Robust Information Retrieval

WSDM 2025 tutorial

Yu-An Liu[a,b], Ruqing Zhang[a,b], Jiafeng Guo[a,b] and **Maarten de Rijke**[c]

https://wsdm2025-robust-information-retrieval.github.io/

March 10, 2025

01:30 – 05:00 PM

[a] Institute of Computing Technology, Chinese Academy of Sciences
[b] University of Chinese Academy of Sciences
[c] University of Amsterdam

Section 4:
Out-of-distribution robustness

SCAN
ME

Revisit the definition of out-of-distribution robustness

Ability of Neural IR models to maintain Top-K ranking performance when exposed to queries and documents that deviate from the distribution seen during training

Definition (Out-of-distribution robustness of information retrieval)

Given an IR model $f_{\mathcal{D}_{\text{train}}}$, an original dataset with training and test data, $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, drawn from the original distribution $\mathcal{G}$, along with a new test dataset $\tilde{\mathcal{D}}_{\text{test}}$ drawn from the new distribution $\tilde{\mathcal{G}}$, and an acceptable error threshold δ, for the top-K ranking result, if

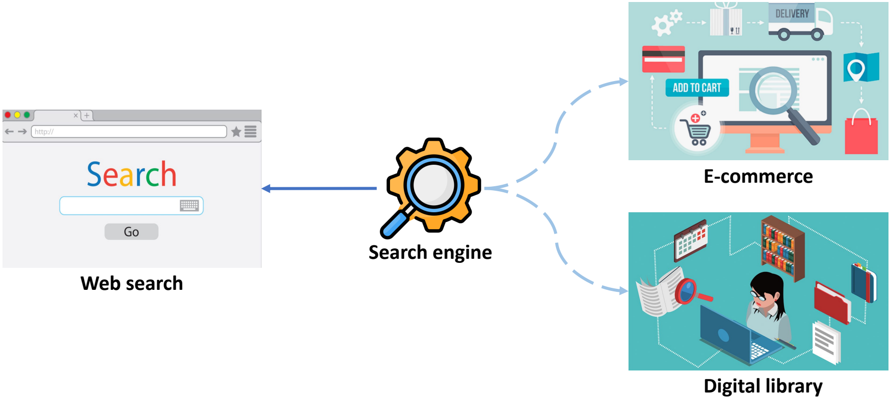
$$\left|\mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}}; \mathcal{D}_{\text{test}}, K\right) - \mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}}; \tilde{\mathcal{D}}_{\text{test}}, K\right)\right| \leq \delta \text{ where } \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}} \sim \mathcal{G}, \tilde{\mathcal{D}}_{\text{test}} \sim \tilde{\mathcal{G}},$$

the model f is considered δ-robust against out-of-distribution data for metric M.

Background: Migration scenarios for search engines

A good search engine can be migrated to various scenarios at a low cost. Difficulty:

- Documents from different domains
- Queries with different types



Web search

Search engine

E-commerce

Digital library

Background: Dynamic scenarios for search engines

A good search engine should keep up with the trends at a low cost. Difficulty:

- Documents on new hotspots
- Queries with new expressions

Image source: New content vs. old webpages

The above are uniformly described as out-of-distribution (OOD) scenarios

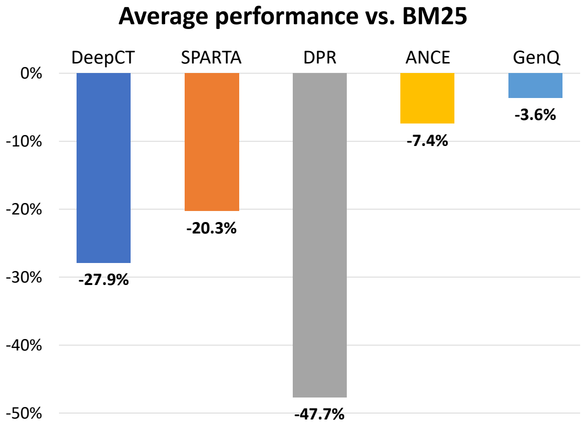
Dilemma: Neural IR models struggle with OOD scenarios

Without retraining, the performance of the neural IR model decreases significantly when faced with OOD data

Data source: [Thakur et al., 2021]

Dilemma: Neural IR models struggle with OOD scenarios

Without retraining, the performance of the neural IR model decreases significantly when faced with OOD data



- Dataset: BEIR
- Senario: OOD corpus
- Observations: The zero-shot performance of neural IR models is worse than traditional IR models

Data source: [Thakur et al., 2021]

A straightforward solution

“Let’s just retrain the neural IR models dynamically in response to OOD data. Problem solved.”

However, neural IR models are data-hungry

Training an effective neural IR model is very costly:

- **Quantity:** Large-scale queries and documents
- **Quality:** Relevance labels provided by experts

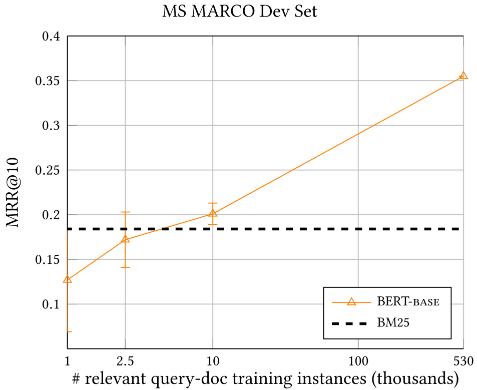
Data source: [Craswell et al., 2021, MacAvaney et al., 2021]

However, neural IR models are data-hungry

Training an effective neural IR model is very costly:

- **Quantity:** Large-scale queries and documents
- **Quality:** Relevance labels provided by experts

Dataset	Year	Query	Corpus
Robust04	2004	250	0.5M
MQ2007	2007	1.7k	25M
Clueweb09-B	2009	150	**50M**
MS MARCO	2017	**367k**	3.3M



