面向多模态文档的语义向量检索技术

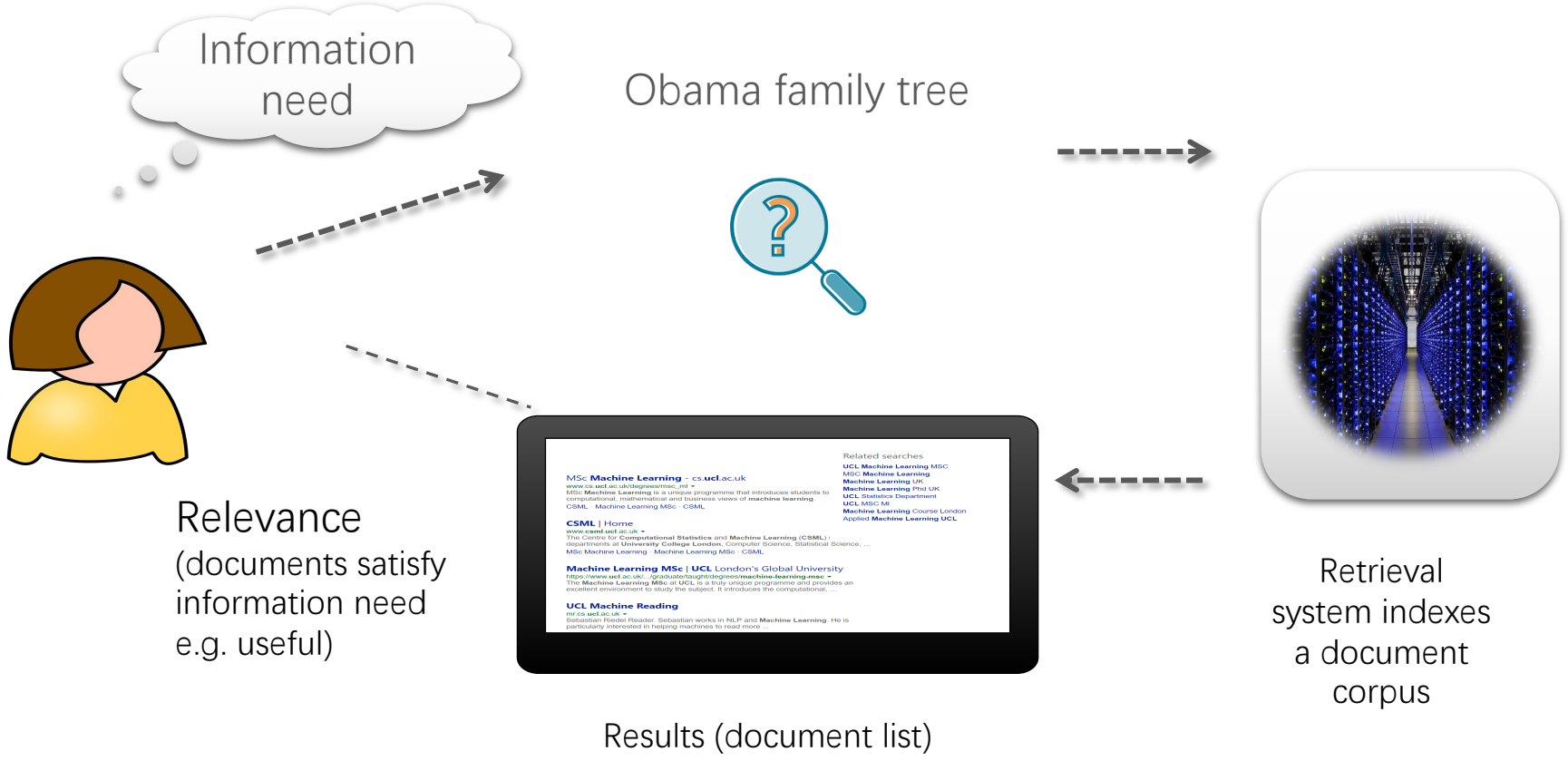
刘正皓

liuzhenghao@mail.neu.edu.cn

CCKS 2023

信息检索及应用

- 信息检索（Information Retrieval, IR）



Relevance between text queries and documents

信息检索及应用

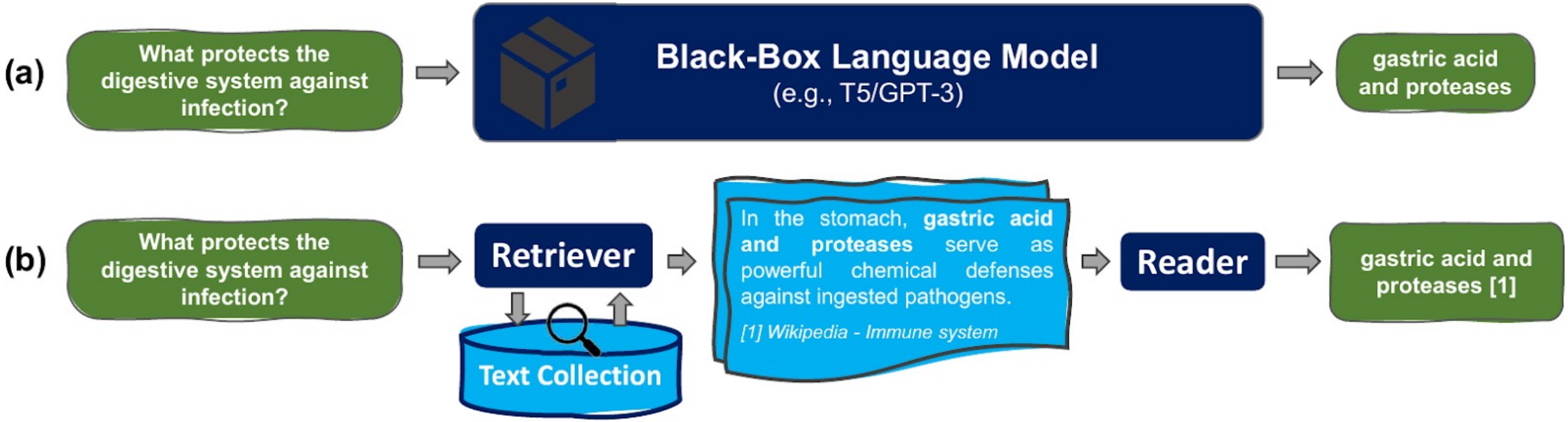
- 信息检索应用
  - 文档检索
    - Query: Obama family tree
    - Document:
      - **Family of Barack Obama** - Wikipedia
      - Barack **Obama Family Tree** along with family connections to other famous kin. Genealogy charts for Barack Obama may include up to 30 generations of ...
  - 自动问答
    - Query: Who is Barack Obama's sister?
    - Answer:

Maya Soetoro-Ng

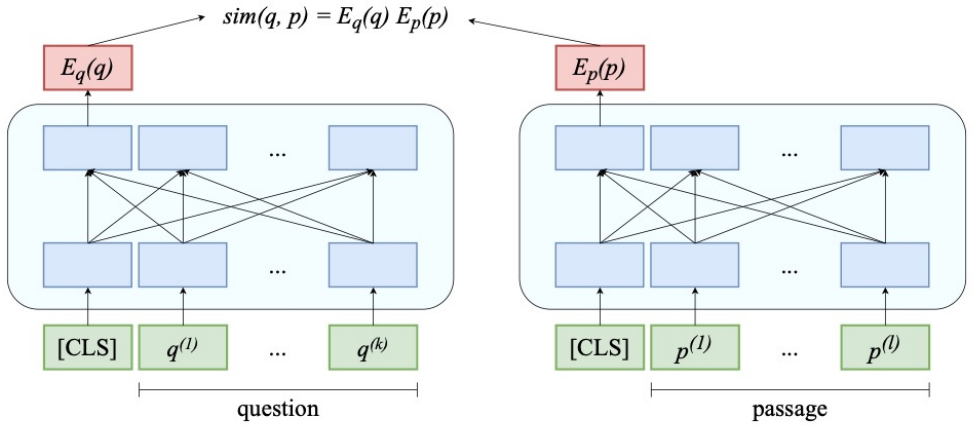
Auma Obama

信息检索及应用

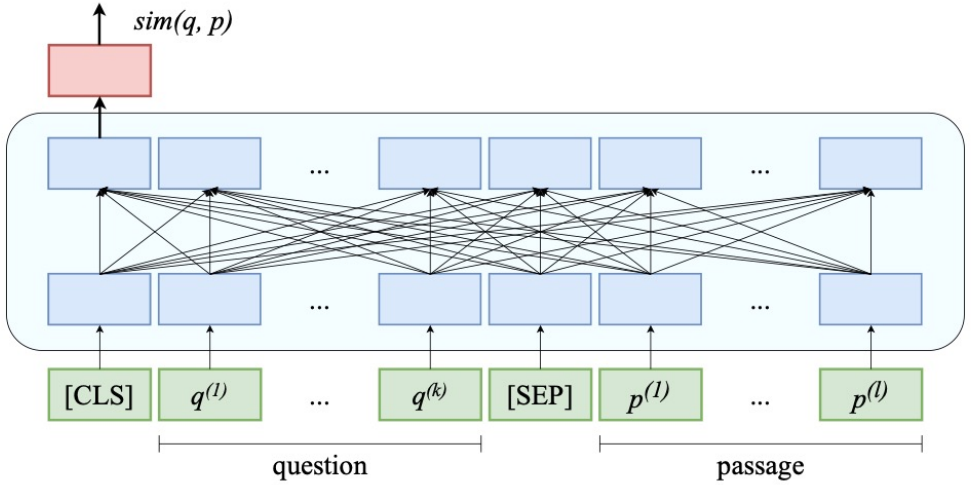
- 信息检索应用
  - 检索增强式语言模型
  - 模型所具有的幻觉问题



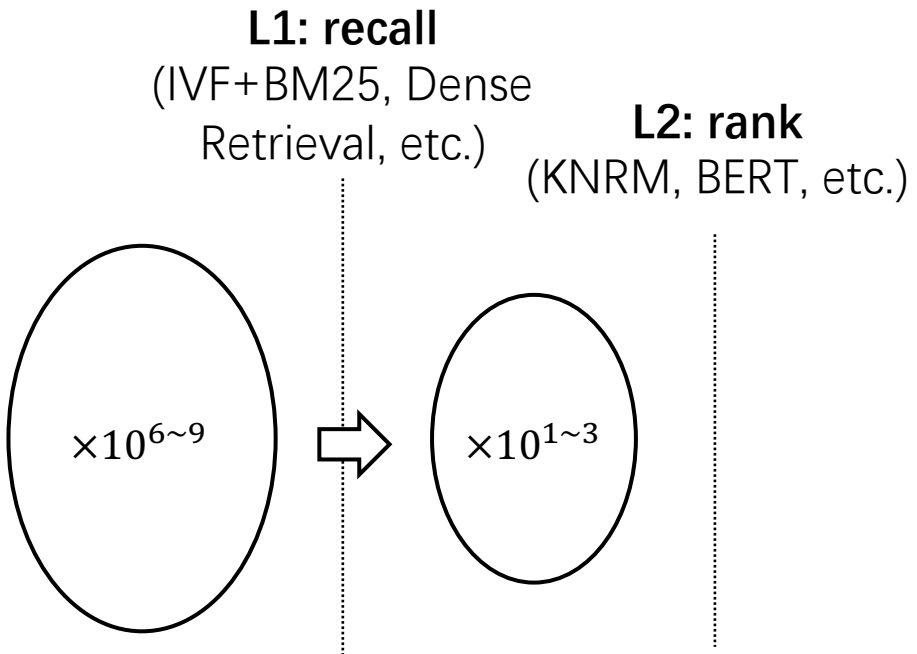
神经信息检索检索



(a) A dual-encoder based on pre-trained LMs.



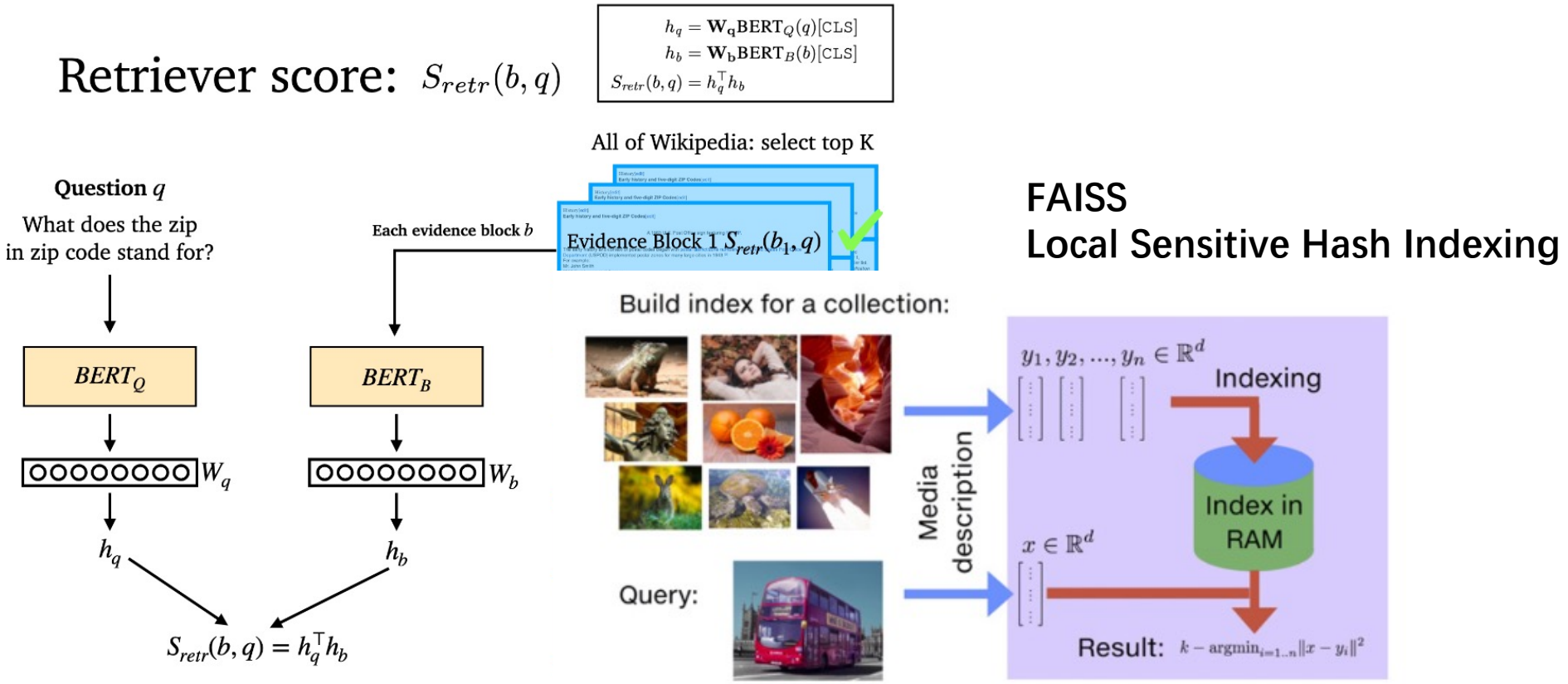
(b) A cross-encoder based on pre-trained LMs.



