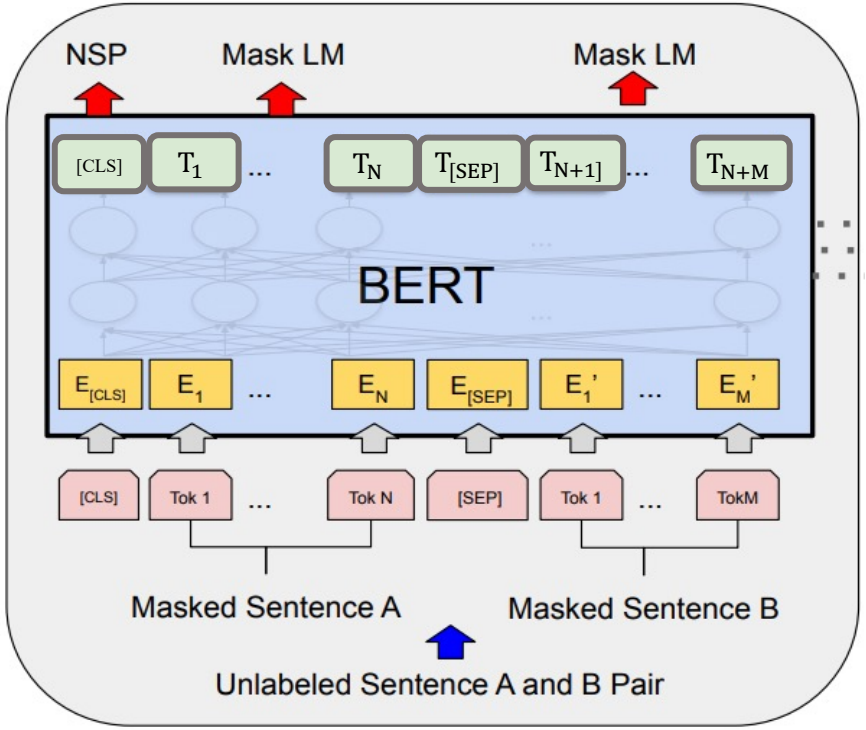
一元语言模型: fibrosis, uip, interstitial, idiopathic, pulmonary, fibrous, inspiratory, auscultation, pulmonology, amiodarone

倾向于文档中的稀有词，但是可能不具有代表性

使用BERT采样代表词

- 上下文语言模型
  - BERT 中的每个标记都从左右上下文累积信息以丰富其表示

- 语义任务（句子、句子对、文档）取得重大成功
  - 语义文本相似性：STS、MRPC
  - 文字分类：AGNews、DBpedia
  - 文档情绪：IMDB、Yelp

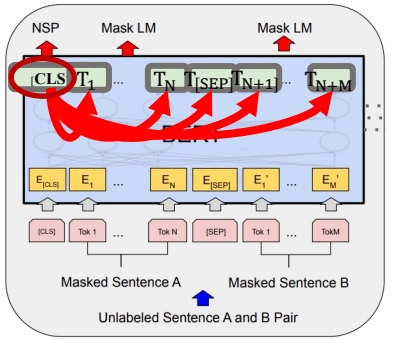


核心思想：利用 BERT 替换经典的一元语言模型来构建 ROP 任务，并向着IR 的定制目标重新训练 BERT

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

使用 BERT 构建 ROP

- 直观的解决方案：根据BERT的[CLS]-Token attention直接采样代表词
  - 特殊标记**[CLS]是整个序列表示的聚合**
  - [CLS]-token attention 表示特定 词 对整个序列的语义贡献度



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 …

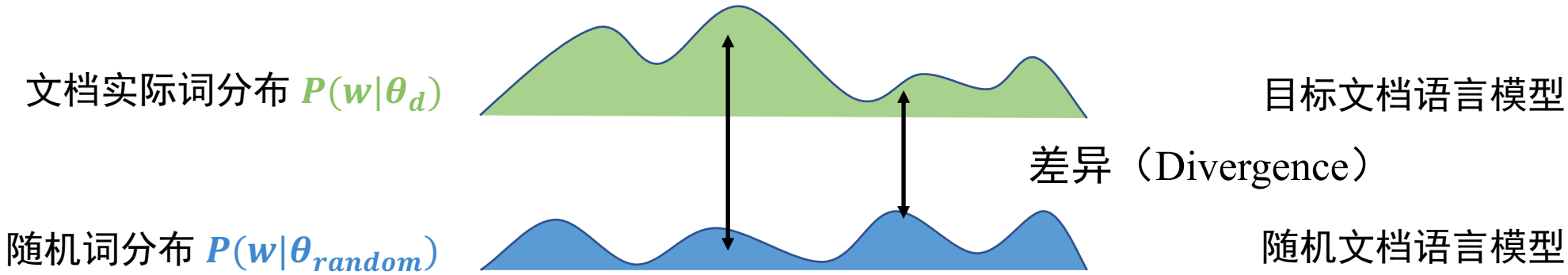
一元语言模型：	fibrosis, uip, interstitial, idiopathic, pulmonary, fibrous, inspiratory, auscultation, pulmonology, amiodarone
原始基于注意力权重的词分布：	pulmonary, fibrosis, in, interstitial, the, of, disease, can, and, lung, chest, is, cause, to,

- 原始[CLS]-Token 基于注意力的词分布 **可以生成有代表性的词，但也有对常用词有较高的权重**
  - BERT 专注于在文档中编码尽可能多的语义信息
  - 从其 原始的 attention 获得的词分布是语义分布，但不一定是代表性/信息量分布

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

随机偏差理论（Divergence from Randomness，DFR）

- 词的**信息量/代表性**可以由**随机过程产生的词分布**和**文档的真实词分布**的差异（divergence）测量。(Amati and Rijsbergen, 2002)



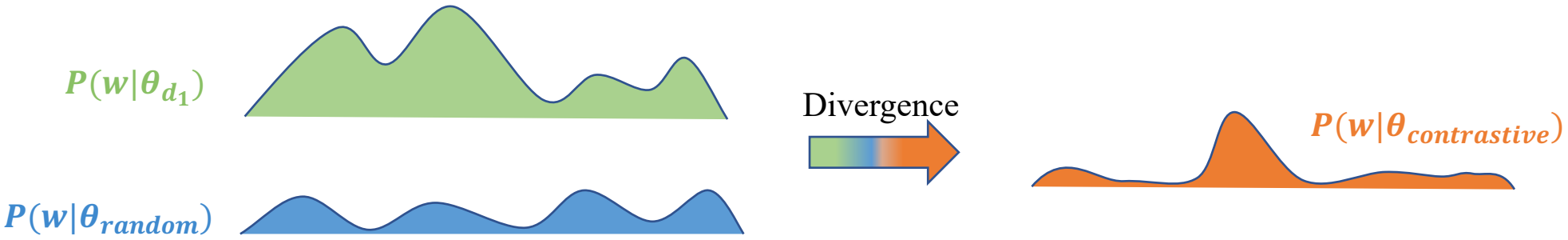
SIGIR'2021,B-PROP: Bootstrapped Pre-training with Representative Words Prediction for Ad-hoc Retrieval

使用 BERT 对 ROP 进行对比式采样

- 基于**对比词分布**进行代表词采样：
  - 计算 **交叉熵 (i.e., the divergence)** ，在文档词分布 $P(w_k|\theta_d)$ 和随机词分布之间 $P(w_k|\theta_{random})$

$$\gamma_{w_k} = CE(\theta_d|\theta_{random}) = -P(w_k|\theta_d)log_2P(w_k|\theta_{random}))$$

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



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

使用 BERT 对 ROP 进行对比式采样

- **文档词分布**
  - 平均多头注意力权重，并将相同词在不同位置的权重相加

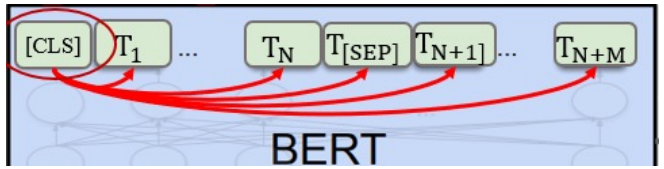
$$a^t = \frac{1}{h}\sum_{i=1}^{h} a_i^t, \qquad where\ a_i^t = softmax(\frac{Q_i^{[CLS]} * K_i^{x_t}}{\sqrt{d/h}})$$

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

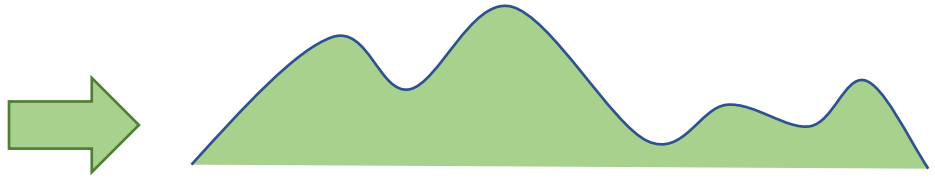
  - 对不同词的 vanilla [CLS]-Token 注意力权重进行饱和和重新规范化