Data source: [Craswell et al., 2021, MacAvaney et al., 2021]

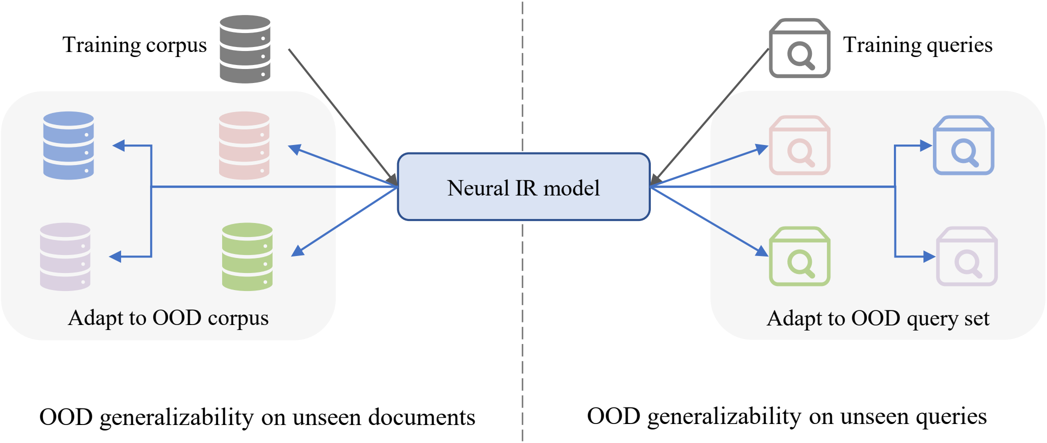
How can we flexibly enhance the OOD robustness of neural IR models?

How can we flexibly enhance the OOD robustness of neural IR models?

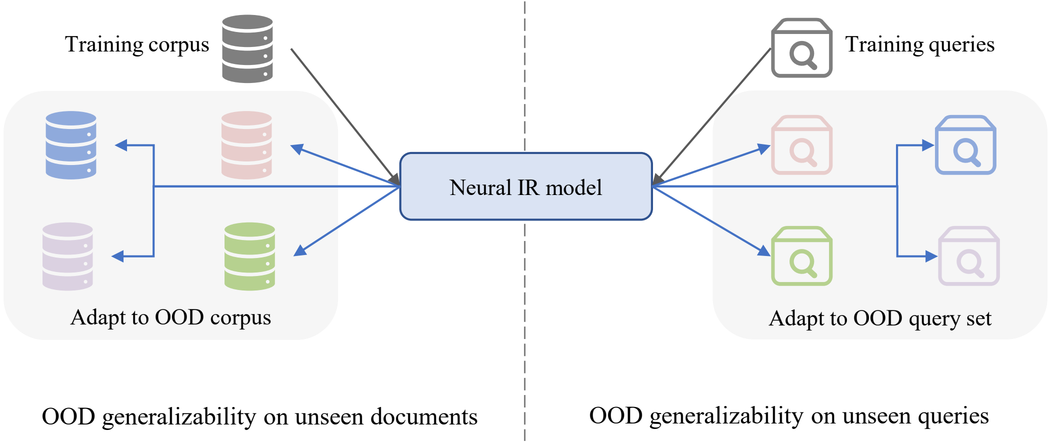
There are two perspectives...

Two perspectives of OOD robustness

The OOD robustness of neural IR models can be categorized into the generalizability on unseen documents and unseen queries



Two perspectives of OOD robustness



- **Unseen documents**: Corpus of new domains, corpus incrementation
- **Unseen querise**: Query variation (typos, etc.), new query types

Outline

We will introduce the OOD robustness through:

- **OOD generalizability on unseen documents**
  - **Benchmarks**
  - **Adaptation to new corpus**
  - **Updates to a corpus**

- **OOD generalizability on unseen queries**
  - **Benchmarks**
  - **Query variation**
  - **Unseen query type**

OOD generalizability on unseen documents

IR systems need to adapt to different environments and variations in the corpus

OOD generalizability on unseen documents

IR systems need to adapt to different environments and variations in the corpus

There are two scenarios:

- **Adaptation to new corpus:** Neural IR models trained on the original corpus are migrated to the new domain corpus

OOD generalizability on unseen documents

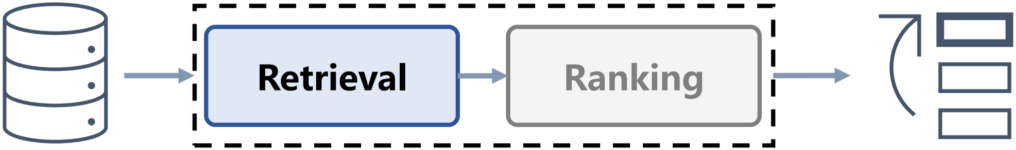
IR systems need to adapt to different environments and variations in the corpus

There are two scenarios:

- **Adaptation to new corpus:** Neural IR models trained on the original corpus are migrated to the new domain corpus
- **Updates to a corpus:** Neural IR models trained on the original corpus, adapted to the continuous growth of documents in the corpus

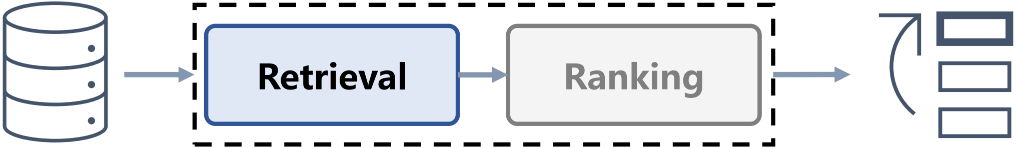
OOD generalizability on unseen documents

The above scenarios have a direct impact on the performance of the retrieval stage



OOD generalizability on unseen documents

The above scenarios have a direct impact on the performance of the retrieval stage



Existing work mainly focuses on neural retrieval models, i.e., dense retrieval models (DRMs) and generative retrieval models (GRMs)

OOD generalizability on unseen documents: Benchmarks

Adaptation to new corpus typically aggregates multiple existing domain IR datasets.