- **Rank**: high-precision, KPI-oriented
- **Recall**: fast, accurate, relevance-oriented

稠密向量检索

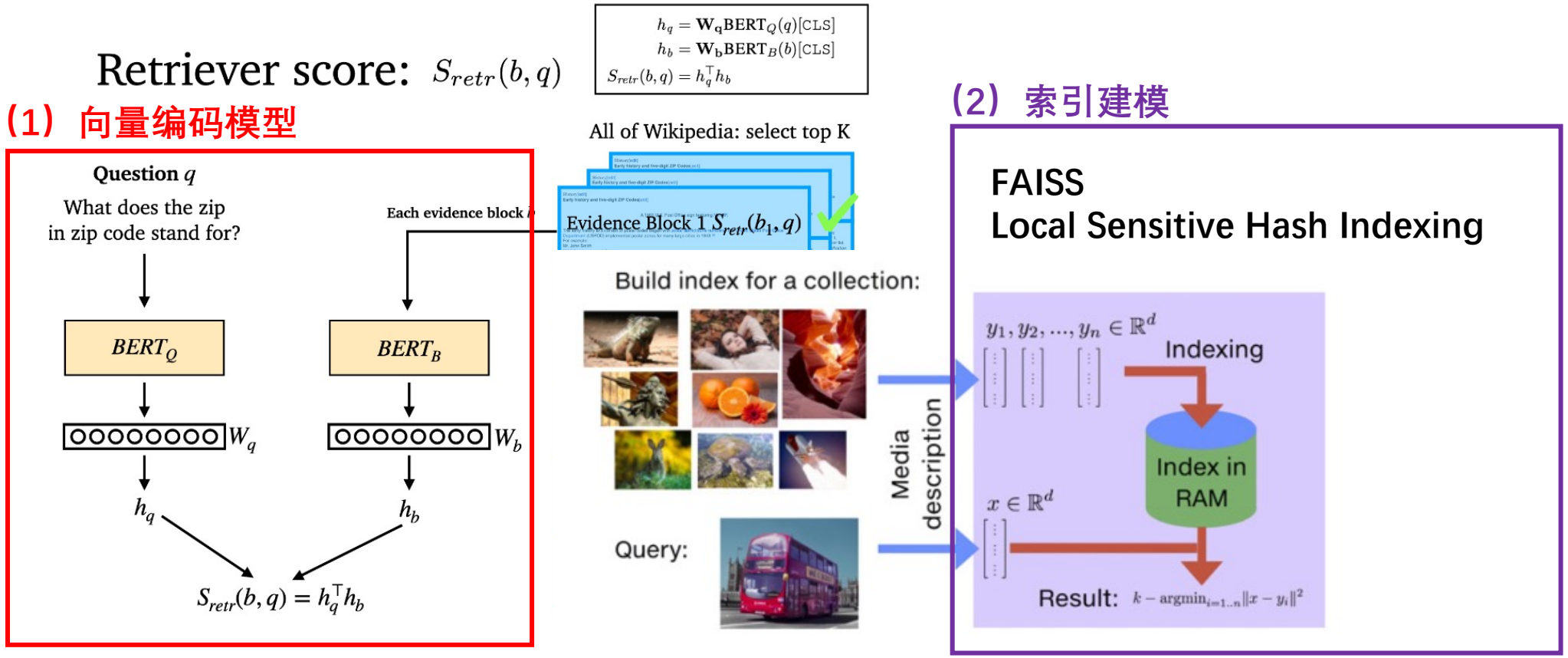
- **DPR** (Karpukhin et al., 2020), **ColBERT** (Khattab and Zaharia, 2020), **ANCE** (Xiong et al., 2021), **Contriever** (Izacard et al., 2022)



(FAISS) Johnson et al - 2017 - Billion-scale similarity search with GPUs
https://github.com/danqi/acl2020-openqa-tutorial/blob/master/slides/part5-dense-retriever-e2e-training.pdf

稠密向量检索

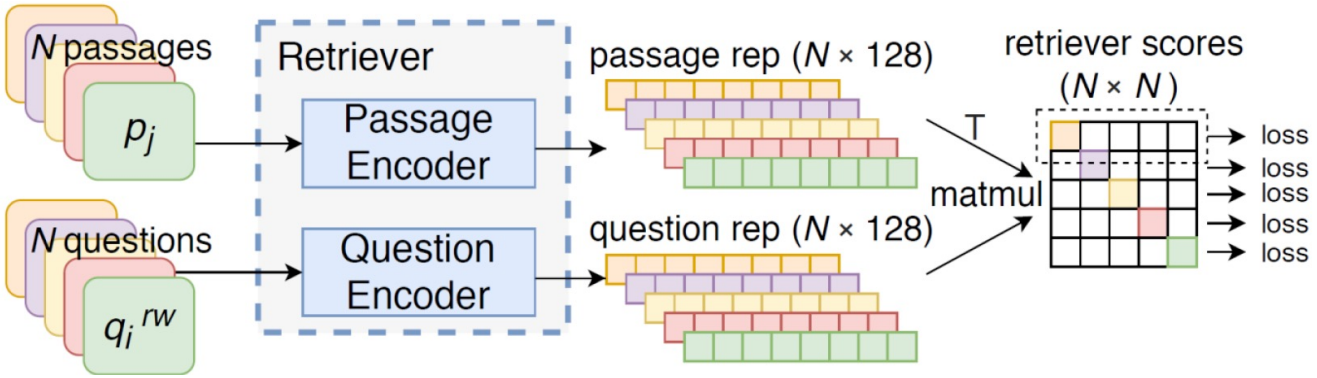
- 模型架构
  - 向量编码模型（BERT、RoBERTa、T5、GPT2）
  - 索引建模（PQ、IVF）



(FAISS) Johnson et al - 2017 - Billion-scale similarity search with GPUs
https://github.com/danqi/acl2020-openqa-tutorial/blob/master/slides/part5-dense-retriever-e2e-training.pdf

稠密向量检索

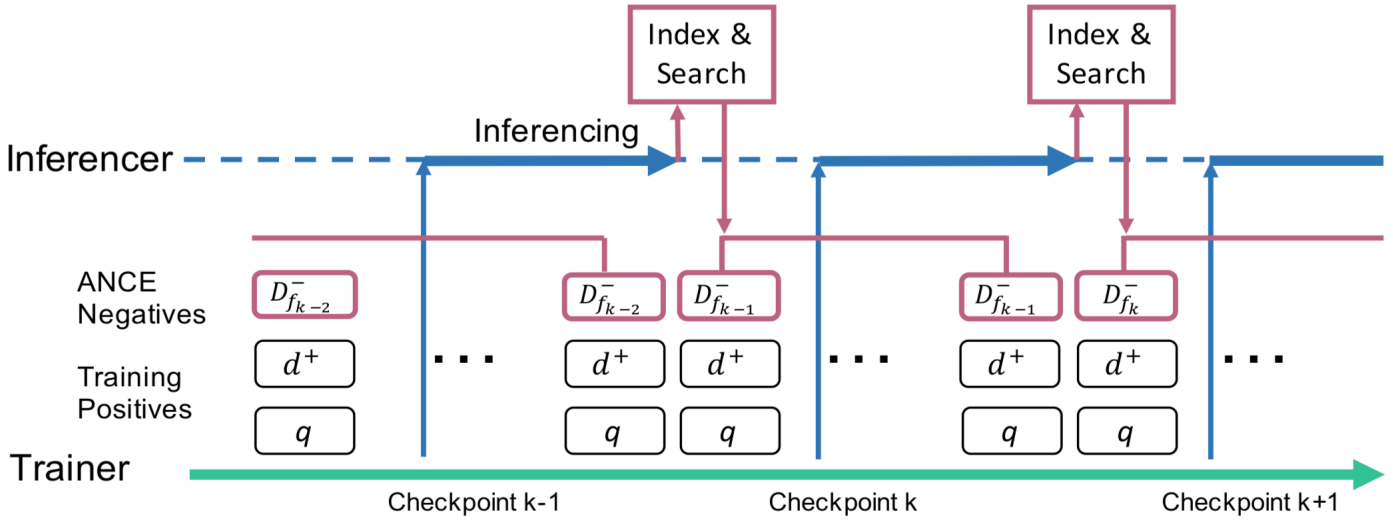
- 对比学习训练
  - 正例选择
    - 数据集中所提供
    - 采用BM25检索出来的包含答案的段落
  - 负例选择
    - 随机负例选取：语料库中随机选择的段落文本
    - BM25负例选取：BM25检索得到的不包含答案字符串的段落文本
    - 批内负例选取：与其他问题相关的段落文本



Karpukhin et al., Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.

稠密向量检索

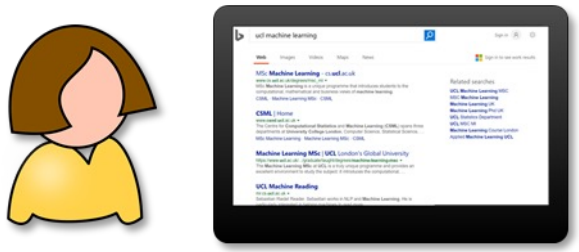
- ANCE
  - ANCE提供高效的编码方法
    - 异步更新的 ANN 索引
  - 使用稠密向量检索器检索难负例文档进行训练
    - 避免梯度递减



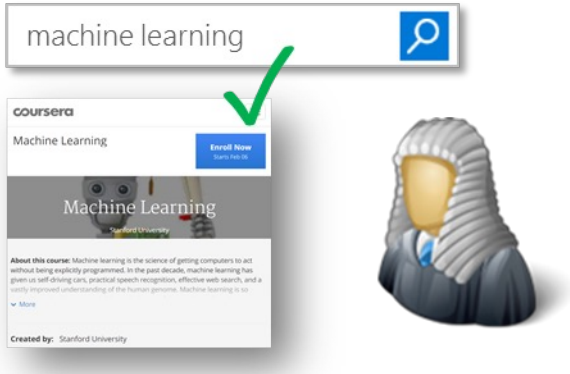
Xiong et al., Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. ICLR 2021.

神经信息检索面临的挑战

- 神经信息检索模型是完全监督训练的
  - 传统的神经信息检索模型使用人类标签作为评估的基本事实并希望在人类标签上训练我们的排名模型
  - 来自垂直领域的用户交互数据通常难以获得
    - 例如：TREC-COVID数据仅有50条标注数据
  - 漏标问题（Hole Rate）
  - 隐私问题



user interaction / click data

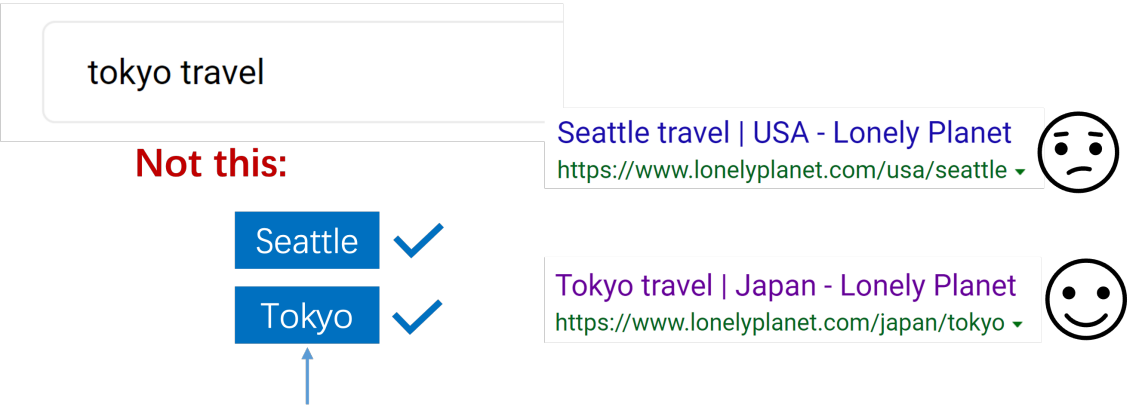


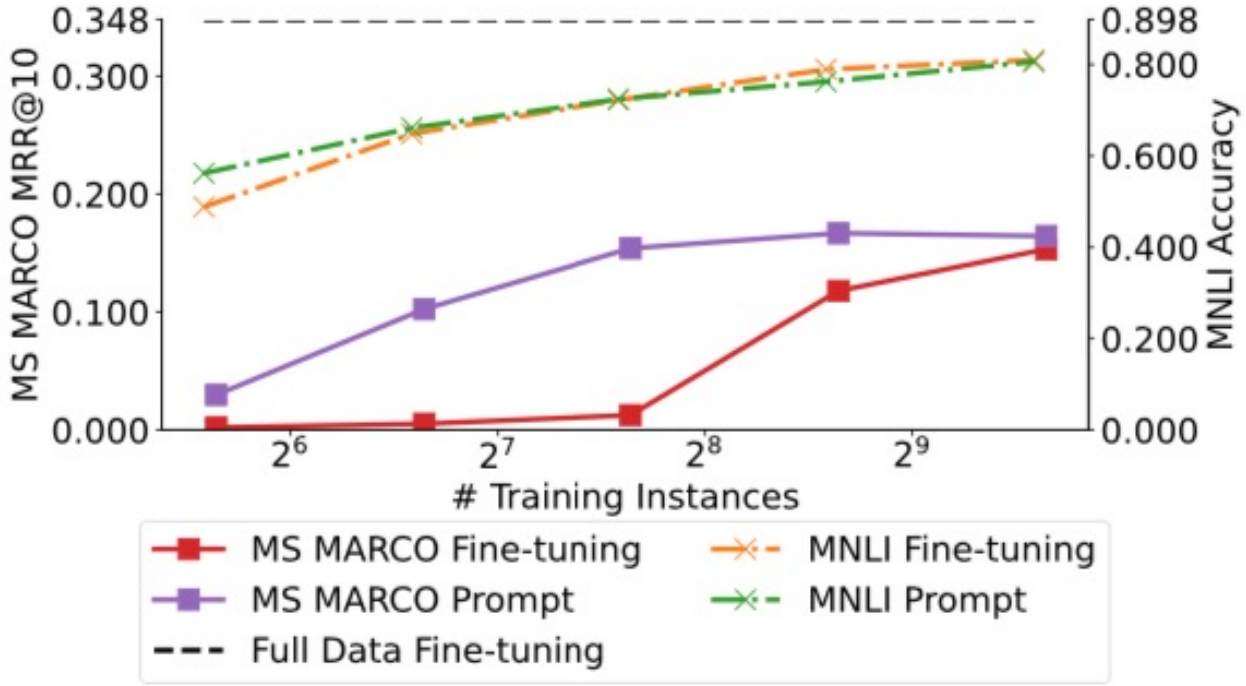
human annotated labels

基于数据增强的神经信息检索模型训练方法

（1）基于大规模预训练语言模型的文本表征方法

- 预训练语言模型需要相关性信号进行微调





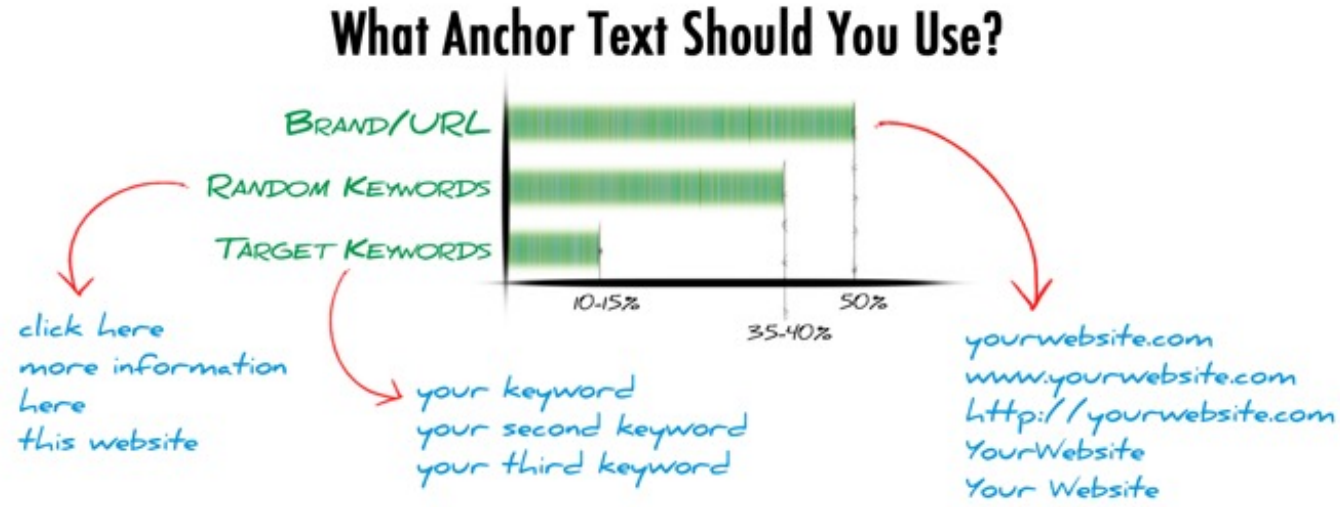
Hu et al P3 Ranker: Mitigating the Gaps between Pre-training and Ranking Fine-tuning with Prompt-based Learning and Pre-finetuning. SIGIR 2022.