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

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



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

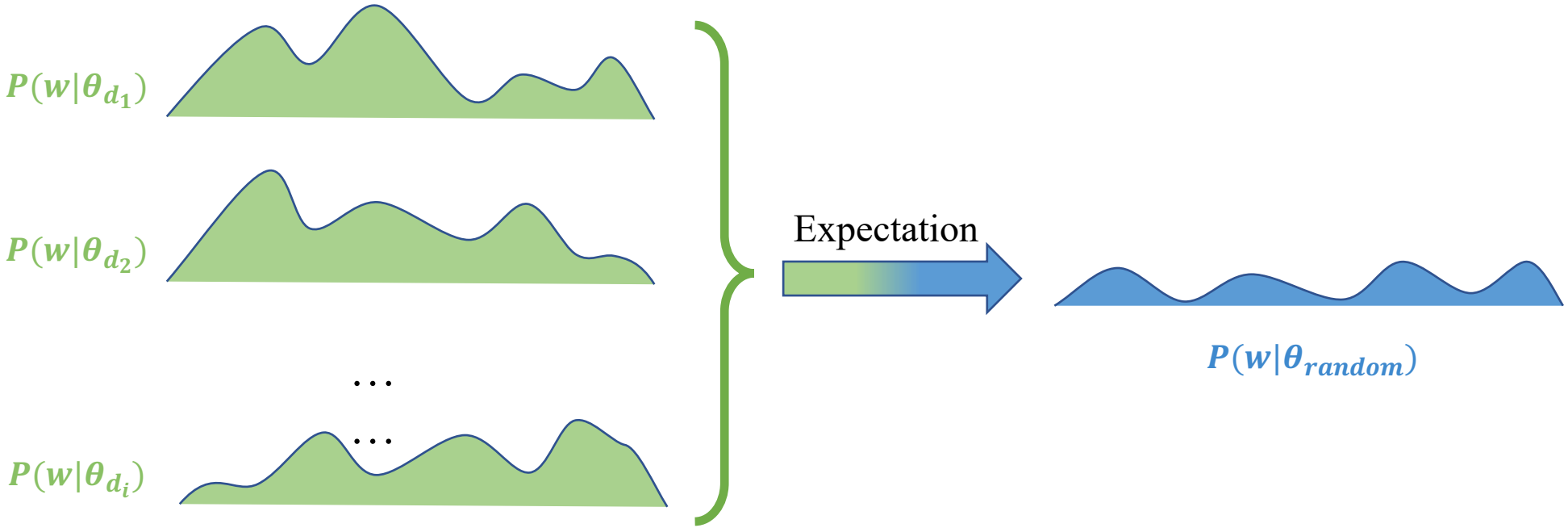


文档词分布$P(w|\theta_d)$

使用 BERT 对 ROP 进行对比式采样

- **随机词分布**:
  - 对文档集合中的所有文档词分布进行期望

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



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

使用 BERT 对 ROP 进行对比式采样

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 …

一元语言模型：	fibrosis, uip, interstitial, idiopathic, pulmonary, fibrous, inspiratory, auscultation, pulmonology, amiodarone
原始基于注意力权重的词分布：	pulmonary, fibrosis, in, interstitial, the, of, disease, can, and, lung, chest, is, cause, to,
对比词分布：	fibrosis, pulmonary, interstitial, idiopathic, lung, chest, disease, diseases, cause, patients, x-ray, scars,

- 对比词分布现在可以获得文档的代表词
  - 通过使用对比方法消除常用词（即停用词）的影响
- 对比词分布在代表性方面优于一元语言模型
  - （lung，chest，...）*vs.*（breath，diagnosed，...）

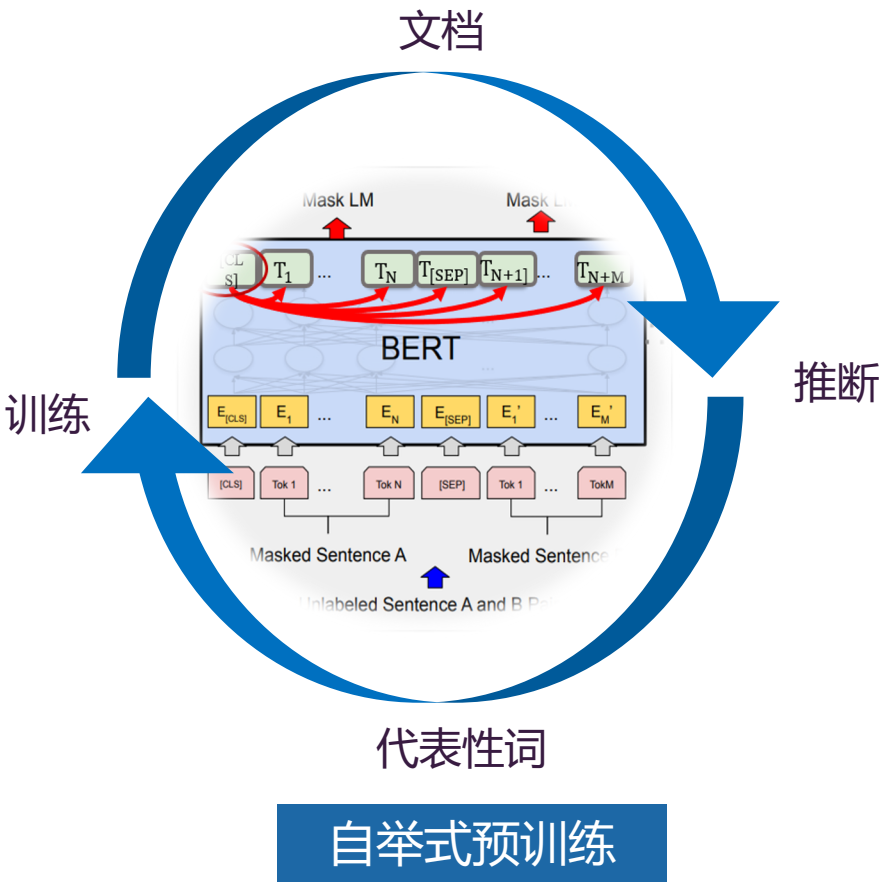
SIGIR'2021,B-PROP: Bootstrapped Pre-training with Representative Words Prediction for Ad-hoc Retrieval

B-PROP：使用代表性词预测任务的自举式预训练模型

- 使用面向IR的预训练目标**重新训练BERT**

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

- **ROP**: 成对偏序学习任务
  - Pairwise Loss: $\mathcal{L}_{ROP} = \max(0, 1 - P(S_1|D) + P(S_2|D))$
- **MLM**: 掩码语言模型
  - Cross-entropy Loss: $\mathcal{L}_{MLM} = -\sum_{\hat{x}\in X} \log p(\hat{x}|X_{\setminus\hat{x}})$



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

小数据集上的结果

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	0.415	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 在小数据集上的表现比 PROP 和其他预训练模型好很多
- 对相似领域的预训练可以带来更好的微调性能

大数据集的实验结果

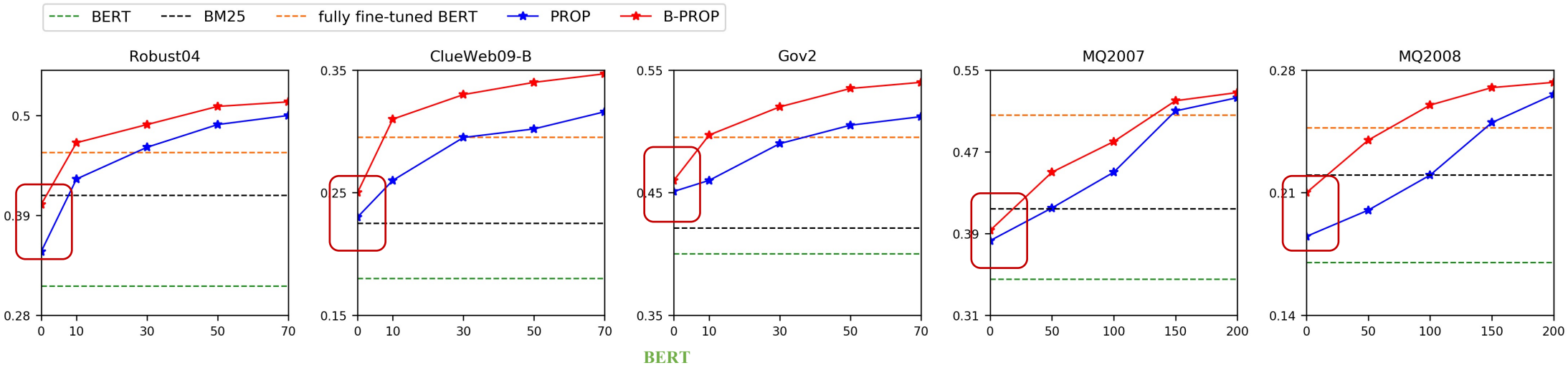
Table 4: Comparisons between B-PROP and the baselines on 2 large-scale datasets. Two-tailed t-tests demonstrate the improvements of B-PROP to the best baseline PROP are statistically significant ($*$ indicates $p \leq 0.05$).