Image source: [Thakur et al., 2021]

OOD generalizability on unseen documents: Benchmarks

Adaptation to new corpus typically aggregates multiple existing domain IR datasets.

BEIR is the most typical, it includes 18 datasets from 9 different retrieval tasks, such as news retrieval, entity retrieval.

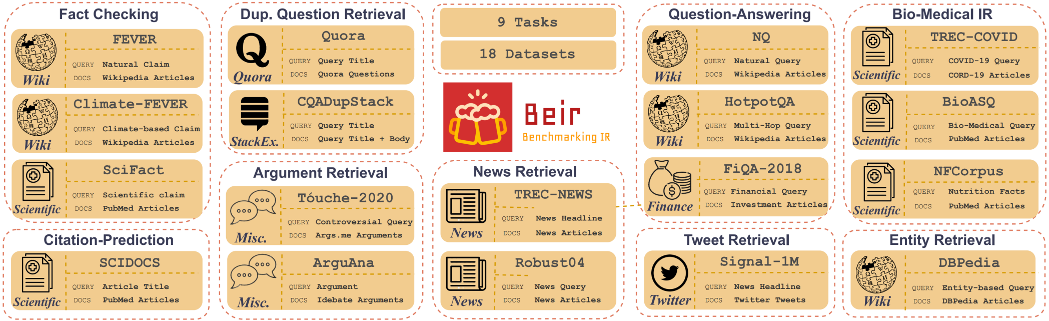


Image source: [Thakur et al., 2021]

OOD generalizability on unseen documents: Benchmarks

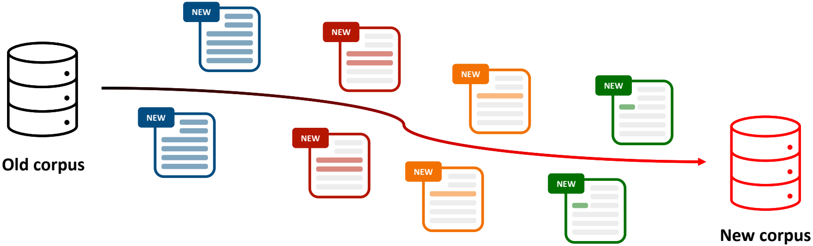
Updates to a corpus mainly slices or expands the existing dataset

OOD generalizability on unseen documents: Benchmarks

Updates to a corpus mainly slices or expands the existing dataset

For example, CDI-MS first randomly sampled 60% documents from the whole corpus as the base documents

Then, it randomly samples 10% documents from the remaining corpus as the new document set, and repeated 4 times



OOD generalizability on unseen documents: Benchmarks

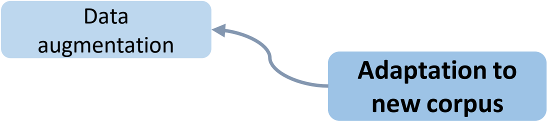
Type	**Dataset**	**#Retrieval task**	**#Corpus**
Adaptation to new corpus	BEIR [Thakur et al., 2021]	9	18

Type	**Dataset**	**#D**	**#Q**$_{\text{train}}$	**#Q**$_{\text{dev}}$	**#Q**$_{\text{eval}}$
Updates to original corpus	CDI-MS [Chen et al., 2023]	3.2M	370K	5,193	5,793
	CDI-NQ [Chen et al., 2023]	8.8M	500K	6,980	6,837
	LL-LoTTE [Cai et al., 2023]	5.5M	16K	8.5k	8.6k
	LL-MultiCPR [Cai et al., 2023]	3.0M	136K	15k	15k

Classification of adaptation to new corpus

Adaptation to new corpus

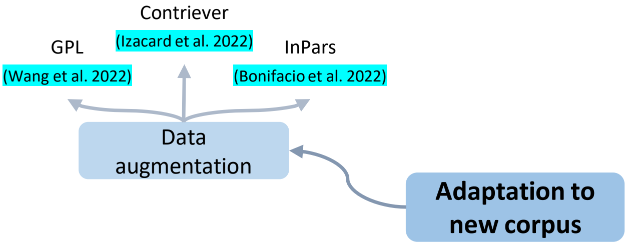
Classification of adaptation to new corpus



Data augmentation

Adaptation to new corpus

Classification of adaptation to new corpus



Contriever

(Izacard et al. 2022)

GPL

(Wang et al. 2022)

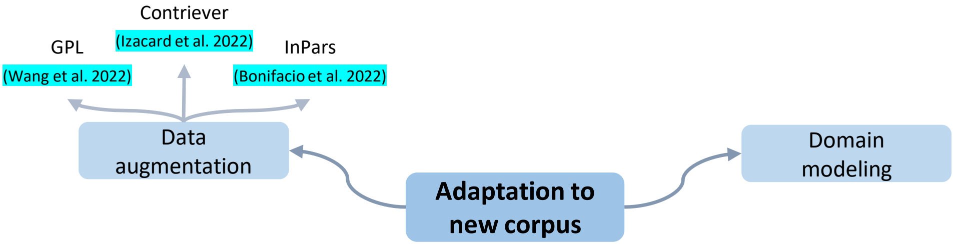
InPars

(Bonifacio et al. 2022)

Data augmentation

Adaptation to new corpus

Classification of adaptation to new corpus



Contriever

(Izacard et al. 2022)

GPL

(Wang et al. 2022)