基于数据增强的神经信息检索模型训练方法

- 锚文本与查询文本相似
- 锚定文档数据可能非常嘈杂，并且噪声数据可能会影响神经网络信息检索方法的性能

New York City Transit Police

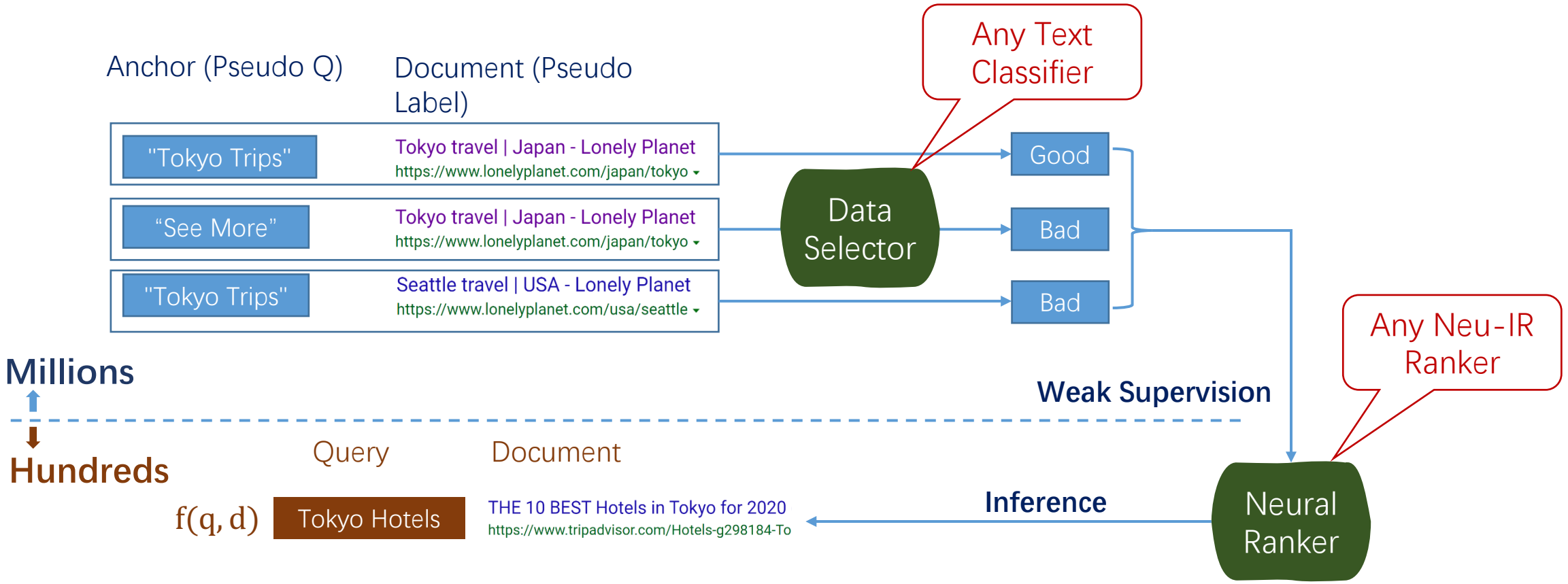
The **New York City Transit Police Department** was a law enforcement agency in New York City that existed from 1953 to 1995, and is currently part of the NYPD. The roots of this organization go back to 1936 when Mayor Fiorello H. La Guardia authorized the hiring



Zhang et al., Selective Weak Supervision for Neural Information Retrieval. WebConf 2020.

基于数据增强的神经信息检索模型训练方法

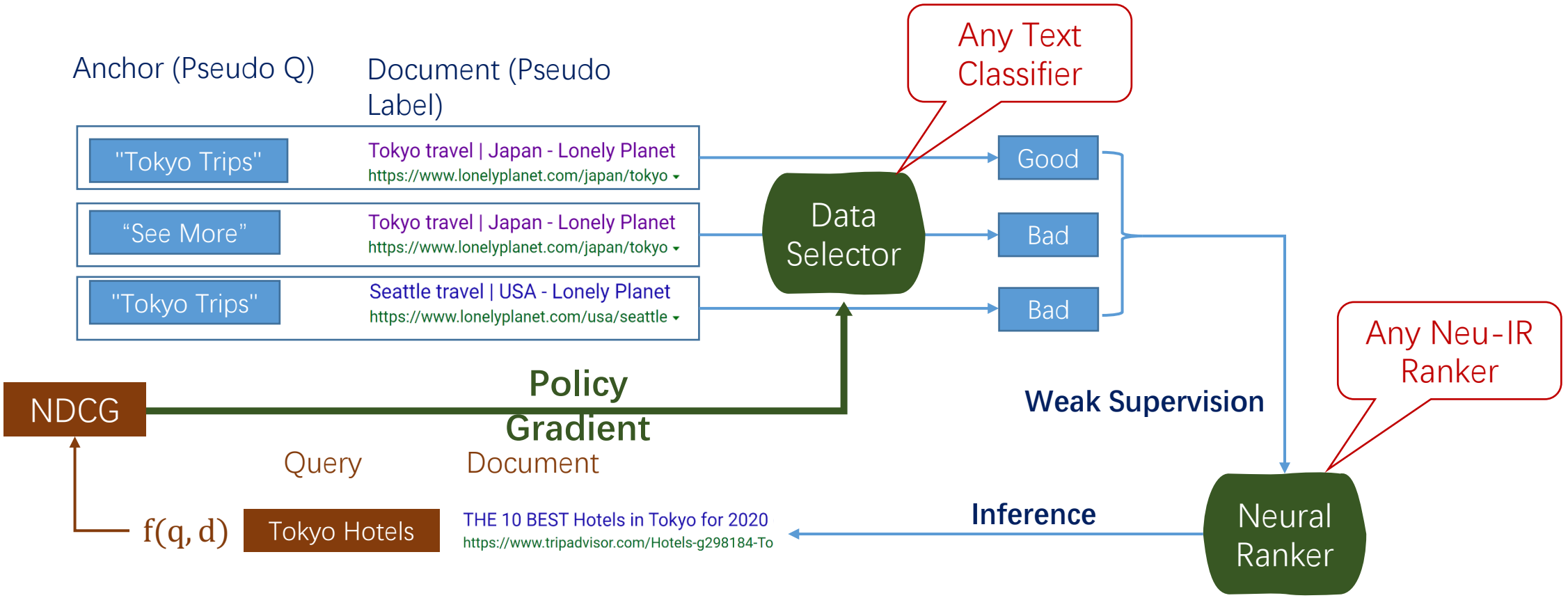
- 基于强化学习的数据筛选模型 (ReinfoSelect)



Zhang et al., Selective Weak Supervision for Neural Information Retrieval. WebConf 2020.

基于数据增强的神经信息检索模型训练方法

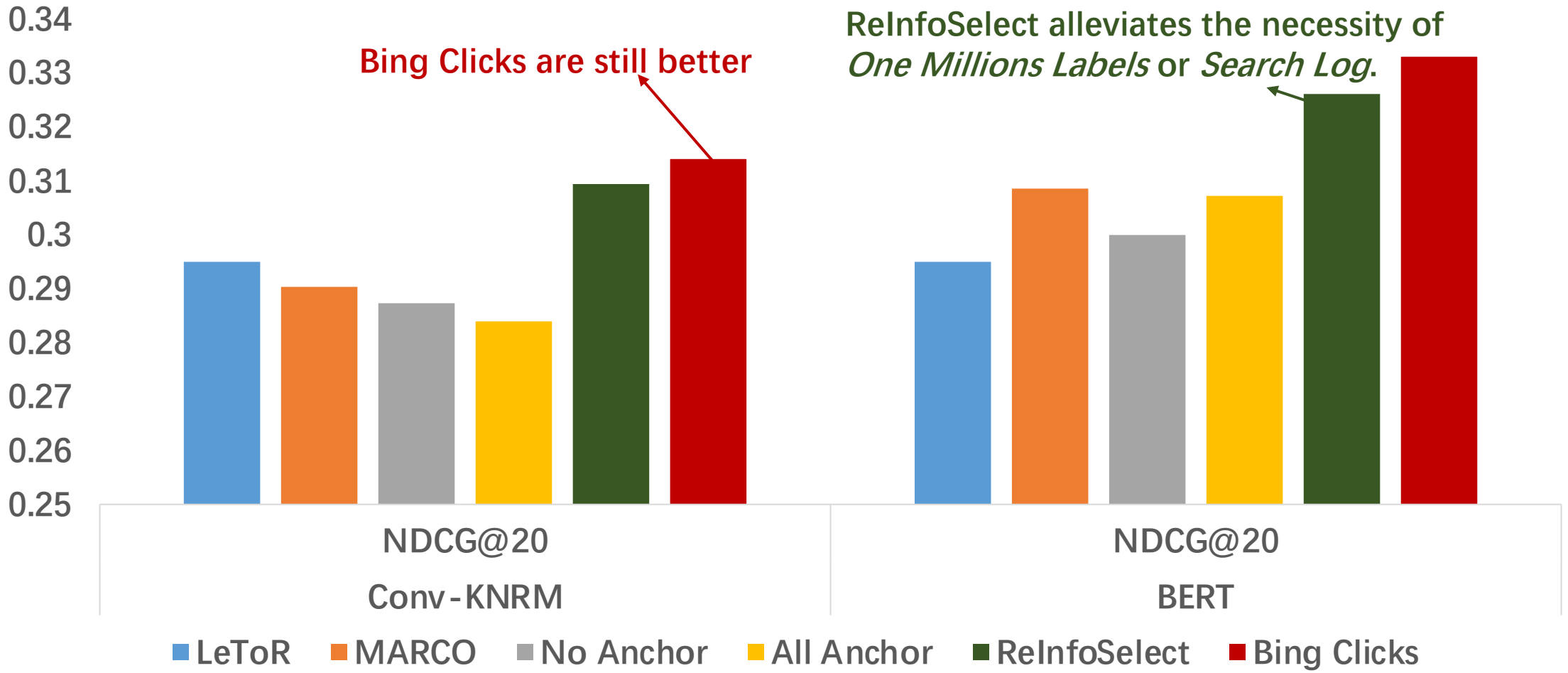
- 基于强化学习的数据筛选模型 (ReinfoSelect)



Zhang et al., Selective Weak Supervision for Neural Information Retrieval. WebConf 2020.

基于数据增强的神经信息检索模型训练方法

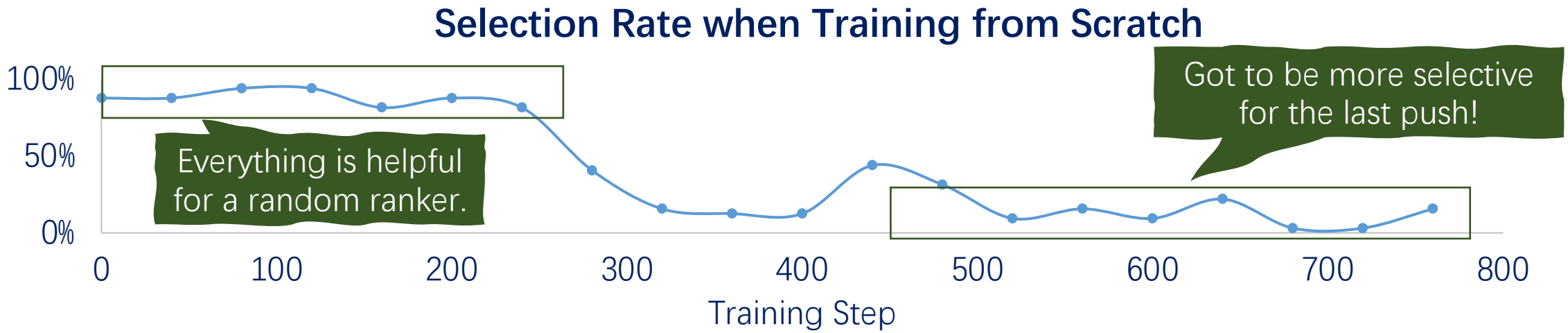
- 基于强化学习的数据筛选模型 (ReinfoSelect)

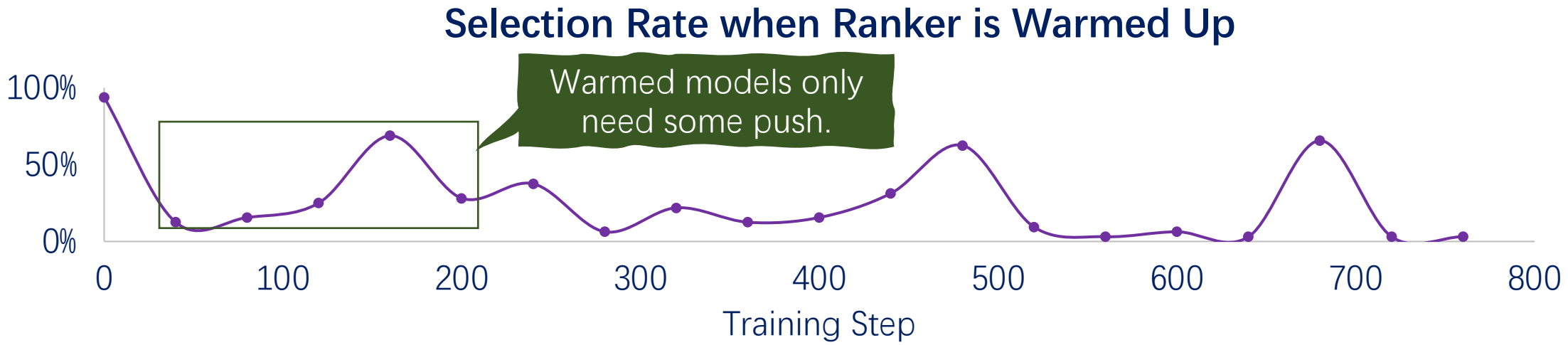


Zhang et al., Selective Weak Supervision for Neural Information Retrieval. WebConf 2020.