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 在大型数据集上的表现优于 PROP 和其他预训练模型，但改进更小

零资源和低资源场景

- 黑色虚线: BM25
- 绿色虚线: BERT
- 橙色虚线: Fully fine-tuned BERT
- 蓝色实线: PROP
- 红色实线: B-PROP



- B-PROP 在使用相同数量的有限监督数据下，在所有数据集上显著优于 PROP
- 在有限的监督数据上微调的 B-PROP 可以实现与在完整监督数据上微调的 BERT 相当/更好的性能
- 在零资源设置下，B-PROP 在所有数据集上的表现都优于 PROP

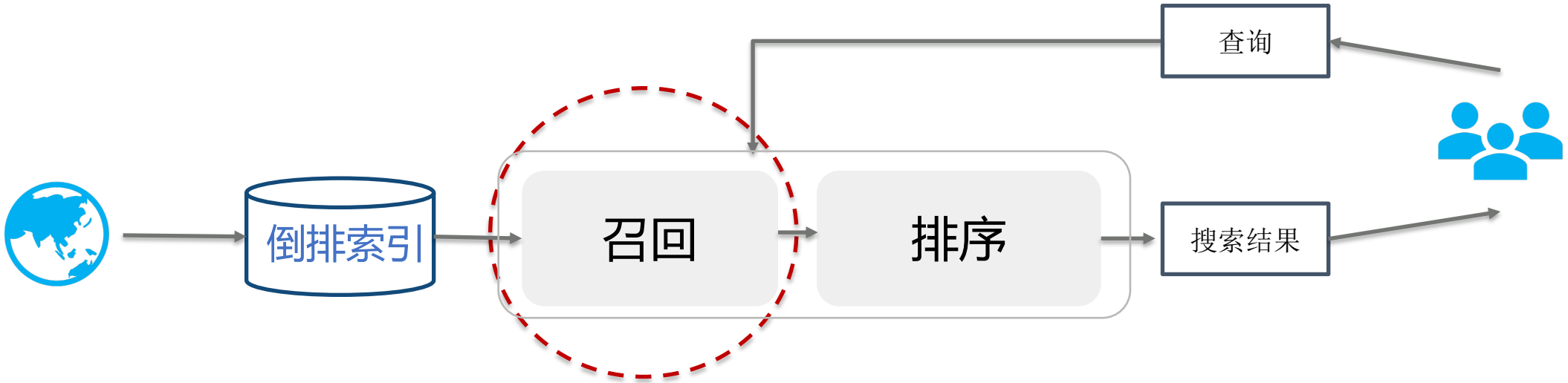
面向排序阶段的预训练

- 总结
  - ① 受启发于经典统计语言模型，区别于NLP建模语义连贯性，为IR提出了建模语义代表性
  - ② 设计了了代表词预测任务
    - 通过使用**统计文档语言模型**对代表性打分，首次提出了面向检索排序阶段的预训练模型PROP
    - 通过使用**BERT中的CLS-Token**权重，进一步改进并优化PROP，提出了B-PROP
  - ③ PROP曾两次登顶国际最大规模排序榜单 MS MARCO（2021/1/2，2021/4/25）

主要研究内容与成果

- 基于词级代表性学习的排序模型预训练
  - ① PROP: 使用代表词预测任务预训练模型 (WSDM'2021)
  - ② B-PROP：自举式预训练代表词预测任务模型(SIGIR'2021)
- 基于文档级对比学习的召回模型预训练
  - ① 基于文档词分布的对比学习方法 (CIKM'2021)
  - ② 基于文档片段的组级别对比学习方法 (SIGIR'2022)
- 基于多层级旁路的参数高效微调方法(CIKM'2022)
  - ① 稳训练的参数高效方法

ROP 预训练任务在召回阶段的表现如何？



ROP在召回阶段的效果

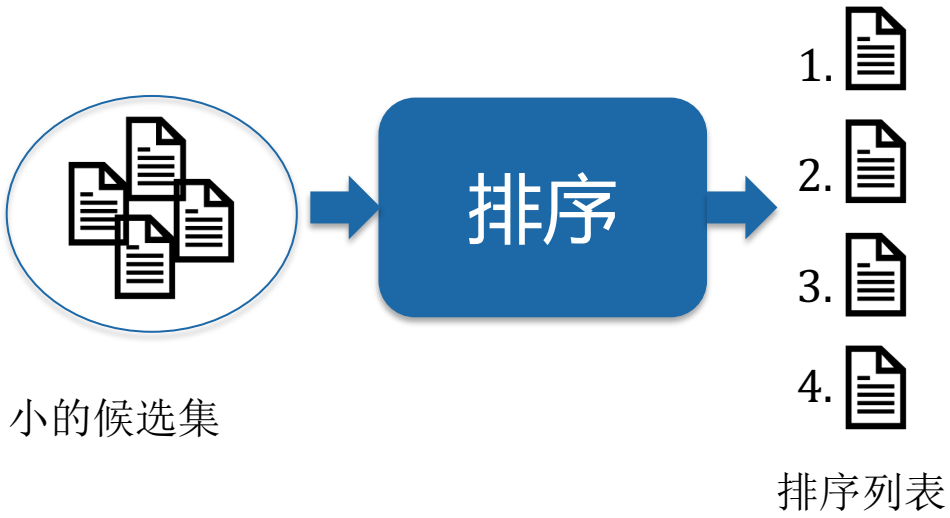
	MARCO Passage	TREC 2019 Passage	MARCO Doc	TREC 2019 Doc
	Recall@1000		Recall@100	
BM25	0.857	0.745	0.808	0.395
DeepCT	0.907	-	-	-
BERT	0.941	0.704	0.869	0.266
ICT	0.938	0.705	**0.873**	**0.273**
PROP	**0.948**	**0.709**	0.871	0.269

ROP 任务的性能略好于 BERT，但**提升非常有限**。

Pre-train a Discriminative Text Encoder for Dense Retrieval via Contrastive Span Prediction, SIGIR 2022

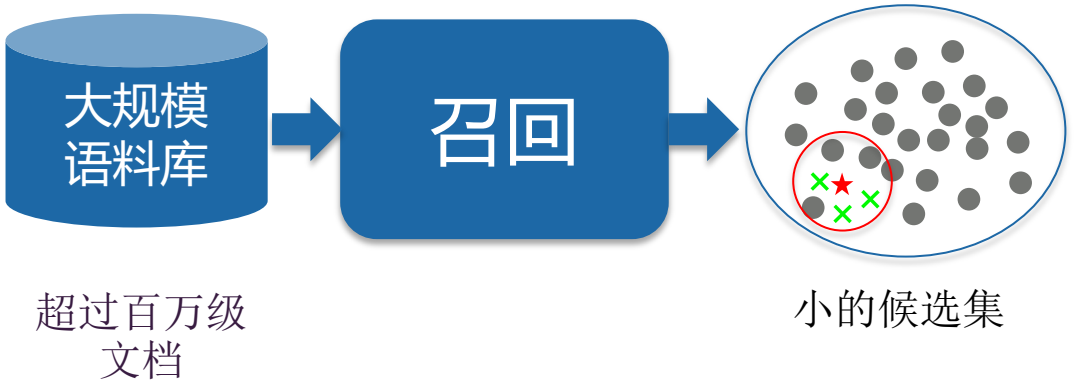
排序和召回的不同需求

排序的需求(ROP 任务):



VS.

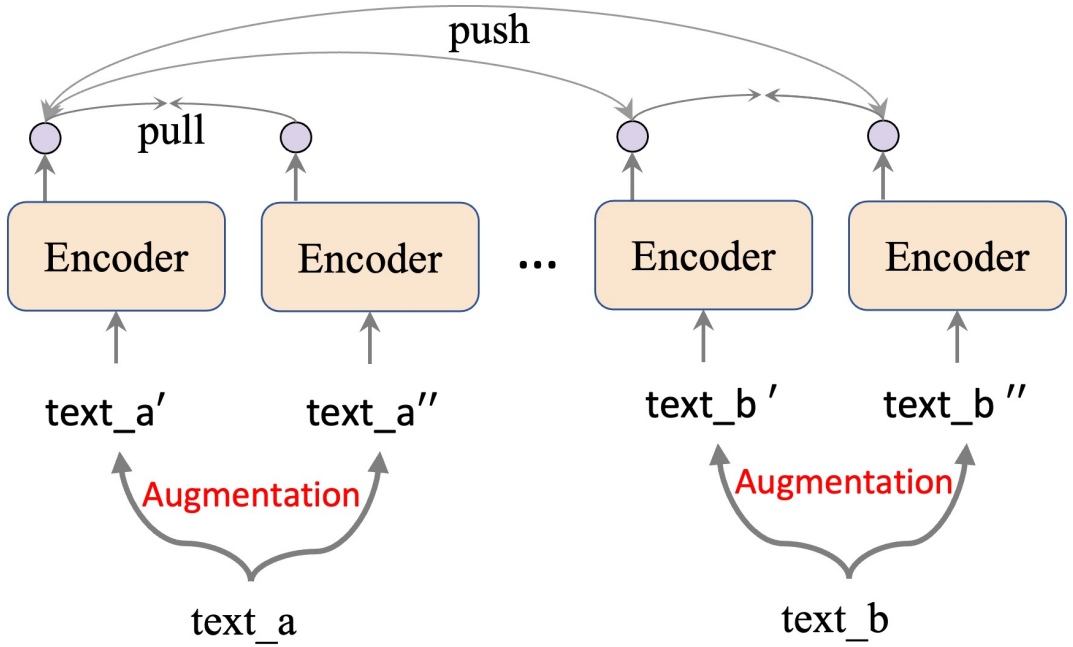
召回的需求:



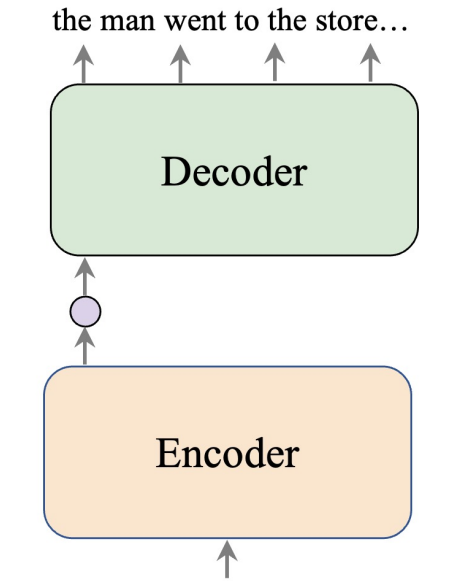
排序:在查询和一小组候选文档之间进行细粒度的相关性匹配

召回:以粗粒度的方式从超过百万个文档中区分出一小组候选文档

两大表达学习方法：对比学习和自编码器



(a) Contrastive Learning



(b) Auto-encoder

关键：**数据增广技术**

已有工作问题： 针对长文档表达学习工作较少

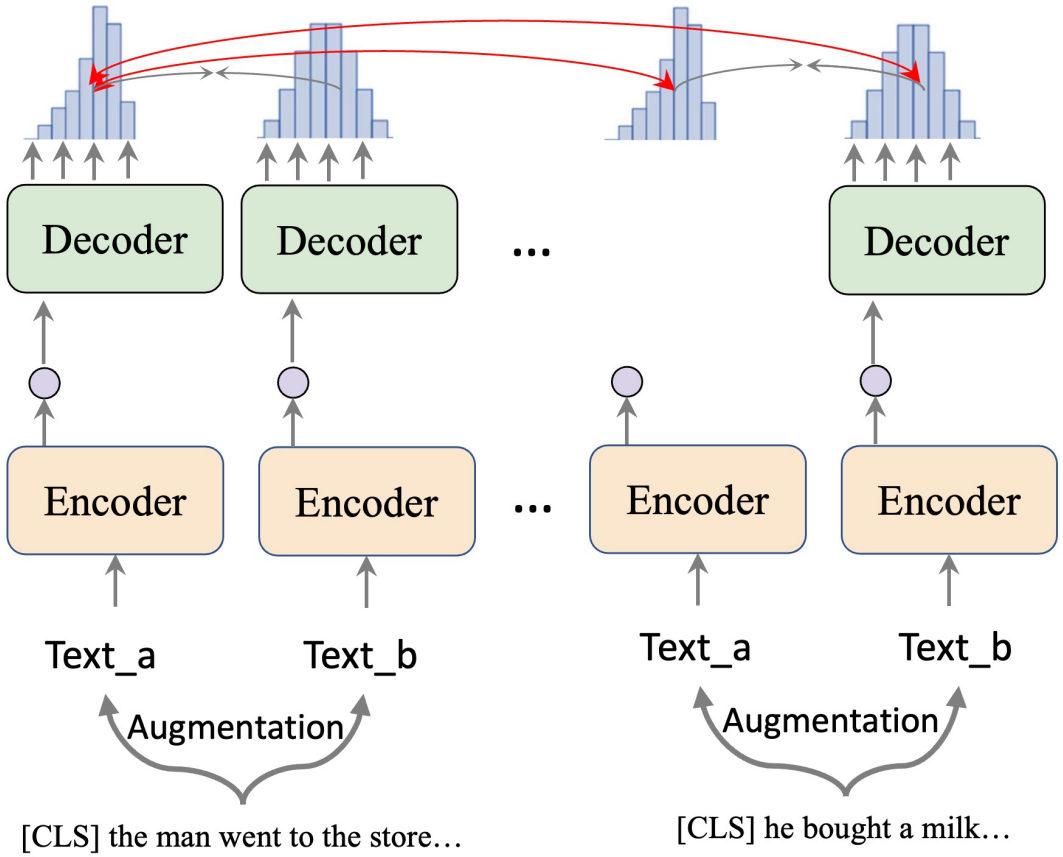
SentBERT、SimCSE、ICT、coCondenser、…

关键：**重建方式**

已有工作问题： （1）绕过效应；（2）表达的区别力

Optimus、SEED、TSALDE、Condenser、…

判别式自编码器的对比预训练方法



基于文档词分布的对比预训练方法

① 随机mask增广两个长文本

② 使用MLP解码器替代自回归解码器

③ 使用KL散度计算词分布之间的距离

A Contrastive Pre-training Approach to Discriminative Autoencoder for Dense Retrieval, CIKM 2022

理论分析

- 当使用KL散度进行对比学习中的 *pull and push*

$$-JS(P,Q) = -\sum_{x\in|V|} p(x)log(p(x)) - \sum_{x\in|V|} q(x)log(q(x)) + \sum_{x\in|V|} (p(x)+q(x))log(p(x)+q(x)).$$

- 推开常见词：

$$-JS(P,Q)_{x\in S} = -\sum_{x\in S} 2p(x)log(p(x)) + \sum_{x\in S} 2p(x)log(2p(x)) = \sum_{x\in S} 2p(x)log2.$$

大于0，常见词的$p(x)$将会被减小

- 拉近代表性词：

$$-JS(P,Q) = -\sum_{x\in \complement VS} p(x)log(p(x))$$

大于$\frac{1}{e}$，$x\log x$单调递增，代表词$p(x)$将会被增大

(1) 基于词分布的对比预训练将在解码时 **抑制常见词的概率。**
(2) 基于词分布的对比预训练将在解码时 **突出代表性/信息性词的概率**

A Contrastive Pre-training Approach to Discriminative Autoencoder for Dense Retrieval, CIKM 2022

实验结果

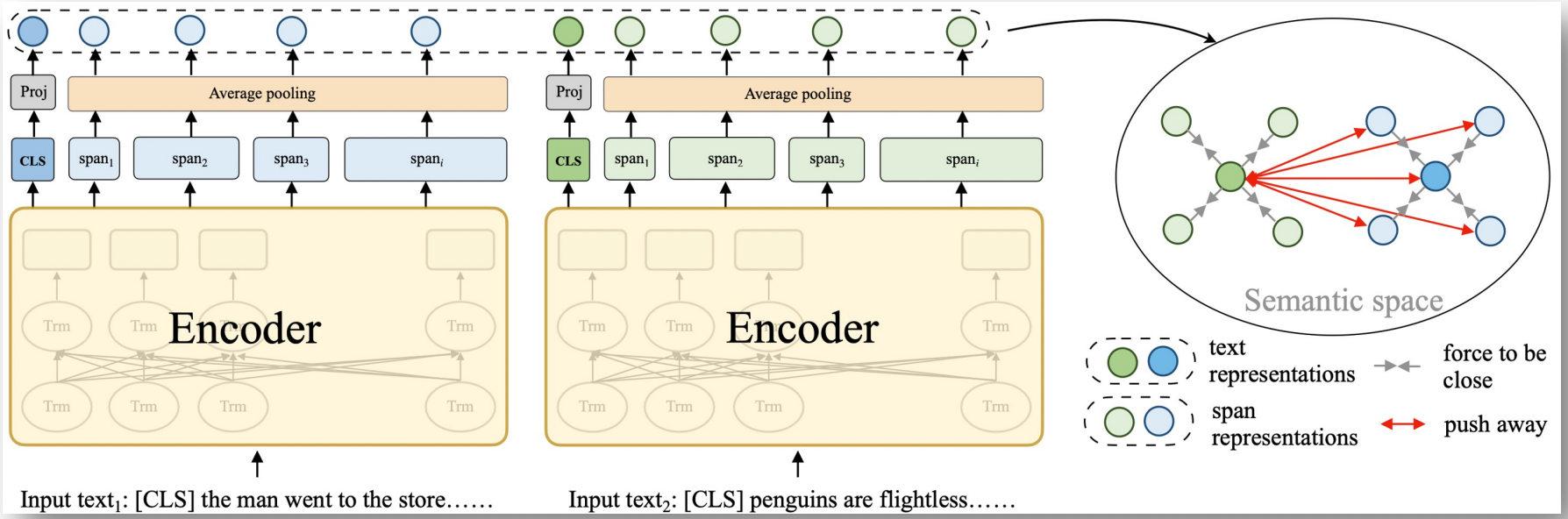
Model	MARCO Dev Passage		TREC2019 Passage		MARCO Dev Doc		TREC2019 Doc	
	MRR@10	Recall@1000	NDCG@10	Recall@1000	MRR@100	Recall@100	NDCG@10	Recall@100
BM25	0.187	0.857	0.501	0.745	0.277	0.808	0.519	**0.395**
Best TREC Trad[5]	-	-	0.554	-	-	-	0.549	-
BERT	0.335	0.957	0.661	0.769	0.389	0.877	0.594	0.301
SimCSE	0.335	0.955	0.662	0.766	0.391	0.879	0.598	0.302
ICT	0.339	0.955	0.670	0.775	0.396	0.882	0.605	0.303
PROP	0.337	0.951	0.673	0.771	0.394	0.884	0.596	0.298
SEED	0.342*	0.963	0.679*	0.782*†	0.396	0.902*	0.605*	0.307
CPDAE_R	0.350*†	0.965*†	0.686*†	0.789*†	0.402*	**0.909***†	0.609*	0.311*
CPDAE	**0.355***†‡	**0.968***†	**0.696***†‡	**0.799***†‡	**0.408***†‡	0.907*†	**0.615***†‡	0.315*†

- 在召回阶段，相比于已有的对比学习方法和自编码器的方法有所提升
- 词分布是基于词带模型的方法，词与词之间的语义关系难以建模

A Contrastive Pre-training Approach to Discriminative Autoencoder for Dense Retrieval, CIKM 2022