InPars

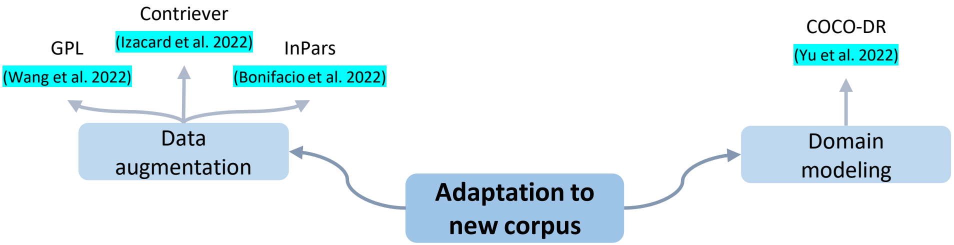
(Bonifacio et al. 2022)

Data augmentation

Domain modeling

Adaptation to new corpus

Classification of adaptation to new corpus



Adaptation to new corpus

Data augmentation

GPL (Wang et al. 2022)

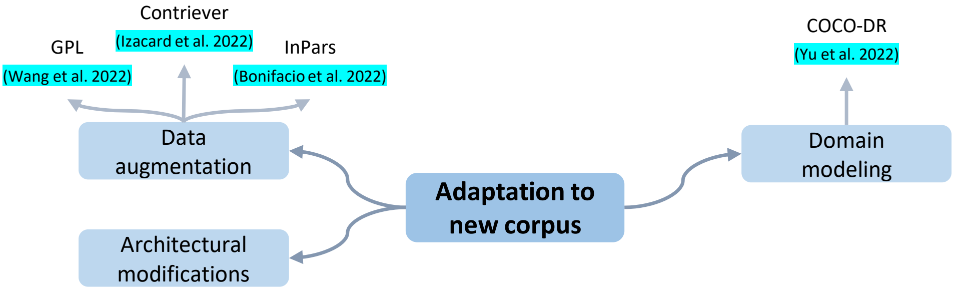
Contriever (Izacard et al. 2022)

InPars (Bonifacio et al. 2022)

Domain modeling

COCO-DR (Yu et al. 2022)

Classification of adaptation to new corpus



Adaptation to new corpus

Data augmentation

GPL (Wang et al. 2022)

Contriever (Izacard et al. 2022)

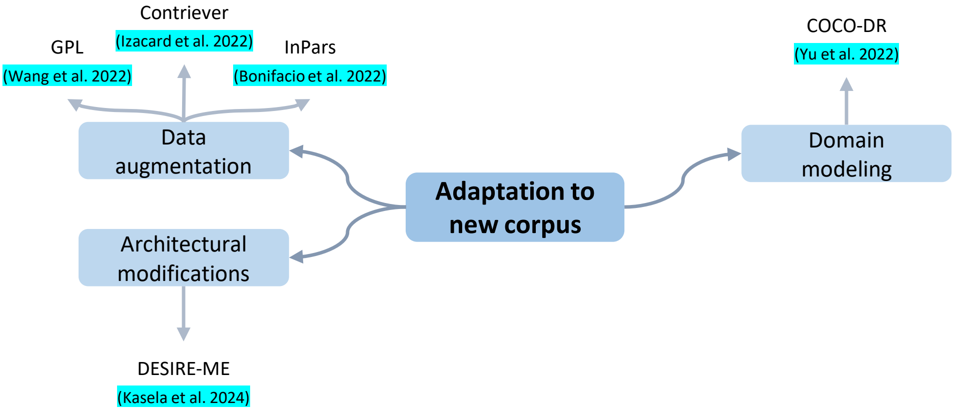
InPars (Bonifacio et al. 2022)

Architectural modifications

Domain modeling

COCO-DR (Yu et al. 2022)

Classification of adaptation to new corpus



Adaptation to new corpus

Data augmentation

GPL (Wang et al. 2022)

Contriever (Izacard et al. 2022)

InPars (Bonifacio et al. 2022)

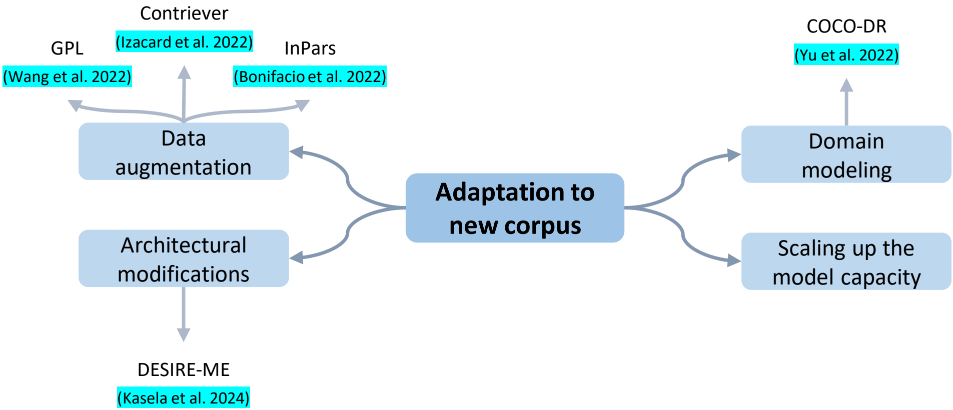
Architectural modifications

DESIRE-ME (Kasela et al. 2024)

Domain modeling

COCO-DR (Yu et al. 2022)

Classification of adaptation to new corpus


Adaptation to new corpus

Data augmentation

GPL (Wang et al. 2022)

Contriever (Izacard et al. 2022)

InPars (Bonifacio et al. 2022)