基于数据增强的神经信息检索模型训练方法

- 基于强化学习的数据筛选模型 (ReinfoSelect)





Zhang et al., Selective Weak Supervision for Neural Information Retrieval. WebConf 2020.

基于数据增强的神经信息检索模型训练方法

- 但是，锚定文档数据仅在Web域中可用

我们能为不同的检索场景生成一些相关性标签吗？

Enterprise Search

Extreme Verticals

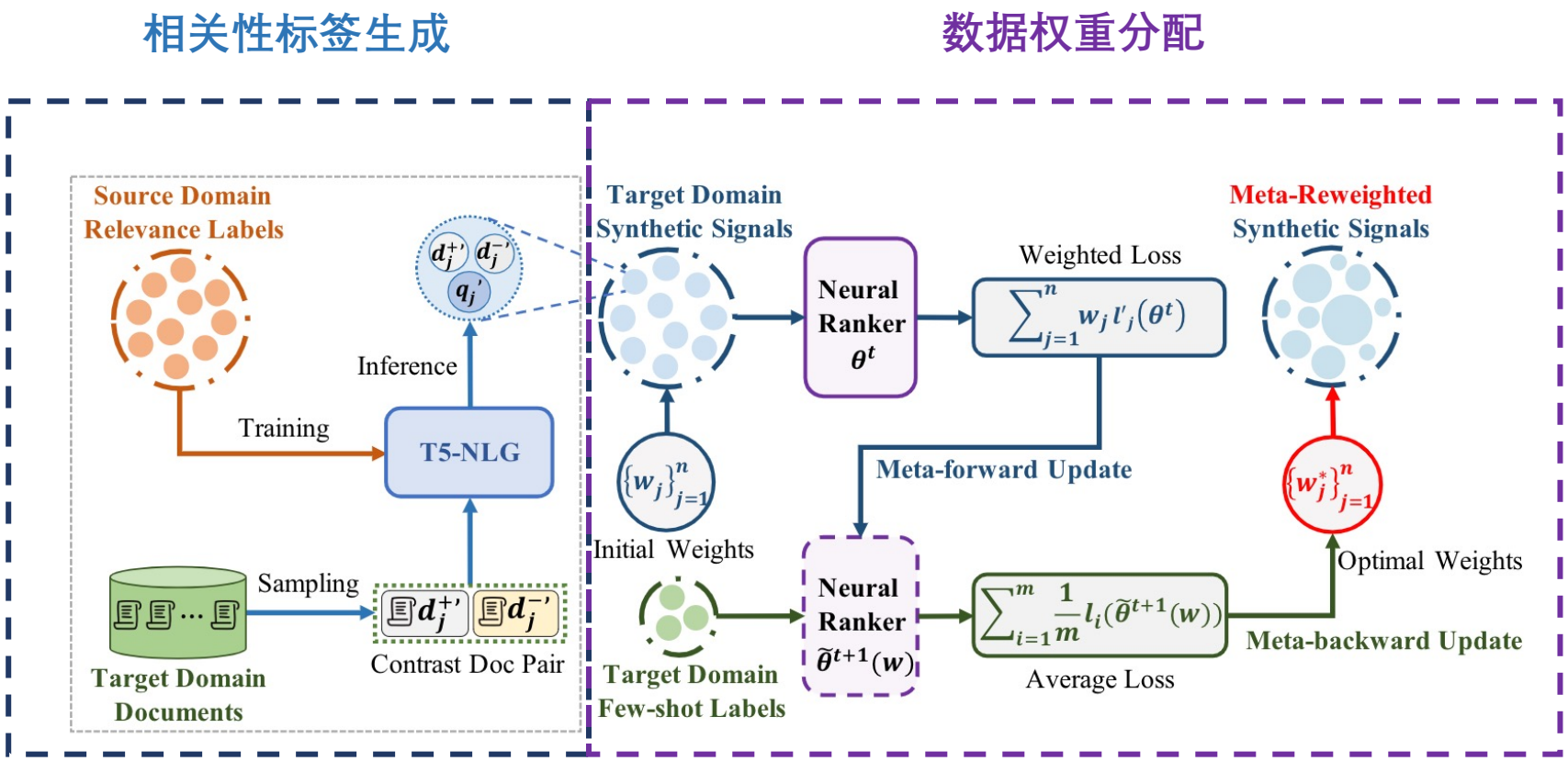
Cloud Search

Personalized Search

IR Community

基于数据增强的神经信息检索模型训练方法

- MetaAdaptRank



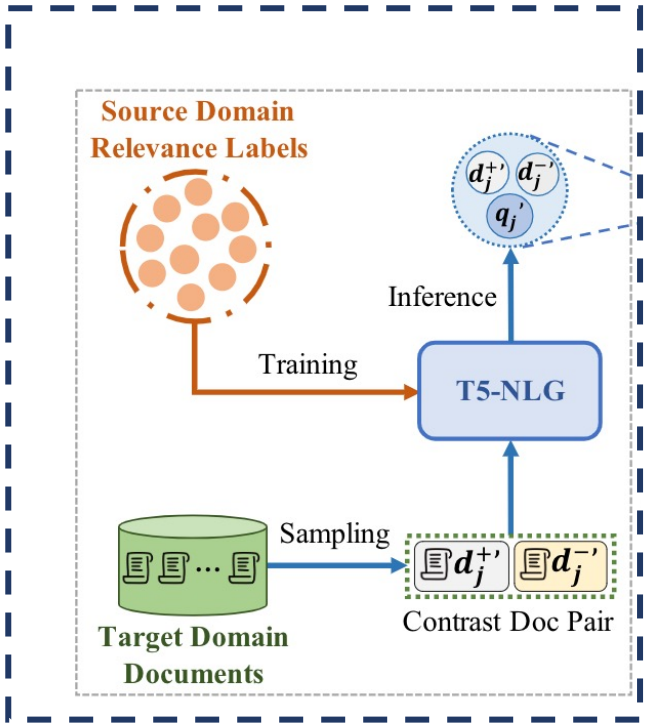
Sun et al., Few-Shot Text Ranking with Meta Adapted Synthetic Weak Supervision. ACL 2021.

基于数据增强的神经信息检索模型训练方法

- 问题生成
  - 基于通用领域大规模语料库的训练生成器
  - 为目标域的文档生成用户问题
  - 使用生成模型弥合领域差距
  - 缺点：
    - 生成的查询过于笼统
    - 这些查询可能与多个文档有关

5 (↑)	CTSyncSup: how does **quarantine prevent** covid outbreak SyncSup: covid outbreak **symptoms**	... the importance of the timing of **quarantine** measures before **symptom** onset to **prevent** covid-19 outbreaks how **quarantine**-based measures can **prevent** or suppress an outbreak ...	... furthermore, the effect of infectiousness prior to **symptom** onset combined with a significant proportion we evaluate two procedures: monitoring individuals for **symptoms** onset ...
6 (↓)	CTSyncSup: **covid-19 pandemic** effects on society SyncSup: what is the antiasia *sentiment* in the united states	... examination of community *sentiment* dynamics due to **covid-19 pandemic**: the outbreak of covid-19 has caused unprecedented impacts to people's daily life around the world. virus may cause different mental health issues to people such as depression, anxiety, sadness ...	... mood of india during covid-19 - an interactive web portal based on emotion analysis of twitter data the **covid-19 pandemic** has affected many countries across the world, and disrupted the day to day activities of many people ...

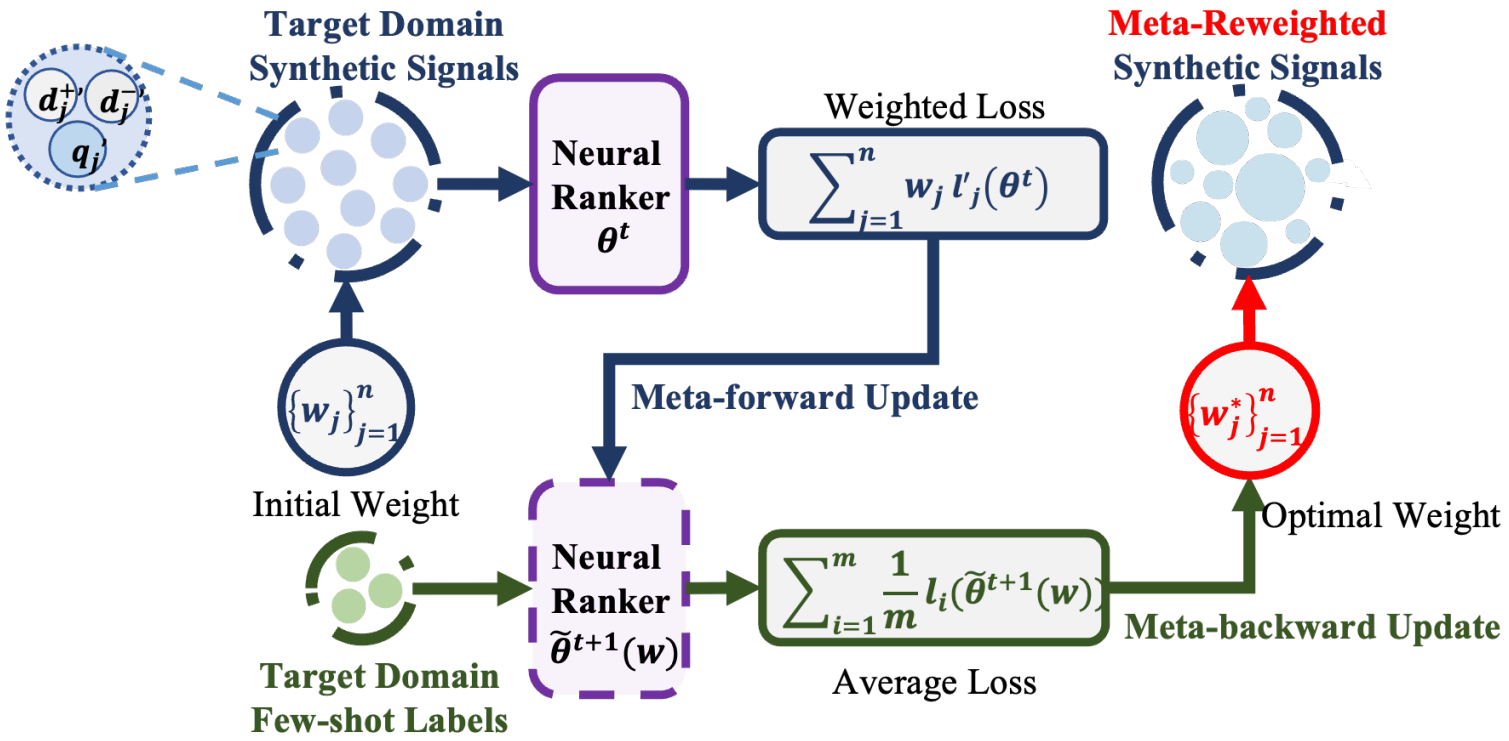
相关性标签生成



Sun et al., Few-Shot Text Ranking with Meta Adapted Synthetic Weak Supervision. ACL 2021.

基于数据增强的神经信息检索模型训练方法

- 重新加权相关性标签
  - 为相关性标签指定初始权重
  - 前向元更新：伪更新神经信息检索模型
  - 后向元更新：计算实际权重
  - 使用元重加权合成信号训练神经信息检索模型



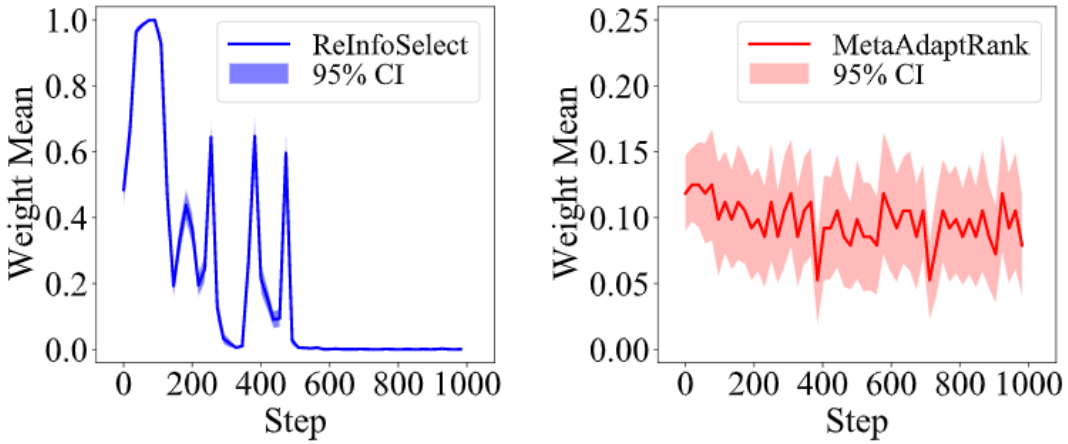
Sun et al., Few-Shot Text Ranking with Meta Adapted Synthetic Weak Supervision. ACL 2021.

基于数据增强的神经信息检索模型训练方法

- MetaAdaptRank可以为弱监督数据分配更细粒度的权重

Methods (Supervision Sources)		ClueWeb09-B (Web)		Robust04 (News)		TREC-COVID (BioMed)	
		NDCG@20	ERR@20	NDCG@20	ERR@20	NDCG@20	P@20
(a)	ReInfoSelect (MS MARCO)	0.3294	0.1760	0.4756	0.1291	$0.8229^{\ddagger}$	$0.8780^{\ddagger}$
(b)	ReInfoSelect (Anchor)	0.3261	0.1669	0.4703	0.1313	0.7891	0.8430
(c)	ReInfoSelect (CTSyncSup)	0.3243	0.1742	$0.4816^{\ddagger}$	0.1334	$0.8230^{\ddagger}$	$0.8800^{\ddagger}$
(d)	MetaAdaptRank (MS MARCO)	$0.3453^{\dagger\ddagger\flat}$	$\mathbf{0.2018}^{\dagger\ddagger\flat\sharp}$	$0.4853^{\ddagger}$	0.1331	$0.8354^{\ddagger\sharp}$	$0.8730^{\ddagger}$
(e)	MetaAdaptRank (Anchor)	0.3374	0.1730	0.4797	0.1314	0.8045	0.8650
(f)	MetaAdaptRank (CTSyncSup)	$0.3416^{\flat}$	$0.1893^{\ddagger\sharp}$	$0.4916^{\dagger\ddagger\sharp}$	$0.1362^{\dagger\sharp}$	$0.8378^{\ddagger\sharp}$	$0.8790^{\ddagger}$
(g)	MetaAdaptRank (MARCO + CTSyncSup)	$\mathbf{0.3498}^{\dagger\ddagger\sharp}$	$0.1926^{\ddagger\flat\sharp}$	$\mathbf{0.4989}^{\dagger\ddagger\flat\natural\sharp}$	$\mathbf{0.1366}^{\dagger\natural}$	$\mathbf{0.8488}^{\dagger\ddagger\flat\natural\sharp}$	$\mathbf{0.8910}^{\ddagger\natural\sharp}$

Table 5: Ranking accuracy of ReInfoSelect and MetaAdaptRank using different supervision sources. Superscripts $\dagger, \ddagger, \flat, \natural, \sharp, \S$ indicate statistically significant improvements over *(a)*†, *(b)*‡, *(c)*$^{\flat}$, *(d)*$^{\natural}$, *(e)*$^{\sharp}$ and *(f)*§.