COSTA：对比片段预测任务

- 核心思想：基于组级别的对比损失，强制编码器生成接近自己文档内部随机跨度的文本表示，同时远离其他表示。

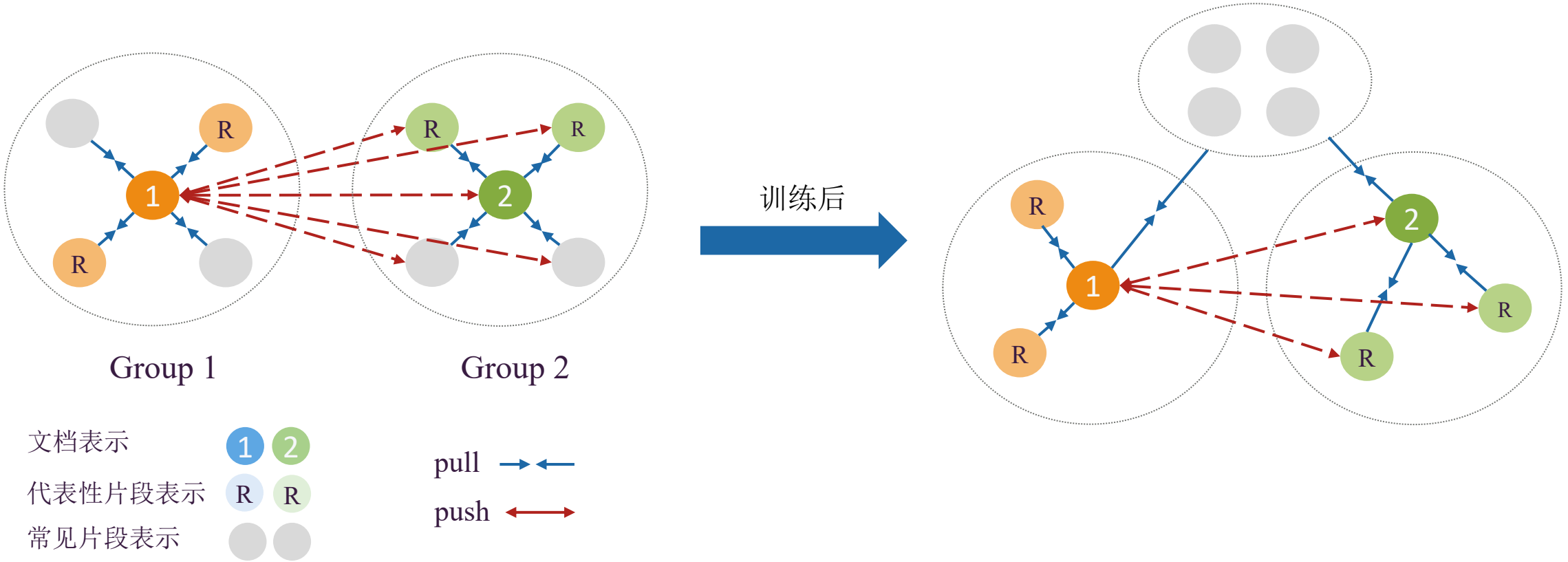


① 只使用编码器，抛弃解码器
② 使用文档表示“重构”出内部片段的表示
③ 组级别的对比损失

- ***Pull*** 给定文本在语义空间的表示与它自己内部采样的多个随机片段表示
- ***Push*** 给定文本的表示远离所有其他文档及其片段的表示

Pre-train a Discriminative Text Encoder for Dense Retrieval via Contrastive Span Prediction, SIGIR 2022

COSTA：对比片段预测任务



- ***Pull*** 给定文本在语义空间的表示与它自己内部采样的多个随机片段表示
- ***Push*** 给定文本的表示远离所有其他文档及其片段的表示

Pre-train a Discriminative Text Encoder for Dense Retrieval via Contrastive Span Prediction, SIGIR 2022

预训练目标

- MLM任务

$$\mathcal{L}_{MLM} = -\sum_{\hat{x}\in X} \log p(\hat{x}|X_{\setminus\hat{x}})$$

- 组级别的对比片段预测任务

$$L_{GWC} = \sum_{i=1}^{N} -\frac{1}{4T} \sum_{p\in S(i)} \frac{\exp(sim(z_i, z_p)/\tau)}{\sum_{j=1}^{N*(4T+1)} \mathbb{I}_{[i\neq j]} \exp(sim(z_i, z_p)/\tau)}$$

- 最终的损失函数:

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{GWC} + \mathcal{L}_{MLM}$$

Pre-train a Discriminative Text Encoder for Dense Retrieval via Contrastive Span Prediction, SIGIR 2022

主要实验结果

Model	MARCO Dev Passage		TREC2019 Passage	
	MRR@10	R@1000	NDCG@10	R@1000
Sparse retrieval models				
BM25	0.187	0.857	0.501	0.745
DeepCT[6]	0.243	0.905	0.551	-
Best TREC Trad[5]	-	-	0.554	-
Fine-tuning with official BM25 negatives				
BERT	0.316	0.941	0.616	0.704
ICT	0.324	0.938	0.618	0.705
PROP	0.320	0.948	0.586	0.709
B-PROP	0.321	0.945	0.603	0.705
SEED[29]	0.329	0.953	-	-
SEED(ours)	0.331*	0.950*	0.625*	0.733*†
COSTA	**0.342***†‡	**0.959***†	**0.635***†‡	**0.773***†‡
Fine-tuning with static hard negatives				
BERT	0.335	0.957	0.661	0.769
ICT	0.339	0.955	0.670	0.775
PROP	0.337	0.951	0.673	0.771
B-PROP	0.339	0.952	0.672	0.774
SEED	0.342*	0.963	0.679*	0.782*†
COSTA	**0.366***†‡	**0.971***†	**0.704***†‡	**0.816***†‡

Model	MARCO Dev Doc		TREC2019 Doc	
	MRR@100	R@100	NDCG@10	R@100
Sparse retrieval models				
BM25	0.277	0.808	0.519	0.395
DeepCT[6]	0.320	-	0.544	-
Best TREC Trad[5]	-	-	0.549	-
1st iteration: Fine-tuning with static hard negatives				
BERT	0.358	0.869	0.563	0.266
ICT	0.364	0.873	0.566	0.273
PROP	0.361	0.871	0.565	0.269
B-PROP	0.365	0.871	0.567	0.268
SEED	0.372*	0.879*	0.573*	0.272
COSTA	**0.395***†‡	**0.894***†‡	**0.582***†‡	**0.278***
2nd iteration: Fine-tuning with static hard negatives				
BERT	0.389	0.877	0.594	0.301
ICT	0.396	0.882	0.605	0.303
PROP	0.394	0.884	0.596	0.298
B-PROP	0.395	0.883	0.601	0.305
SEED	0.396	0.902*	0.605*	0.307
COSTA	**0.422***†‡	**0.919***†‡	**0.626***†‡	**0.320***†‡

Use BM25 negatives

Mine hard negatives use the current model

Initialized from the passage model and mine hard negatives use the current model

Pre-train a Discriminative Text Encoder for Dense Retrieval via Contrastive Span Prediction, SIGIR 2022

COSTA的文档间代表性

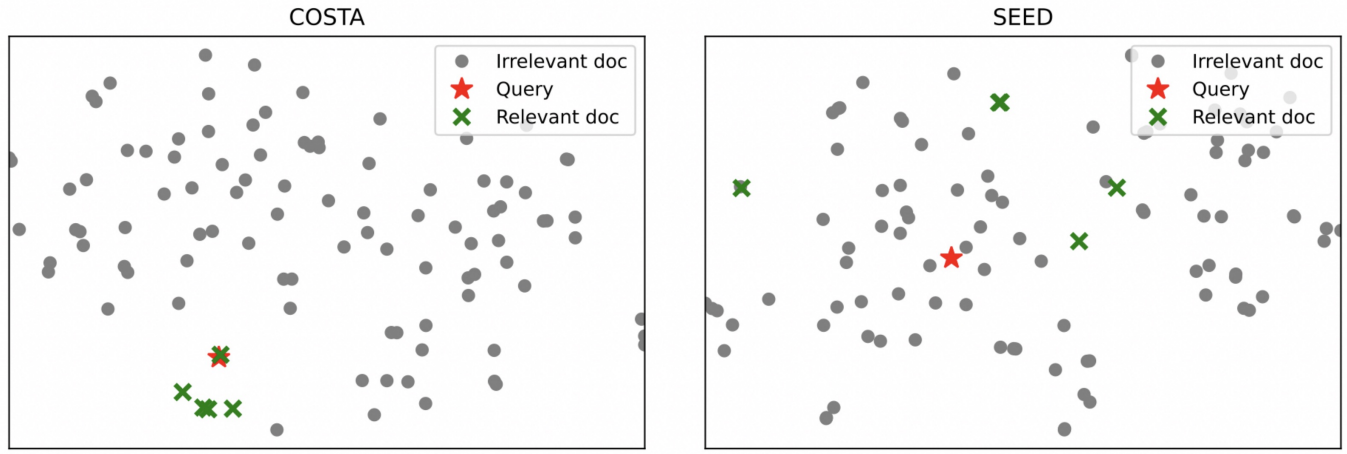


Figure 3: The t-SNE plot of query and document representations for SEED and COSTA. The QID is 47923 and is from TREC2019 Passage test set.