Architectural modifications

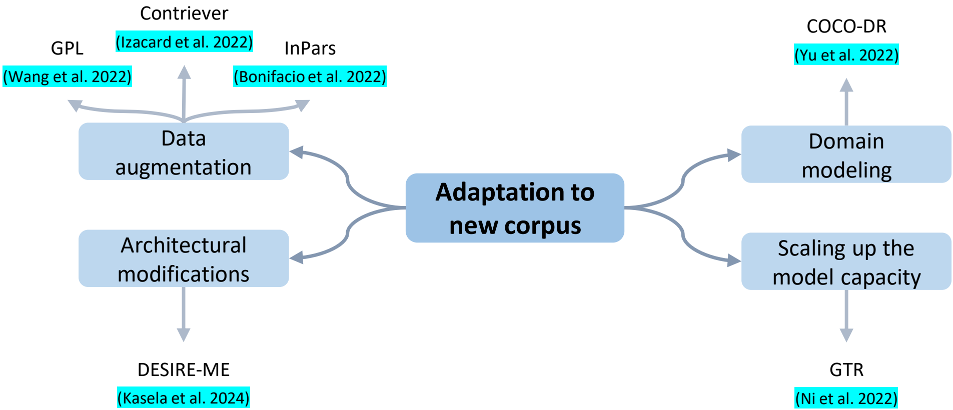
DESIRE-ME (Kasela et al. 2024)

Domain modeling

COCO-DR (Yu et al. 2022)

Scaling up the model capacity

Classification of adaptation to new corpus



Adaptation to new corpus

Data augmentation

GPL (Wang et al. 2022)

Contriever (Izacard et al. 2022)

InPars (Bonifacio et al. 2022)

Architectural modifications

DESIRE-ME (Kasela et al. 2024)

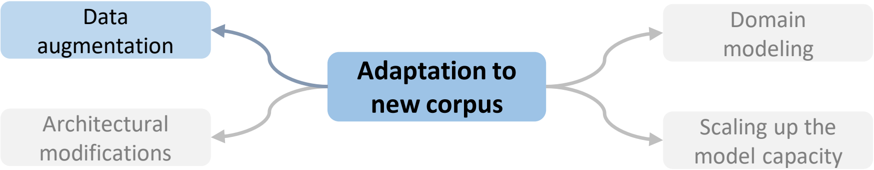
Domain modeling

COCO-DR (Yu et al. 2022)

Scaling up the model capacity

GTR (Ni et al. 2022)

Adaptation to new corpus: Data augmentation



Data augmentation

Architectural modifications

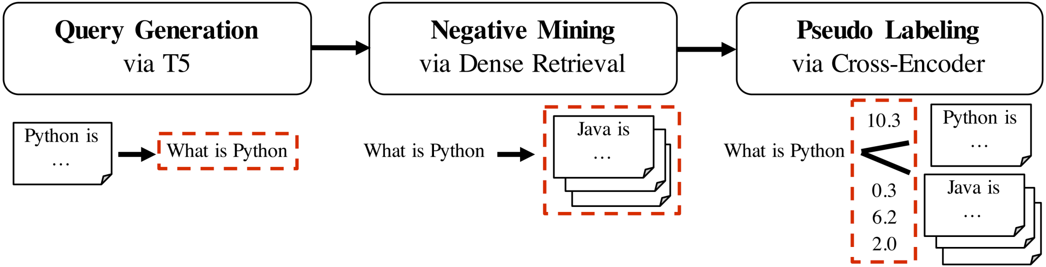
Adaptation to new corpus

Domain modeling

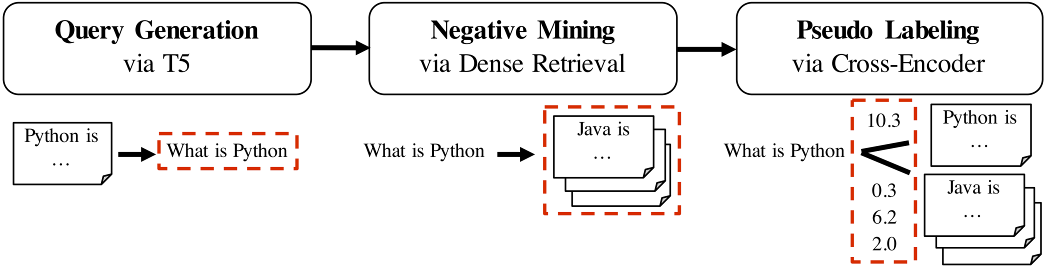
Scaling up the model capacity

Adaptation to new corpus: Data augmentation

Generative pseudo labeling (GPL) combines a query generator with pseudo labeling from a cross-encoder to generate additional training data [Wang et al., 2022]

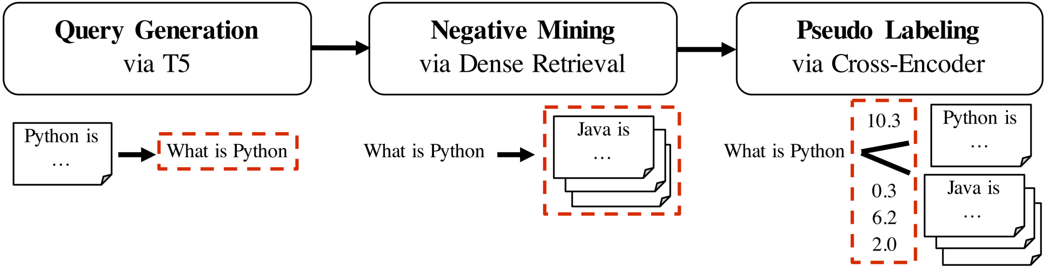


Adaptation to new corpus: Data augmentation



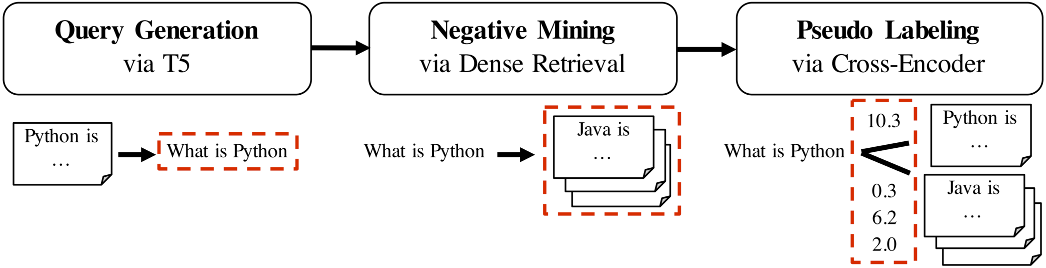
- Synthetic queries are generated for each passage from the target corpus