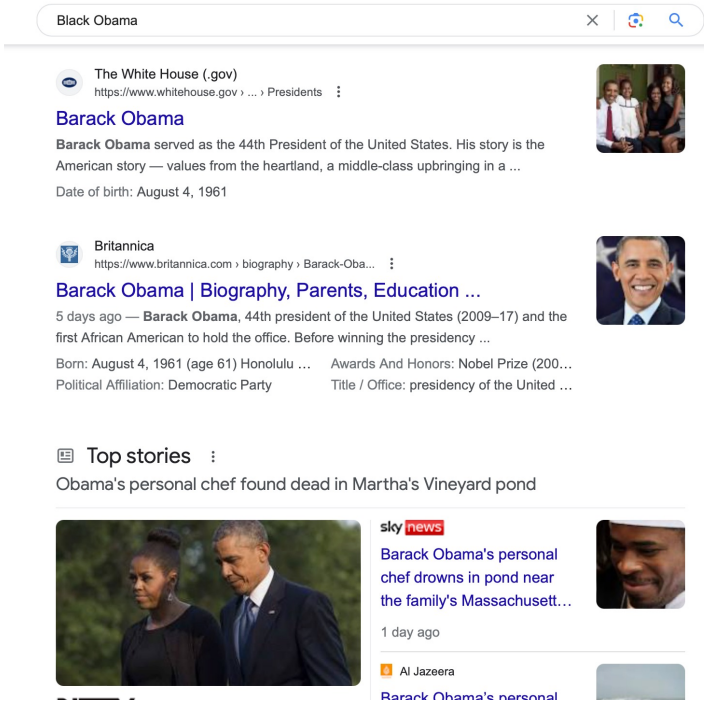


(a) ReInfoSelect. (b) MetaAdaptRank.

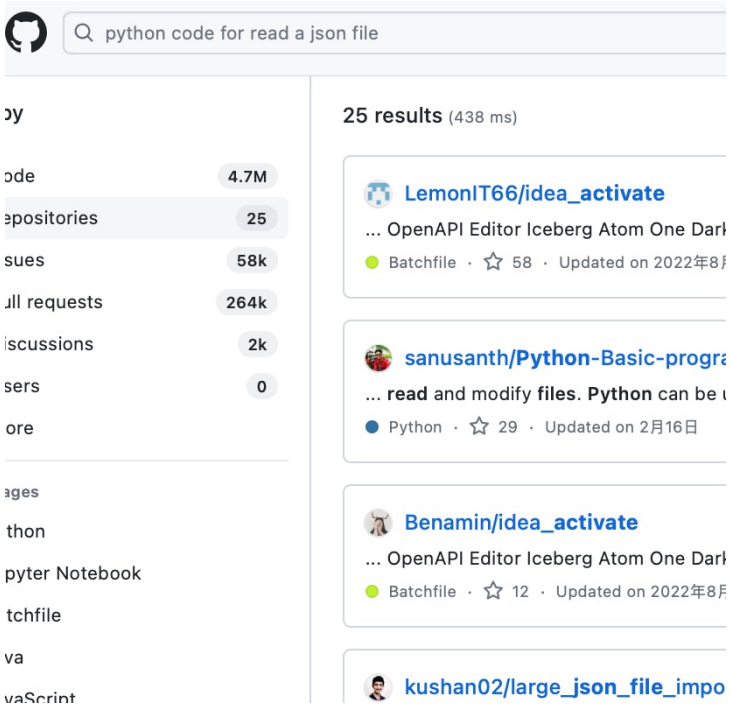
Sun et al., Few-Shot Text Ranking with Meta Adapted Synthetic Weak Supervision. ACL 2021.

面向多模态文档的稠密向量检索模型

- 多模态数据在网络搜索中随处可见



文本和图片

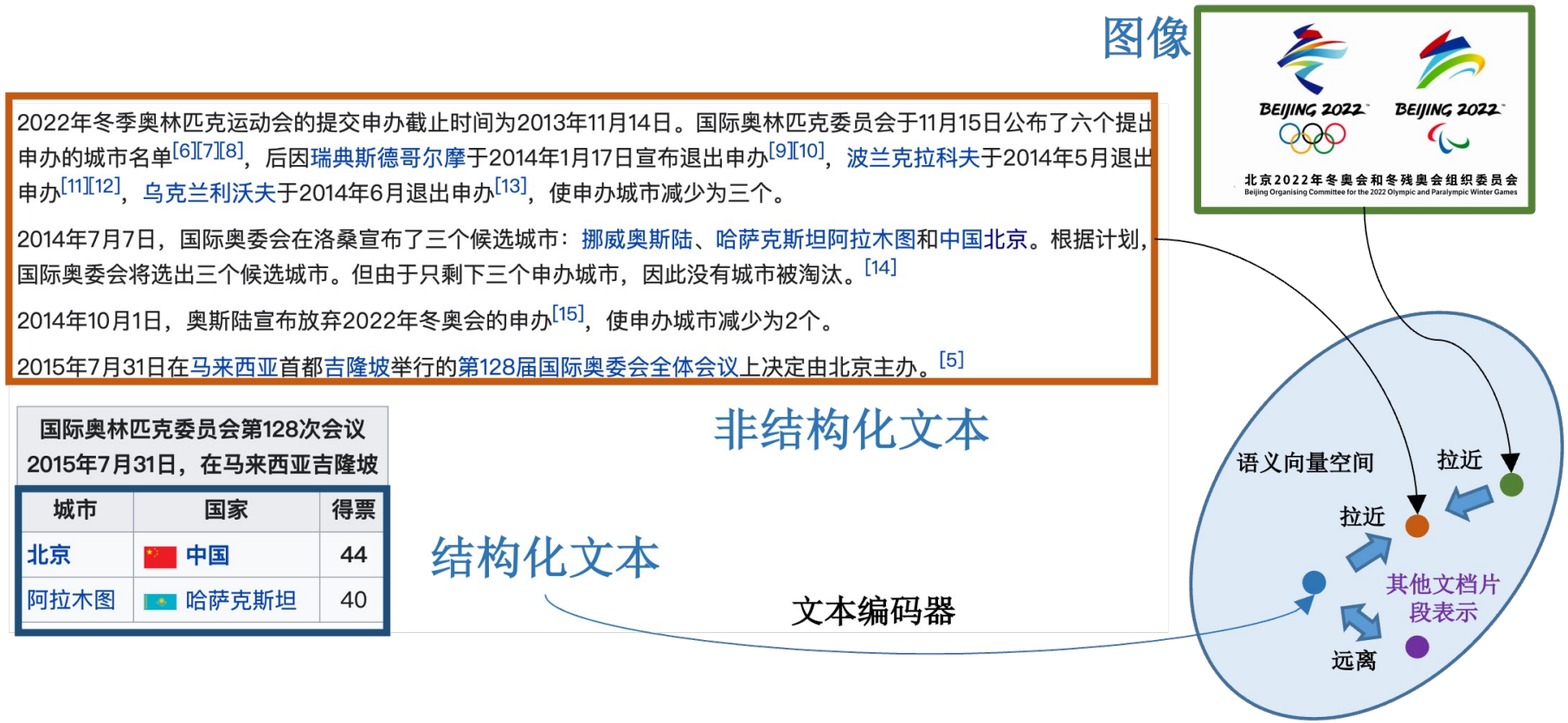


代码

结构化文本

面向多模态文档的稠密向量检索模型

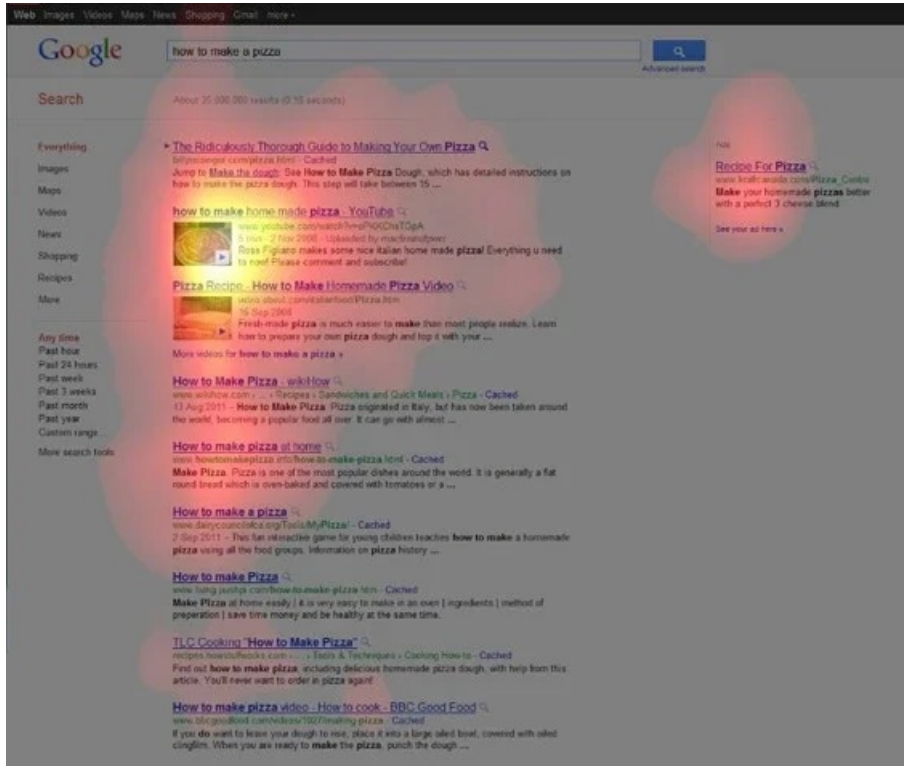
- 多模态数据在网络搜索中随处可见
  - 对结构化文本、非结构化文本和图像进行语义向量表示
  - 对多模态检索结果融合



面向多模态文档的稠密向量检索模型

- 多模态检索任务侧重于如下方面
  - 问题文档相关性建模（relevance modeling）；单/跨模态匹配（single/cross modality matching）；以及模态路由（modality routing）

网页浏览中眼动追踪



多模态检索

Query: What water-related object is sitting in front of the Torre del Reloj?
Retrieval Candidates:

Image1

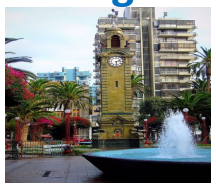
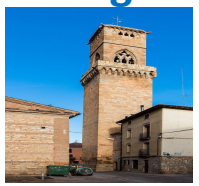


Image2

Image3



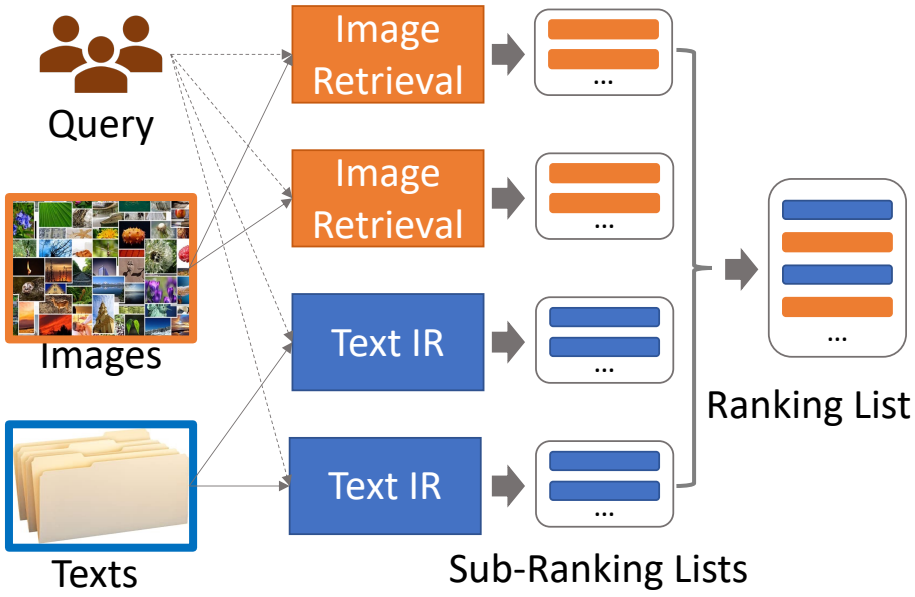
Text1: The Torre del Reloj Spanish is the main city gate of the historic center of Cartagena de Indias.
Text2: The Torre del Reloj is the clock tower, known as Arquillo Clock, and is one of the most emblematic buildings of Chiclana.
Text3: Other landmarks in the city include the Torre del Reloj (Clock Tower).

Liu et al., Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval. ICLR 2023.