- COSTA 在语义空间中会将查询及其相关文档**更紧密地映射到一起**。

Pre-train a Discriminative Text Encoder for Dense Retrieval via Contrastive Span Prediction, SIGIR 2022

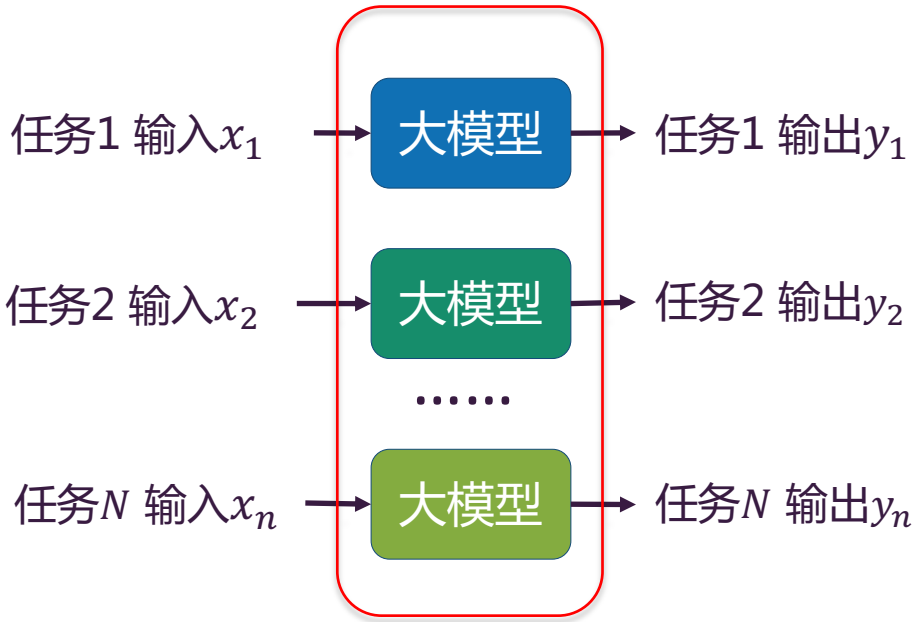
召回模型预训练

- 总结
  - ① 区别于排序阶段建模词粒度语义代表性的预训练方法，提出了面向召回阶段建模文档粒度语义代表性的对比预训练方法
  - ② 设计了两种对比预训练任务
    - 基于自编码器的文档词分布的对比方法
    - 仅使用编码器的文档片段的组级别对比方法
  - ③ 显著提升了预训练模型在召回排序上的性能，并验证了其判别能力

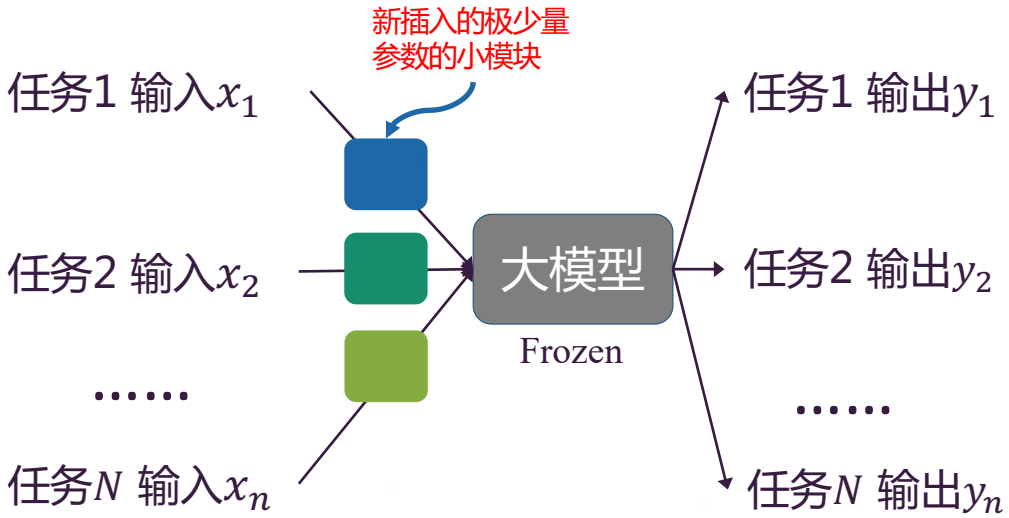
主要研究内容与成果

- 基于词级代表性学习的排序模型预训练
  - ① PROP: 使用代表词预测任务预训练模型 (WSDM'2021)
  - ② B-PROP：自举式预训练代表词预测任务模型(SIGIR'2021)
- 基于文档级对比学习的召回模型预训练
  - ① 基于文档词分布的对比学习方法 (CIKM'2021)
  - ② 基于文档片段的组级别对比学习方法 (SIGIR'2022)
- 基于多层级旁路的参数高效微调方法(CIKM'2022)
  - ① 稳训练的参数高效方法

面向排序和召回的微调



(1) **主流的完全微调方法**，N个任务需要N个大模型同时在线服务

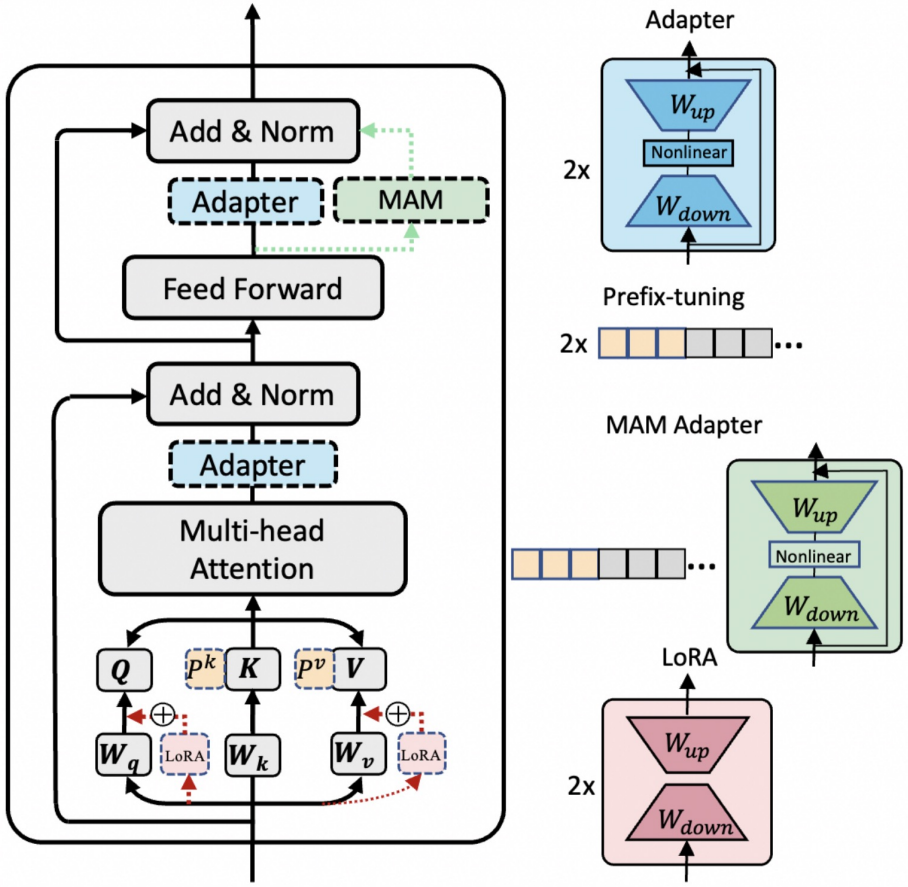


(2) **参数高效的微调方法**，N个任务需要**1**个大模型在线服务

随着模型大小和任务数量的大幅增加，可行性低且成本过高

已有的参数高效的微调方法

- 代表性方法： Adapter (Hously et.al), Prefix-tuning (Liang et.al.), Lora (Hu et.al), MAM-Adapter (He et.al)



Scattered or Connected? An Optimized Parameter-efficient Tuning Approach for Information Retrieval, CIKM 2022

已有方法在 IR 任务的实证分析

- 现有方法在 IR 中的表现是否与在 NLP 中一样好？

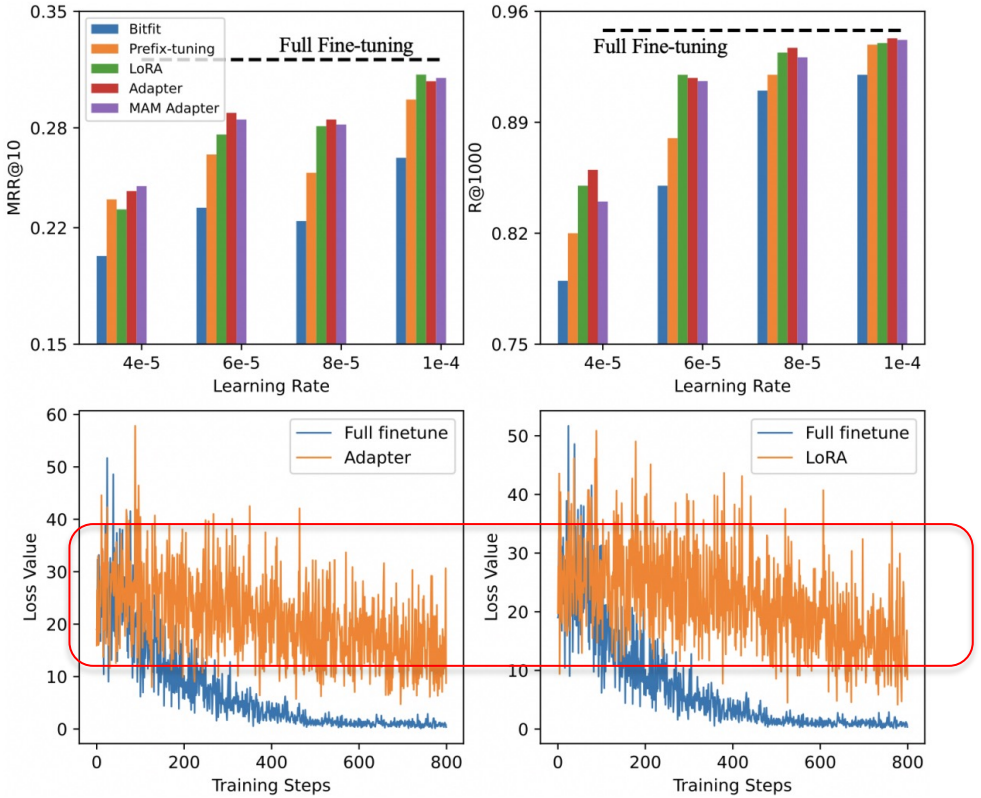
Method	#Params	MARCO Passage		TREC2019 Passage		MARCO Doc		TREC2019 Doc	
		MRR@10	R@1000	nDCG@10	R@100	MRR@100	R@100	nDCG@10	R@100
Full fine-tuning	100%	**0.316**	**0.949**	**0.600**	**0.715**	**0.312**	**0.801**	**0.462**	**0.409**
Bitfit	0.09%	0.262	0.921	0.562	0.677	0.264	0.785	0.437	0.345
Prefix-tuning	0.5% (l=32)	0.294	0.939	0.596	0.692	0.266	0.782	0.423	0.326
Adapter	0.5% (r=16)	0.304	0.941	**0.606**	0.696	0.255	0.770	0.418	0.370
MAM Adapter	0.5% (r=16,l=16)	0.304	0.944	**0.609**	0.712	0.280	0.799	0.458	0.381
LoRA	0.5% (r=16)	0.302	0.943	**0.608**	0.707	0.271	0.794	0.417	0.376
Prefix-tuning	3.6% (l=200)	0.304	0.943	0.580	0.702	0.265	0.775	0.395	0.376
Adapter	6.7% (r=200)	0.316	0.946	0.587	0.687	0.270	0.785	0.433	0.400
MAM Adapter	6.7% (r=200,l=200)	0.314	0.947	**0.616**	**0.720**	0.283	0.792	0.438	0.402
LoRA	6.7% (r=200)	0.316	0.946	0.597	0.715	0.279	0.794	0.417	0.379