Adaptation to new corpus: Data augmentation



- Synthetic queries are generated for each passage from the target corpus
- The generated queries are used for mining negative passages

Adaptation to new corpus: Data augmentation



- Synthetic queries are generated for each passage from the target corpus
- The generated queries are used for mining negative passages
- The query-passage pairs are labeled by a cross-encoder and used to train the domain-adapted dense retriever

Data augmentation: GPL

😃 Straightforward: Access to large amounts of pseudo labeled data

Data augmentation: GPL

😀 Straightforward: Access to large amounts of pseudo labeled data

🤔 Unstable: Not all generated queries are of high quality

Data augmentation: GPL

- 😃 Straightforward: Access to large amounts of pseudo labeled data
- 🤔 Unstable: Not all generated queries are of high quality
- 🤔 Dependent: Over-reliance on cross-coder performance

Adaptation to new corpus: Data augmentation

Contriever explores the limits of contrastive learning as a way to pre-train in an unsupervised way a dense retriever [Izacard et al., 2021]

Adaptation to new corpus: Data augmentation

Contriever explores the limits of contrastive learning as a way to pre-train in an unsupervised way a dense retriever [Izacard et al., 2021]

- Build positive pairs from a single document through the inverse Cloze task

Adaptation to new corpus: Data augmentation

Contriever explores the limits of contrastive learning as a way to pre-train in an unsupervised way a dense retriever [Izacard et al., 2021]

- Build positive pairs from a single document through the inverse Cloze task
- Build a large set of negative pairs, including in-batch negatives and cross-batch negatives

Adaptation to new corpus: Data augmentation

Contriever explores the limits of contrastive learning as a way to pre-train in an unsupervised way a dense retriever [Izacard et al., 2021]

- Build positive pairs from a single document through the inverse Cloze task
- Build a large set of negative pairs, including in-batch negatives and cross-batch negatives
- Perform contrastive learning on the whole constructed training data

Data augmentation: Contriever

😃 Low data costs: Unsupervised construction of a large amount of pre-training data

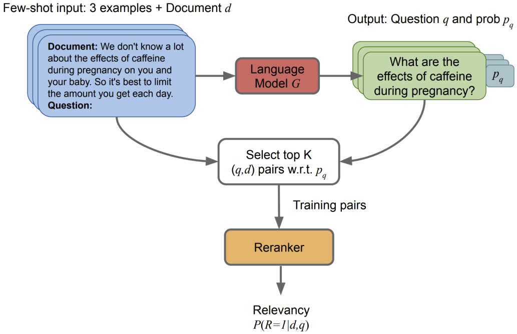
Data augmentation: Contriever

😃 Low data costs: Unsupervised construction of a large amount of pre-training data

🤔 High training costs: High cost of pre-training

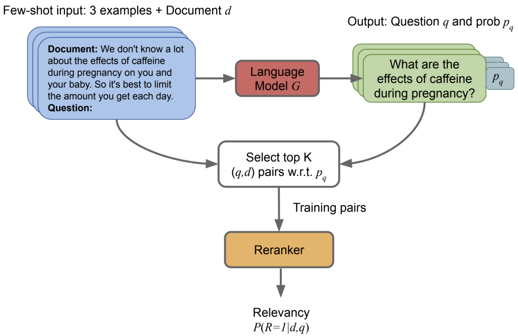
Adaptation to new corpus: Data augmentation

InPars harnesses the few-shot capabilities of large language models as synthetic data generators for IR task [Bonifacio et al., 2022]



Few-shot input: 3 examples + Document d
Document: We don't know a lot about the effects of caffeine during pregnancy on you and your baby. So it's best to limit the amount you get each day.
Question:
Language Model G
Output: Question q and prob p_q
What are the effects of caffeine during pregnancy?
p_q
Select top K (q,d) pairs w.r.t. p_q
Training pairs
Reranker
Relevancy
$P(R=1|d,q)$

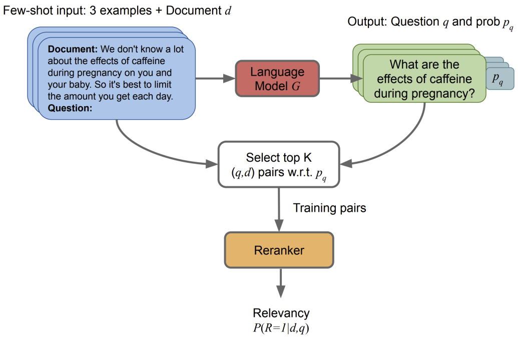
Adaptation to new corpus: Data augmentation



Few-shot input: 3 examples + Document d
Document: We don't know a lot about the effects of caffeine during pregnancy on you and your baby. So it's best to limit the amount you get each day.
Question:
Language Model G
Output: Question q and prob p_q
What are the effects of caffeine during pregnancy?
p_q
Select top K (q,d) pairs w.r.t. p_q
Training pairs
Reranker
Relevancy
$P(R=1|d,q)$

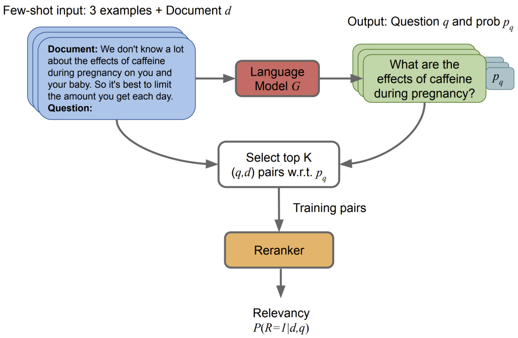
- For a document, 3 sets of q-d pairs are constructed as the instruction