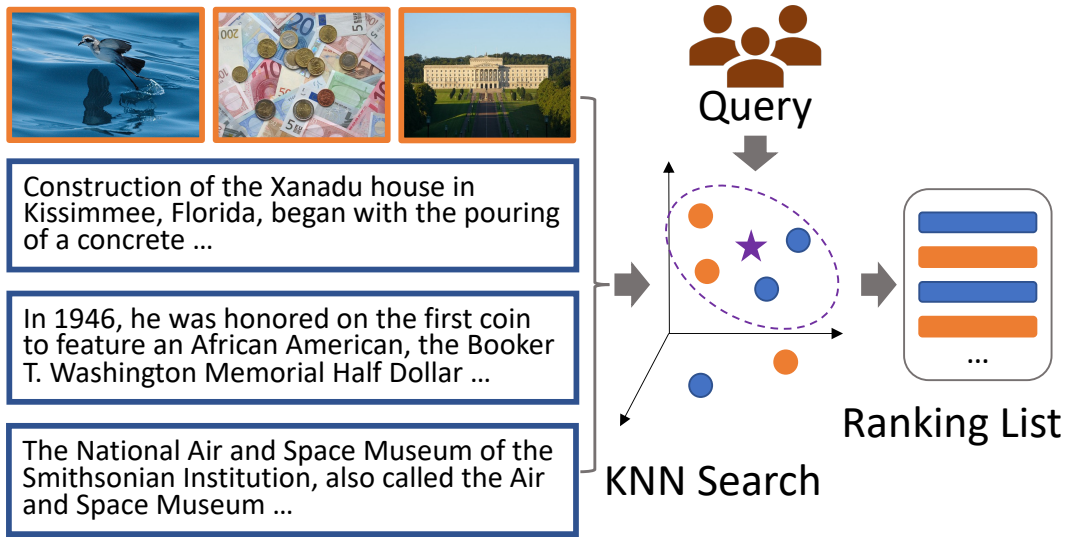
面向多模态文档的稠密向量检索模型

- 多模态文档检索策略
  - 分而治之：从多模态中检索文档，然后融合检索结果
  - 统一视觉语言稠密向量检索：为多模态检索倾斜一个通用嵌入空间，从而统一了相关性建模和检索结果融合阶段

Divide-and-Conquer Retrieval



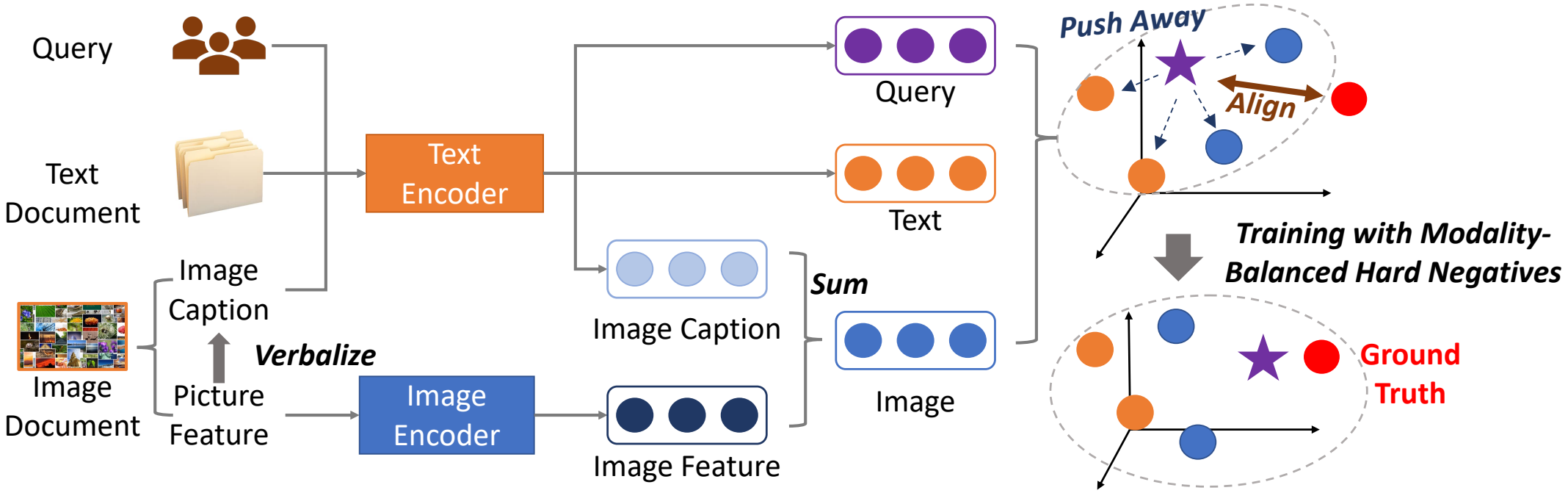
Universal Vision-Language Dense Retrieval



Liu et al., Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval. ICLR 2023.

面向多模态文档的稠密向量检索模型

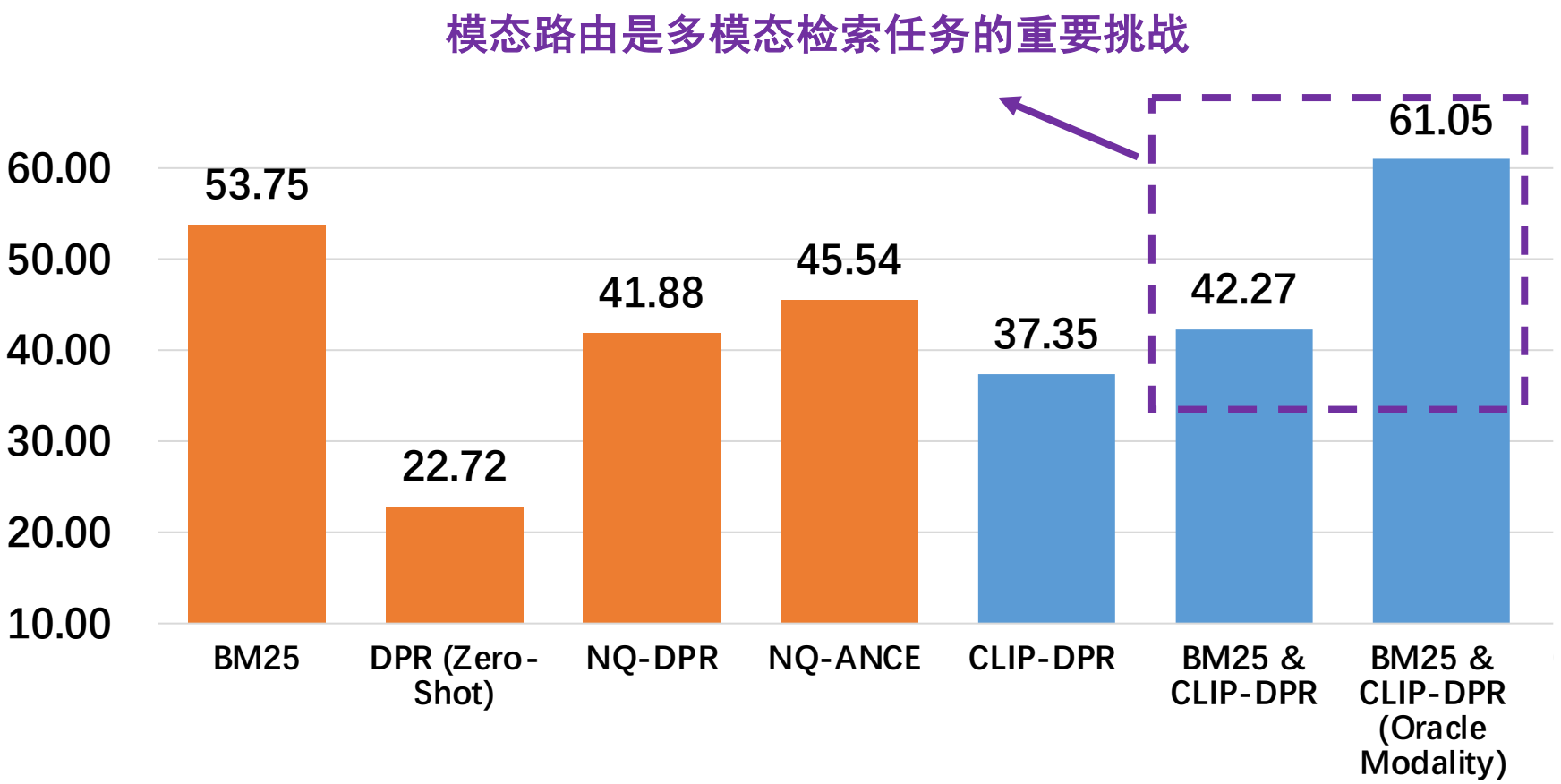
- 统一视觉语言稠密向量检索（Universal Vision-Language Dense Retrieval， UniVL-DR）



Liu et al., Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval. ICLR 2023.

面向多模态文档的稠密向量检索模型

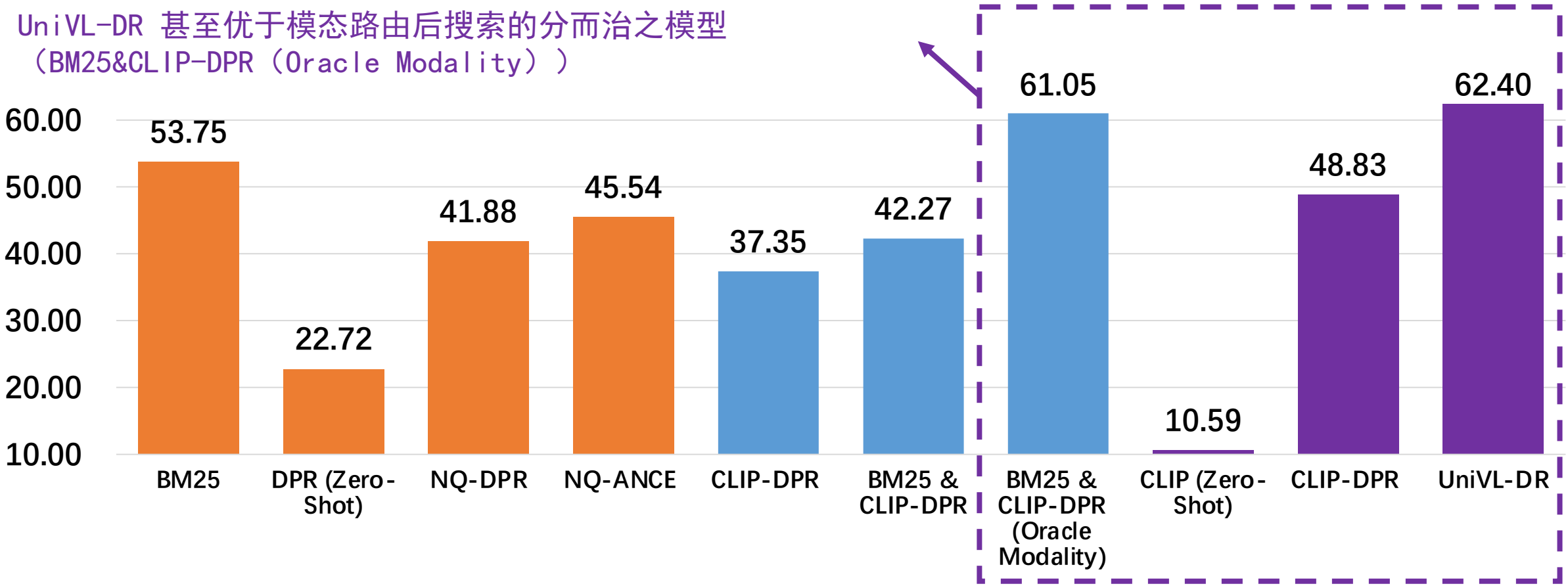
- 多模态检索可以通过分而治之的模型来实现（***Multi-modal->Single/Cross modality retrieval & Fusion***）



Liu et al., Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval. ICLR 2023.

面向多模态文档的稠密向量检索模型

- 学习一种用于查询、文本文档和图像文档的通用嵌入空间（***Multi-modal -> Universal Dense Retrieval***）

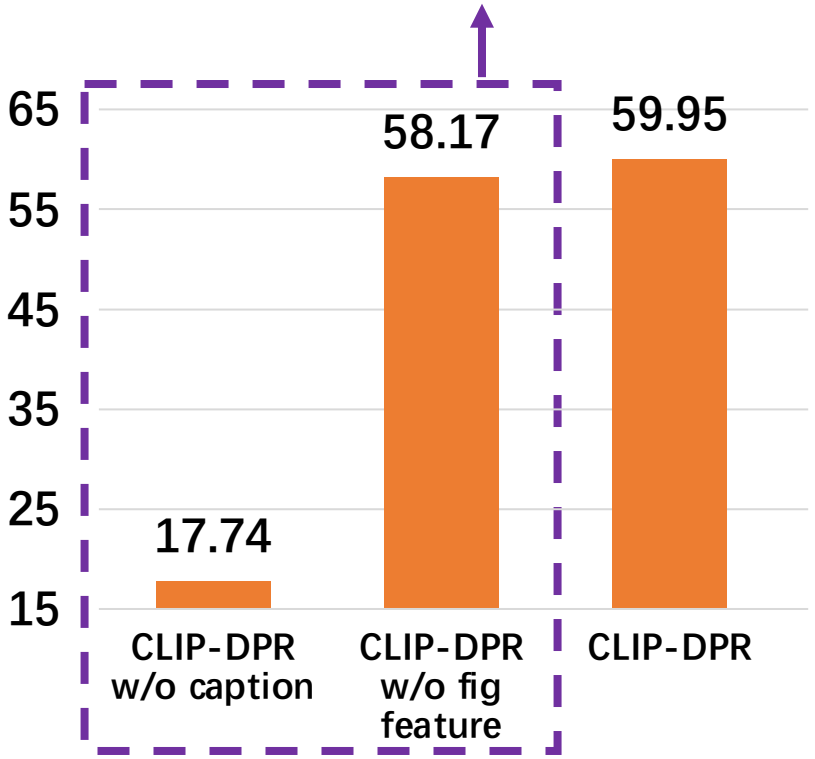


Liu et al., Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval. ICLR 2023.

面向多模态文档的稠密向量检索模型

- 我们对图像检索任务进行实验，以展示如何表示图像文档

图像标题在查询和图像文档之间的相关性建模中发挥着关键作用



问：明尼唐卡杜鹃花的花瓣是杯形的吗？

Figure Features:

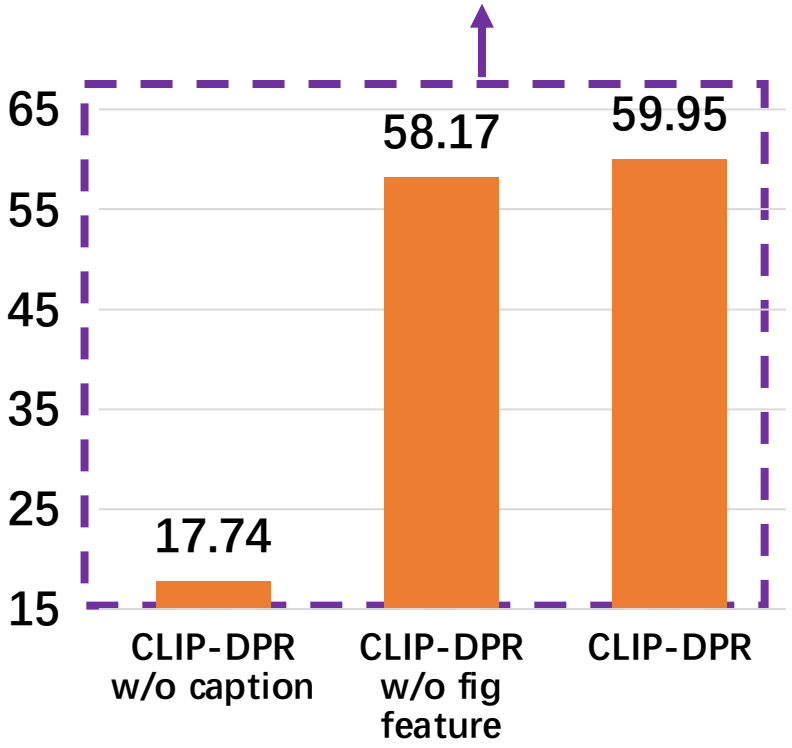
Image Captions:
Minnetonka Rhododendron flower along Tranquility Court …

Liu et al., Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval. ICLR 2023.