- 与 NLP 中令人鼓舞的结果不同，已有方法都无法在所有数据集上使用少于 1% 的模型参数的情况下实现与完全微调相当的性能
- 召回和排序阶段趋势一致

Scattered or Connected? An Optimized Parameter-efficient Tuning Approach for Information Retrieval, CIKM 2022

经验性的观察

- 参数有效但学习效率不高

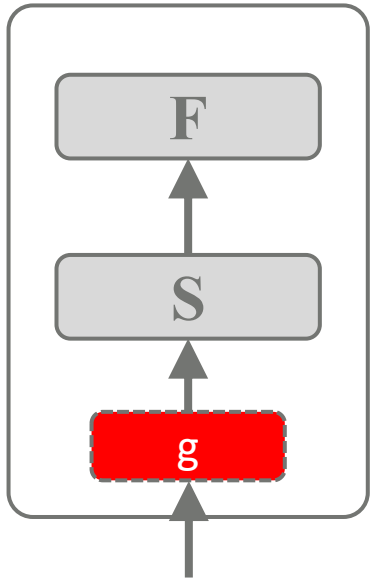


- 对学习率敏感，训练不稳定导致收敛缓慢

Scattered or Connected? An Optimized Parameter-efficient Tuning Approach for Information Retrieval, CIKM 2022

理论分析

- 为什么参数有效调整方法的标准设置在 检索 中效果不好？

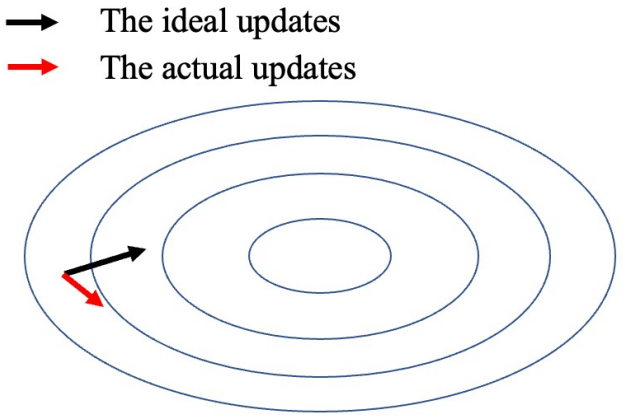


每个模块对最终的优化方向贡献不同

$$\Delta\,(idea) = \Delta\,F, \Delta\,S, \Delta, g$$

VS.

$$\Delta\,(actual) = \Delta\,g$$

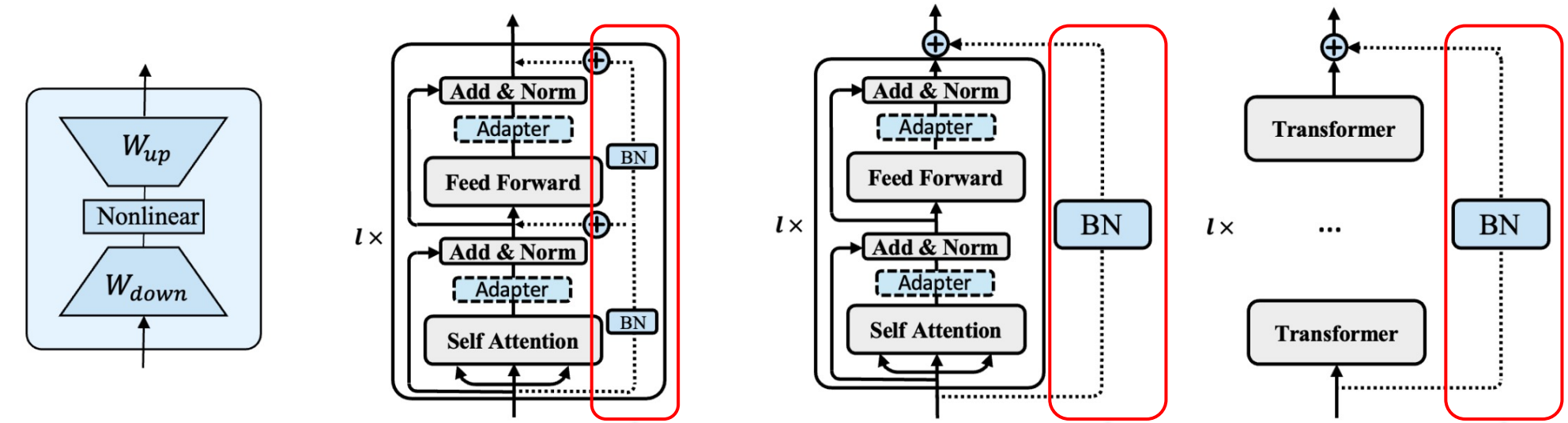


- **理想优化方向与实际更新方向不一致！**

Scattered or Connected? An Optimized Parameter-efficient Tuning Approach for Information Retrieval, CIKM 2022

引入旁路的参数高效微调方法

- 使用 ResNet 的思想，为模型**插入旁路模块**，形成高速通道，稳定梯度流动



(a) The aside module: BN　(b) IAA-S: outside the sub-layer　(c) IAA-L: outside the layer　(d) IAA-M: outside the model

$$\Delta\,(actual) = \Delta\, g = \frac{dL}{dF}\frac{dF}{dS}\frac{dS}{dg} \quad \textit{VS.} \quad \Delta\,(actual) = \Delta\, g \approx \frac{dL}{dg}$$

Scattered or Connected? An Optimized Parameter-efficient Tuning Approach for Information Retrieval, CIKM 2022

实验结果

Table 4: Comparisons between IAA and the baselines at the retrieval stage. Two-tailed t-tests demonstrate the improvements of IAA over baselines are statistically significant ($p \leq 0.05$). $*$ indicate significant improvements over full fine-tuning. † indicate significant improvements over best parameter-efficient tuning methods (PET) at the same setting.

Method	#Params	MARCO Passage		TREC2019 Passage		MARCO Doc		TREC2019 Doc	
		MRR@10	R@1000	nDCG@10	R@100	MRR@100	R@100	nDCG@10	R@100
Full fine-tuning	100%	0.316	0.949	0.600	0.715	**0.312**	**0.801**	**0.462**	**0.409**
Best PET	0.5%	0.304	0.944	0.609	0.712	0.280	0.799	0.458	0.381
IAA-S Adapter	0.5% (r=8,ar=8)	0.312†	0.941	0.605	0.719	0.285	0.785	0.454	0.384
IAA-L Adapter	0.5% (r=12,ar=12)	0.314†	0.943	0.615†	0.735*	0.292	0.792	0.446	0.391
IAA-M Adapter	0.5% (r=15,ar=24)	0.309	0.941	0.602	0.721	0.287	0.782	0.449	0.385
Best PET	6.7%	0.316	0.946	0.616	0.720	0.283	0.792	0.438	0.402
IAA-S Adapter	6.7% (r=100,ar=100)	0.324	0.947	0.581	0.719	0.290	0.798	0.441	0.398
IAA-L Adapter	6.7% (r=50,ar=300)	**0.327**†*	**0.951**	**0.617***	**0.735**†	0.295†	0.795	0.439	0.395
IAA-M Adapter	6.7% (r=185,ar=960)	0.321	0.948	0.592	0.710	0.285	0.793	0.437	0.402

- 在召回阶段， IAA 模型通过**学习不到 1% 的模型参数**，实现了与完全微调相当的性能，并且**明显优于最好的已有的方法**
- 排序阶段的趋势一致

Scattered or Connected? An Optimized Parameter-efficient Tuning Approach for Information Retrieval, CIKM 2022