Adaptation to new corpus: Data augmentation


Few-shot input: 3 examples + Document d
Output: Question q and prob p_q
Document: We don't know a lot about the effects of caffeine during pregnancy on you and your baby. So it's best to limit the amount you get each day.
Question:
Language Model G
What are the effects of caffeine during pregnancy?
p_q
Select top K (q,d) pairs w.r.t. p_q
Training pairs
Reranker
Relevancy
$P(R=1|d,q)$

- For a document, 3 sets of q-d pairs are constructed as the instruction
- Generate query with LLM and get the corresponding generation probability

Adaptation to new corpus: Data augmentation


Few-shot input: 3 examples + Document d
Document: We don't know a lot about the effects of caffeine during pregnancy on you and your baby. So it's best to limit the amount you get each day.
Question:
Language Model G
Output: Question q and prob p_q
What are the effects of caffeine during pregnancy?
p_q
Select top K (q,d) pairs w.r.t. p_q
Training pairs
Reranker
Relevancy
$P(R=1|d,q)$

- For a document, 3 sets of q-d pairs are constructed as the instruction
- Generate query with LLM and get the corresponding generation probability
- Based on this, the corresponding query is generated for each randomly sampled document, constituting a positive sample for training

Data augmentation: InPars

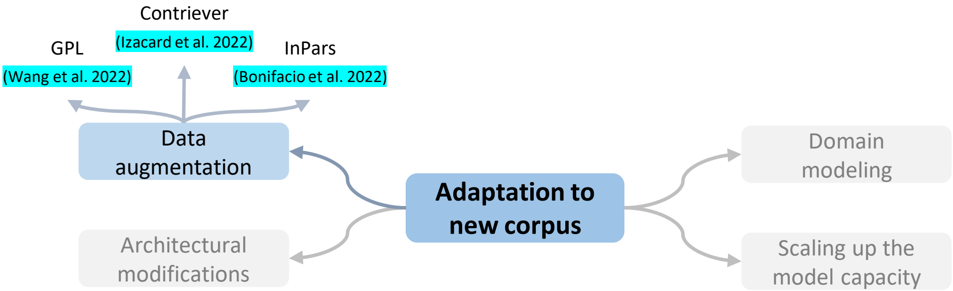
😃 Effective: Constructing positive samples using LLMs

Data augmentation: InPars

😃 Effective: Constructing positive samples using LLMs

🤔 Risky: Low-quality generated queries may occur

Review data augmentation



Contriever

(Izacard et al. 2022)

GPL

(Wang et al. 2022)

InPars

(Bonifacio et al. 2022)

Data augmentation

Adaptation to new corpus

Domain modeling

Architectural modifications

Scaling up the model capacity

Review data augmentation

😃 Effective: Simple way to improve model training

Review data augmentation

- 😀 Effective: Simple way to improve model training
- 🤔 Diverse: There are various ways to synthesize data

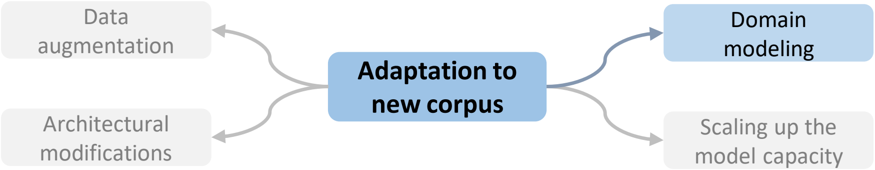
Review data augmentation

😃 Effective: Simple way to improve model training

🤔 Diverse: There are various ways to synthesize data

😟 Risky: Low-quality data is hard to avoid

Adaptation to new corpus: Domain modeling



Adaptation to new corpus: Domain modeling

COCO-DR uses implicit distributionally robust optimization (iDRO) to reweight samples from different source query clusters for improving model robustness over rare queries during fine-tuning [Yu et al., 2022]

A model trained to be more robust on the source domain is likely to better generalize to unseen data

Adaptation to new corpus: Domain modeling

COCO-DR uses implicit distributionally robust optimization (iDRO) to reweight samples from different source query clusters for improving model robustness over rare queries during fine-tuning [Yu et al., 2022]

A model trained to be more robust on the source domain is likely to better generalize to unseen data

- Cluster source queries using K-Means and then optimize the iDRO loss

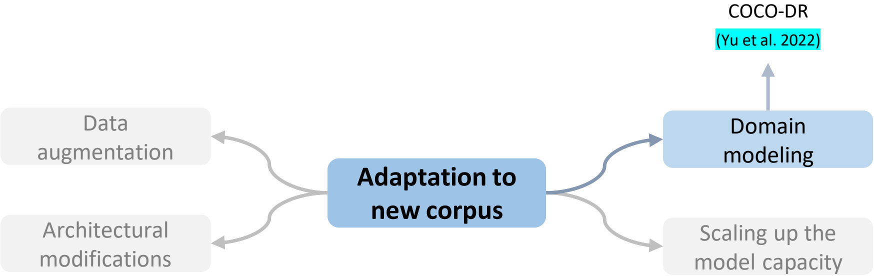
Adaptation to new corpus: Domain modeling

COCO-DR uses implicit distributionally robust optimization (iDRO) to reweight samples from different source query clusters for improving model robustness over rare queries during fine-tuning [Yu et al., 2022]

A model trained to be more robust on the source domain is likely to better generalize to unseen data

- Cluster source queries using K-Means and then optimize the iDRO loss
- Dynamic weight of each cluster during fine-tuning

Review domain modeling


COCO-DR

(Yu et al. 2022)