面向多模态文档的稠密向量检索模型

- 我们对图像检索任务进行实验，以展示如何表示图像文档

图形特征可以帮助更好地理解图像文档的语义



问：明尼唐卡杜鹃花的花瓣是杯形的吗？

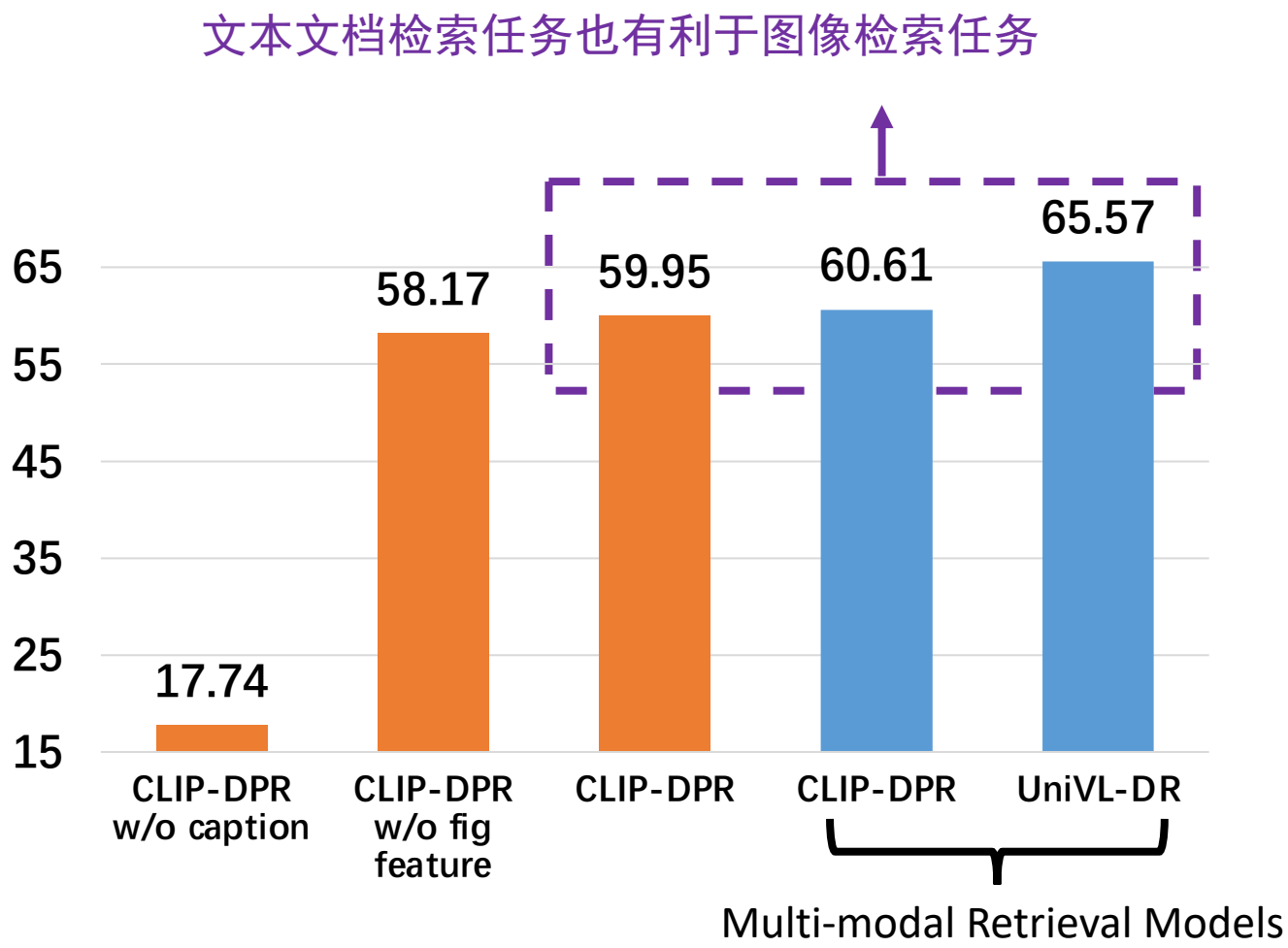
Figure Features:

Image Captions:
Minnetonka Rhododendron flower along Tranquility Court …

Liu et al., Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval. ICLR 2023.

面向多模态文档的稠密向量检索模型

- 然后我们进一步展示多模态检索模型的图像检索性能



问：明尼唐卡杜鹃花的花瓣是杯形的吗？

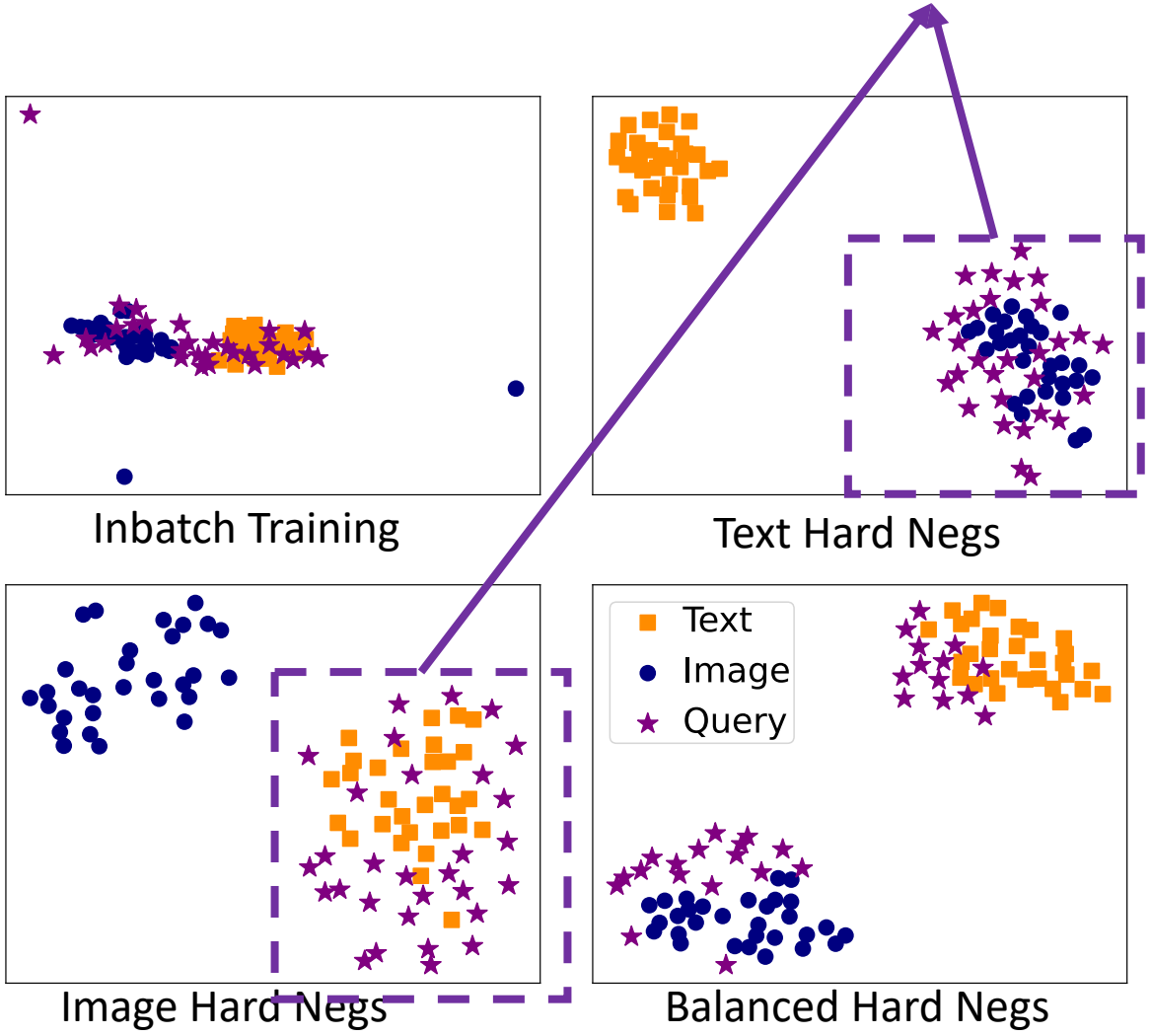
Figure Features:

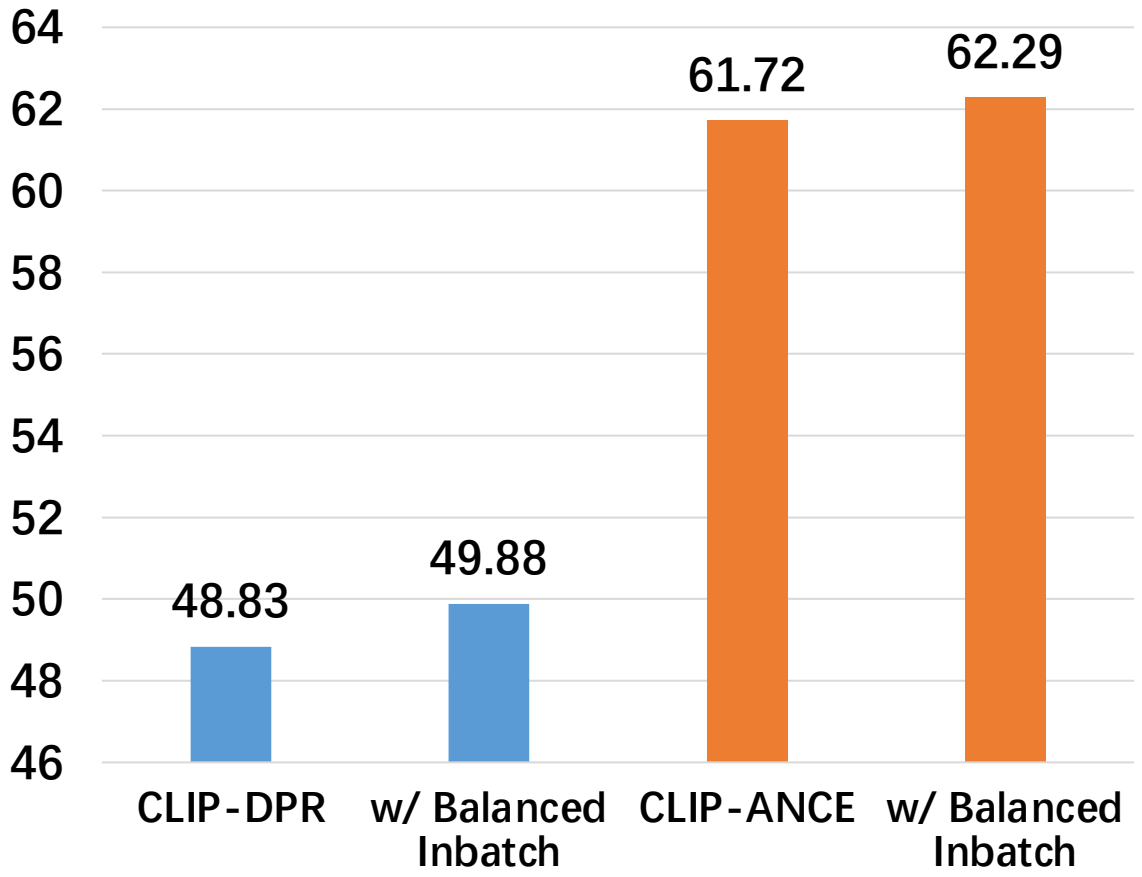
Image Captions:
Minnetonka Rhododendron flower along Tranquility Court …

Liu et al., Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval. ICLR 2023.

面向多模态文档的稠密向量检索模型

平衡难负例的模态可以减轻模型的模态偏好



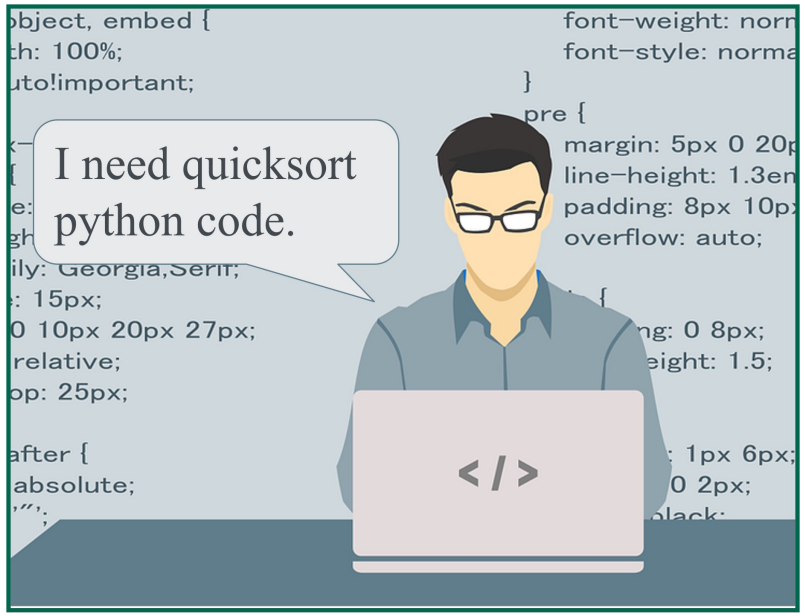


Liu et al., Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval. ICLR 2023.

面向多模态文档的稠密向量检索模型

- 针对结构化文本如何进行搜索？

Code Retrieval



def quicksort(arr): if len(arr) <= 1: return arr…

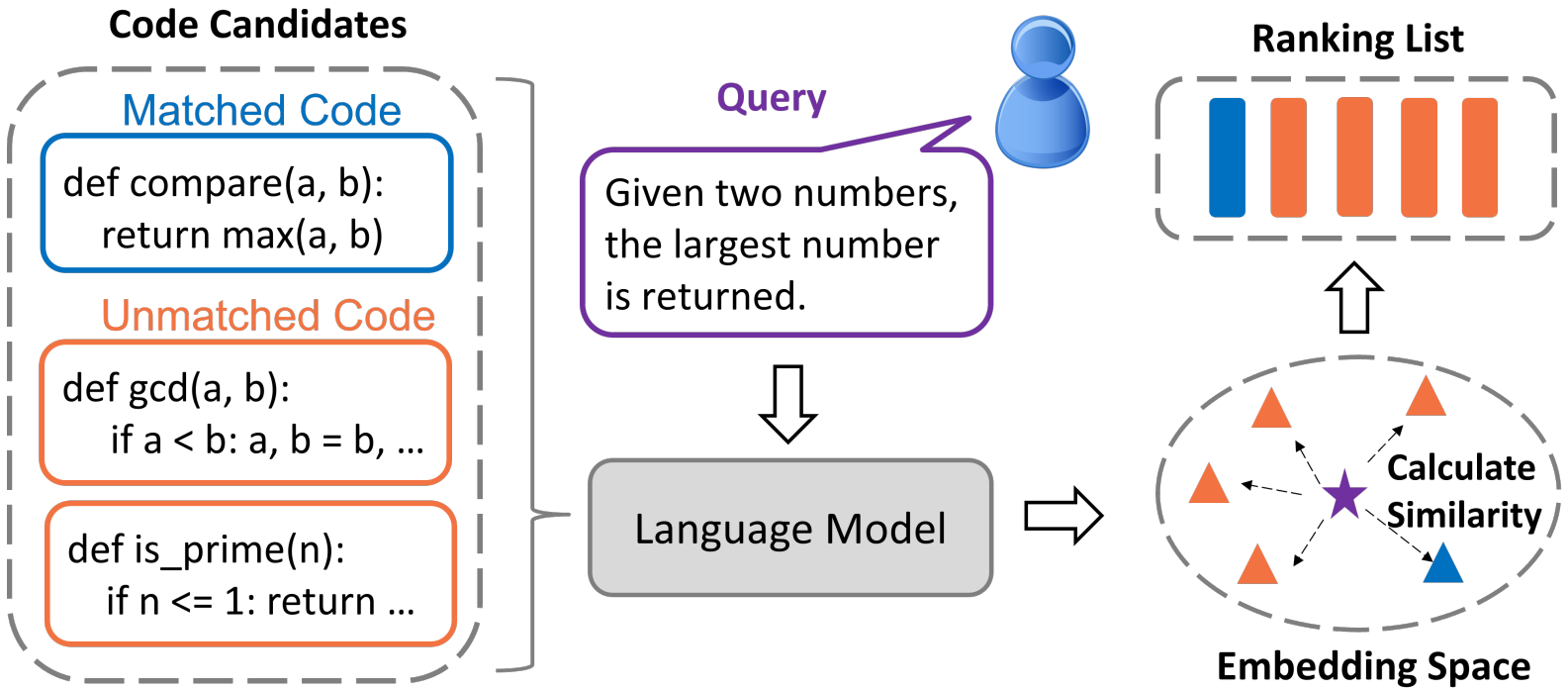
Product Retrieval

Business notebook for office. Brand: HP; Color: White...