旁路对于收敛性的影响

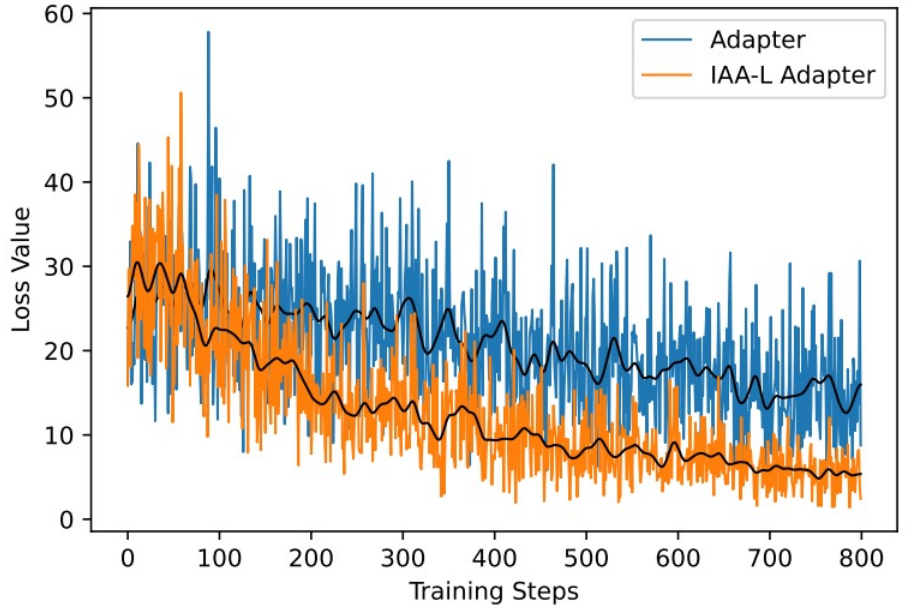


Figure 5: The loss value over training steps.

- **IAA-L Adapter 的 loss 值比 Adapter 低，收敛速度也比 Adapter 快**

Scattered or Connected? An Optimized Parameter-efficient Tuning Approach for Information Retrieval, CIKM 2022

面向排序和召回模型的微调方法

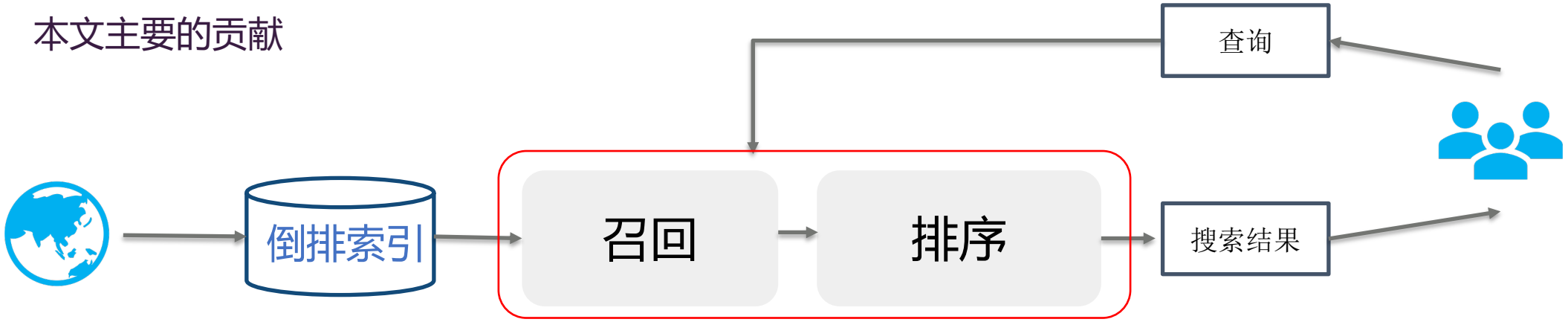
- 总结
  - ① 对已有的参数高效的方法在排序和召回阶段进行了丰富的实证研究
  - ② 发现了已有方法学习不高效的问题，并从理论层面分析了这一现象
  - ③ 提出引入多层级旁路的参数高效学习方法，在检索上使用1%的参数量即可达到甚至超越完全微调

目录

- 研究背景及目标
- 主要研究内容与成果
- 总结与展望
- 攻读博士学位期间的学习和科研情况

总结

- 本文主要的贡献



①基于词级代表性学习的排序模型预训练方法（WSDM'2021,SIGIR'2021）

✓解决对异质文本进行细粒度相关性关系建模的问题

②基于文档级对比学习的召回模型预训练方法（CIKM'2022,SIGIR'2022）

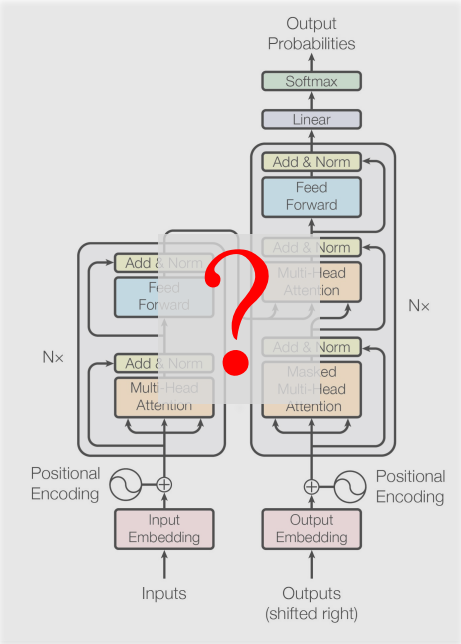
✓解决了学习具有区分力的文档表达的问题

③基于多层级旁路的参数高效微调方法（CIKM'2022）

✓解决已有方法学习不高效的问题

展望

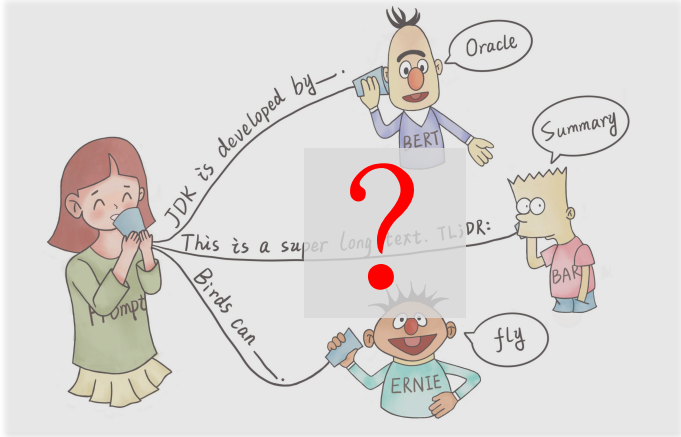
模型结构



+

预训练任务

提示学习



(1) Transformer对IR是最优结构吗？(2) 有没有直接建模相关性的学习方法？ (3) 无需标注数据的提示学习方法？

目标：一个在各个检索领域，各种检索任务，拿来即用的预训练模型

目录

攻读博士学位期间学习情况

1. 课程完成情况
   - 研究生学位课程
     - ✓ 已完成总共有 40.5 学分，其中学位课 25 学分, GPA 3.66
   - 博士学位课程
     - ✓ 已完成所里专业课：6 学分（机器学习、算法设计与分析）
2. 参与项目情况

项目类别	项目/课题名称	参与时间	角色
实验室自研系统	自然语言理解引擎	2018年9月	核心开发人员
实验室自研系统	大数据智能分析系统	2019年1月	项目组成员
国家自然科学基金项目	信息检索与评价	2019年9月	项目组成员
国家自然科学基金杰出青年项目	Web信息检索与数据挖掘	2019年9月	项目组成员
国家重点研发计划项目	数据融合分析与可视化技术研究	2019年9月	项目组成员

攻读博士学位期间学习情况已发表论文及专利

共发表8篇文章申请1个专利：5篇国际顶会一作，1篇国际顶会三作，1篇一作CCF-A中文期刊，1篇合作国际期刊

1. **Xinyu Ma**, Jiafeng Guo, et.al, Pre-train a Discriminative Text Encoder for Dense Retrieval via Contrastive Span Prediction (**SIGIR 2022, Oral, 21%, CCF-A**)
2. **Xinyu Ma**, Jiafeng Guo , et.al, B-PROP: Bootstrapped Pre-training with Representative Words Prediction for Ad-hoc Retrieval (**SIGIR 2021, Oral, 21%, CCF-A**)
3. **Xinyu Ma**, Jiafeng Guo, et.al. PROP: Pre-training with Representative Words Prediction for Ad-hoc Retrieval (**WSDM 2021, Oral, CCF-B**)
4. Yixing Fan, Jiafeng Guo, **Xinyu Ma**, et.al, A Linguistic Study on Relevance Modeling in Information Retrieval (**WWW 2021，Oral, CCF-A**)
5. **Xinyu Ma**, Jiafeng Guo, et.al, Scattered or Connected? An Optimized Parameter-efficient Tuning Approach for Information Retrieval (**CIKM 2022, CCF-B**)
6. **Xinyu Ma**, Ruqing Zhang, et.al, A Contrastive Pre-training Approach to Discriminative Autoencoder for Dense Retrieval (**CIKM 2022, CCF-B**)
7. Yixing Fan, Xiaohui Xie, Yinqiong Cai, Jia Chen, **Xinyu Ma**, Xiangsheng Li, Ruqing Zhang, Jiafeng Guo, Xueqi Cheng, Pre-training Methods in Information Retrieval(**FnTIR期刊**)
8. **马新宇**, 范意兴, 郭嘉丰, 张儒清, 苏立新, 程学旗。关于短文本匹配的泛化性和迁移性的研究分析(**《计算机研究与发展》，CCF-A，中文核心期刊**）
9. 程学旗, 郭嘉丰, 范意兴, 张儒清, 赵恒, **马新宇**。发明名称：基于BERT模型的文档关键词抽取方法及装置，2021.02.02 （已申请，未授权）

谢谢！

请各位老师提问指导