Li et al. Structure-Aware Language Model Pretraining Can Benefit Dense Retrieval on Structured Data. ACL 2023.

面向多模态文档的稠密向量检索模型

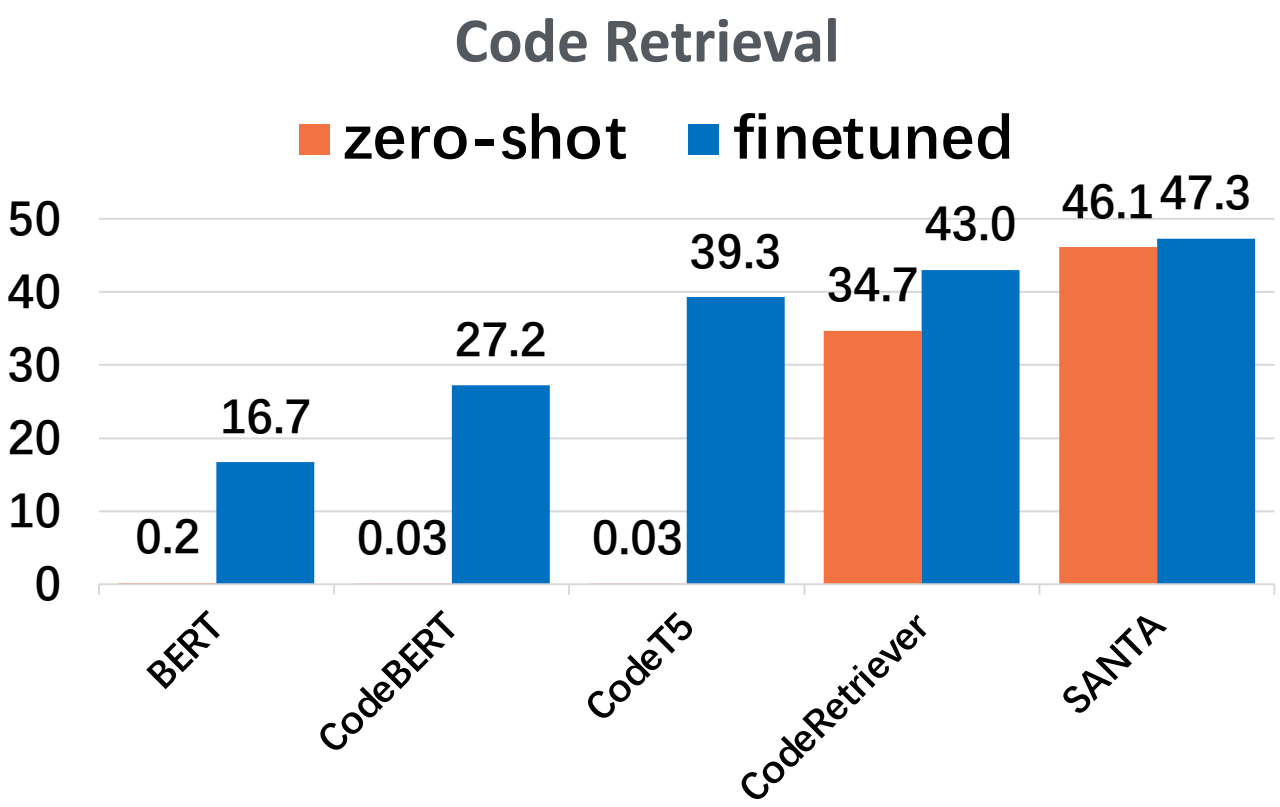
- 基于结构感知的稠密向量检索模型（Structure Aware Dense Retrieval, SANTA）
  - Structured Data Alignment (SDA)：通过将结构化数据与非结构化数据对齐来优化向量表示空间
  - Masked Entity Prediction (MEP)：恢复屏蔽实体来指导语言模型更好地理解结构化数据的语义

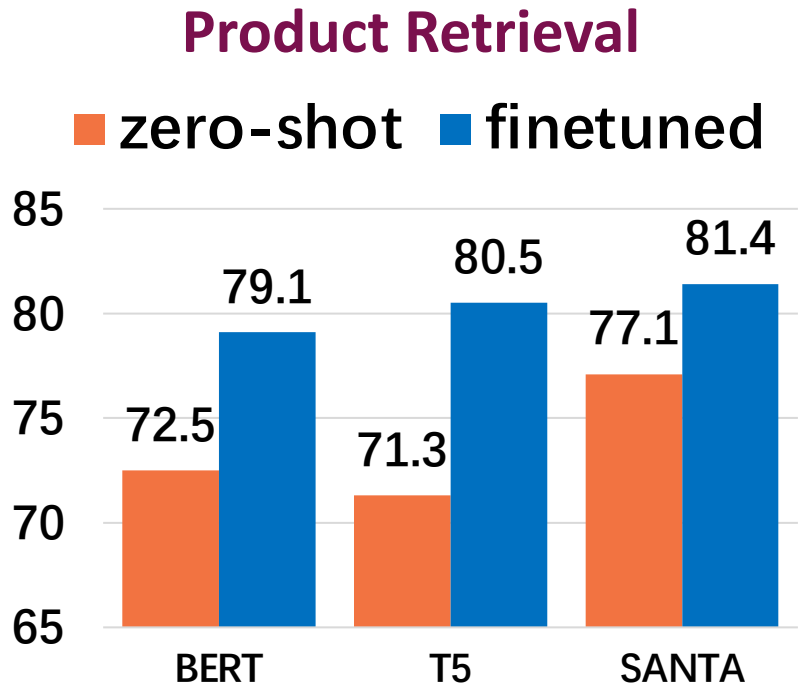


Li et al. Structure-Aware Language Model Pretraining Can Benefit Dense Retrieval on Structured Data. ACL 2023.

面向多模态文档的稠密向量检索模型

- 基于结构感知的稠密向量检索模型（Structure Aware Dense Retrieval, SANTA）
  - Structured Data Alignment (SDA)：通过将结构化数据与非结构化数据对齐来优化向量表示空间
  - Masked Entity Prediction (MEP)：恢复屏蔽实体来指导语言模型更好地理解结构化数据的语义

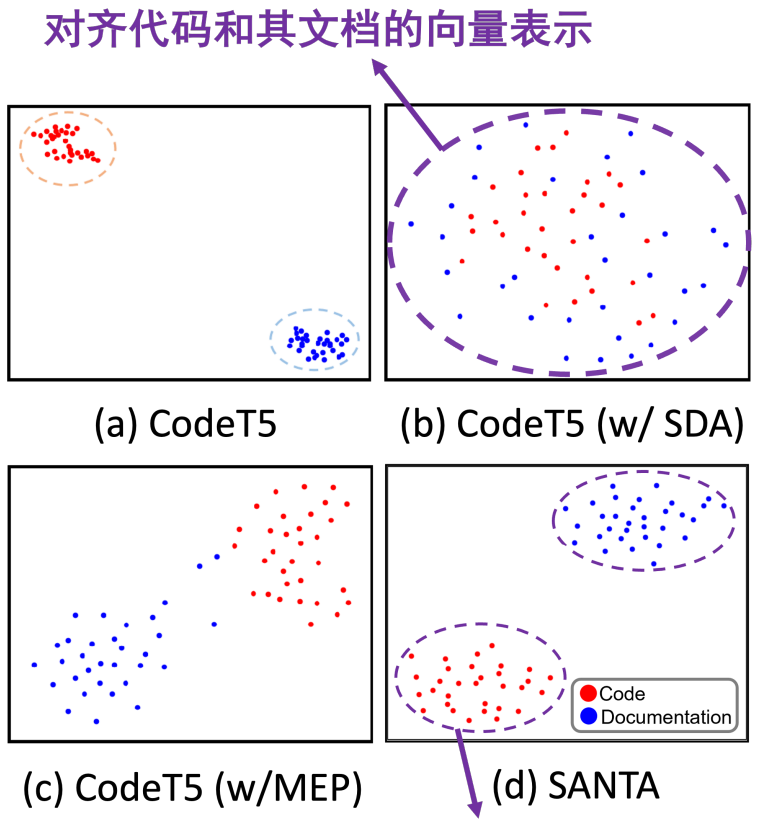


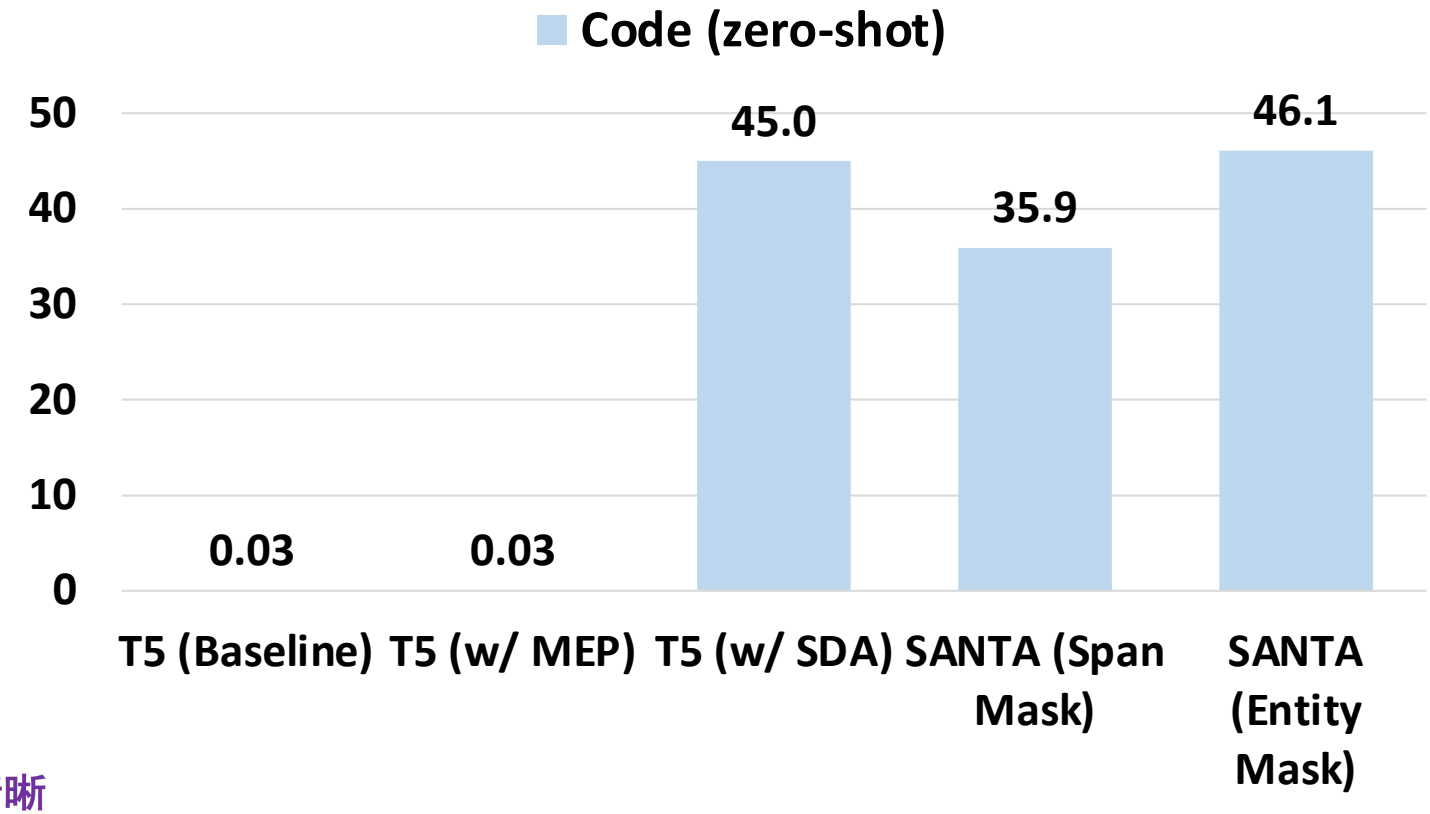


Li et al. Structure-Aware Language Model Pretraining Can Benefit Dense Retrieval on Structured Data. ACL 2023.

面向多模态文档的稠密向量检索模型

- 基于结构感知的稠密向量检索模型（Structure Aware Dense Retrieval, SANTA）
  - Structured Data Alignment (SDA)：通过将结构化数据与非结构化数据对齐来优化向量表示空间
  - Masked Entity Prediction (MEP)：恢复屏蔽实体来指导语言模型更好地理解结构化数据的语义





Li et al. Structure-Aware Language Model Pretraining Can Benefit Dense Retrieval on Structured Data. ACL 2023.

总结

- 当前预训练语言模型在实现信息检索过程中还需要大量的相关性信号
  - 面向Web领域的基于锚文本的数据增广方法
  - 面向通用领域的基于问题生成的相关性数据生成方法
  - 基于强化学习和元学习的相关性数据筛选方法

- 面向多模态数据的信息检索方法
  - 面向多模态数据的向量表示方法
  - 以语言为中心的多模态向量表示预训练方法

总结

- OpenMatch
  - 高效的向量检索建模
  - 最新的稠密检索建模工作（多模态、多任务）
  - 丰富的模型资源

Welcome to OpenMatch Community !

清华与微软团队联合提出基于领域知识迁移学习的神经信息检索

AI 科技评论
2020-09-02 14:45 | 国内顶尖人工智能媒体和产业服务平台，专注AI业界、学术...
关注

Microsoft / AI for Business

Microsoft Biomedical Search surfaces the most relevant results from more than 20 million documents from CORD-19, Microsoft Academic, PubMed and PubMed Central and relies on three interrelated AI efforts: *PubMedBERT, MetaAdaptRank* and *SaliencyMeasure* models.

PubMedBERT is a large-scale language model that was pre-trained on biomedical text rather than on a mix of general-domain language and domain-specific language. The model was pre-trained from scratch with 3 billion words specific to biomedicine. We have found that the model outperforms all prior language models on biomedical natural language processing applications.

MetaAdaptRank helps to accurately determine relevance by alleviating common problems associated with the ranking of research search results. Information retrieval systems often fail to identify all relevant information because queries and documents use different terms to describe the same concept. For example, this mismatch can happen when a searcher is unfamiliar with new terminology. MetaAdaptRank can learn the semantics of specialized domains to more accurately rank results even for topics or keywords for which information is scarce.