Data augmentation

Architectural modifications

Adaptation to new corpus

Domain modeling

Scaling up the model capacity

Review domain modeling

😃 Reliable: Theoretically guaranteed generalization from existing domains to unseen domains

Review domain modeling

😃 Reliable: Theoretically guaranteed generalization from existing domains to unseen domains

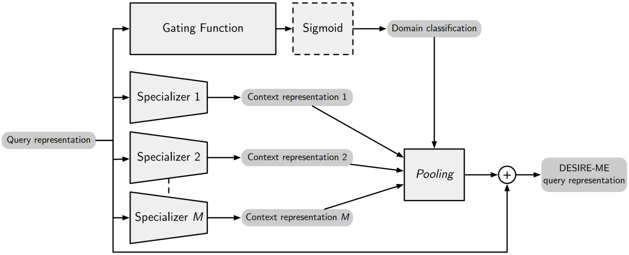
🤔 Complex: Complexity of realization and training process

Adaptation to new corpus: architectural modifications

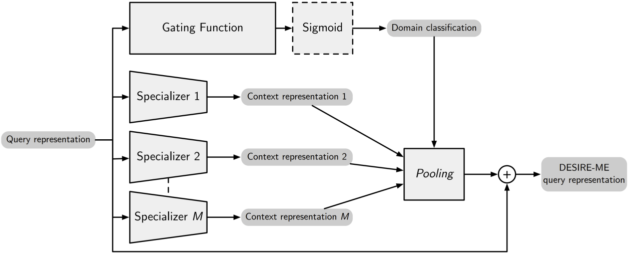
Data augmentation
Architectural modifications
Adaptation to new corpus
Domain modeling
Scaling up the model capacity

Adaptation to new corpus: architectural modifications

DESIRE-ME uses the mixture-of-experts framework to combine multiple specialized neural models [Kasela et al., 2024]

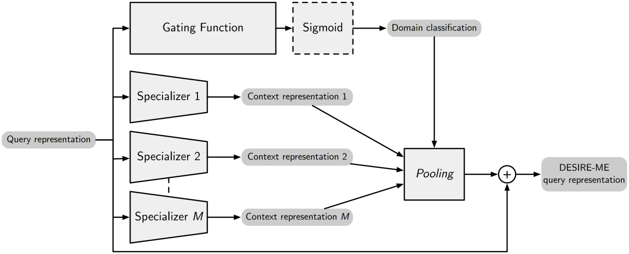


Adaptation to new corpus: architectural modifications



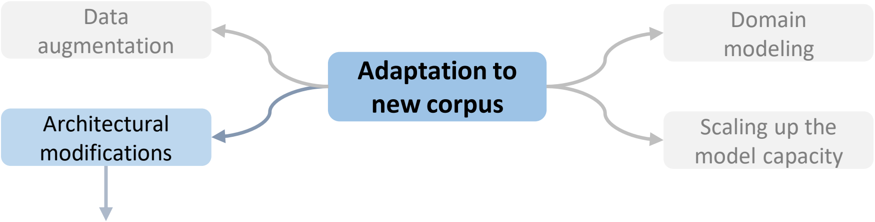
- **Specializers** focus on tuning query representation for the corresponding domain

Adaptation to new corpus: architectural modifications



- **Specializers** focus on tuning query representation for the corresponding domain
- **Pooling module** merges the domain context representations computed by the specializers on the basis of the domain likelihood estimated by the gating function

Review architectural modifications



DESIRE-ME

(Kasela et al. 2024)

Review architectural modifications

😀 Explainable: Explicit modeling domain information

Review architectural modifications

😃 Explainable: Explicit modeling domain information

😟 Restricted: Assumption of having query domain information

Adaptation to new corpus: Scaling up the model capacity

Adaptation to new corpus

Data augmentation

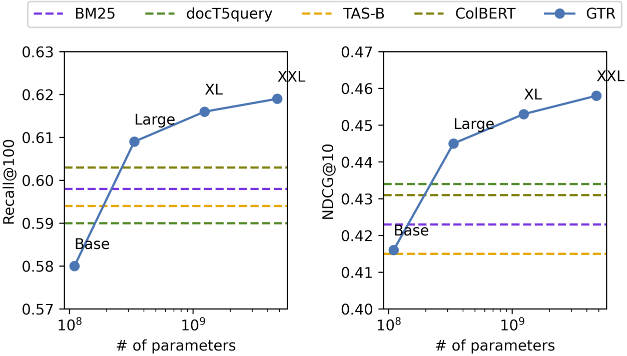
Architectural modifications

Domain modeling

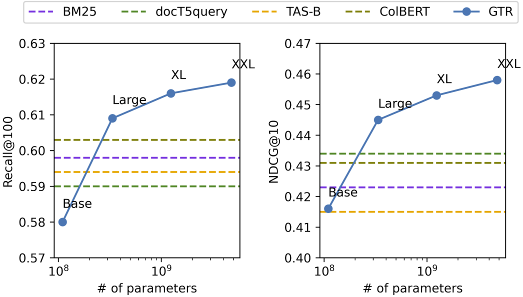
Scaling up the model capacity

Adaptation to new corpus: Scaling up the model capacity

GTR scales up the dual encoder model size while keeping the bottleneck embedding size fixed [Ni et al., 2022]

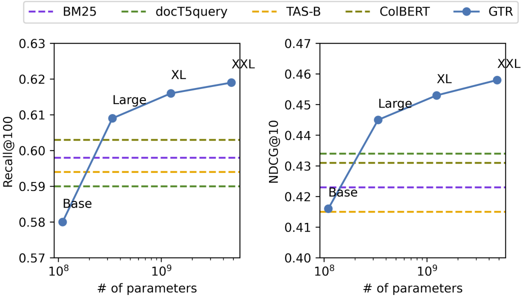


Adaptation to new corpus: Scaling up the model capacity



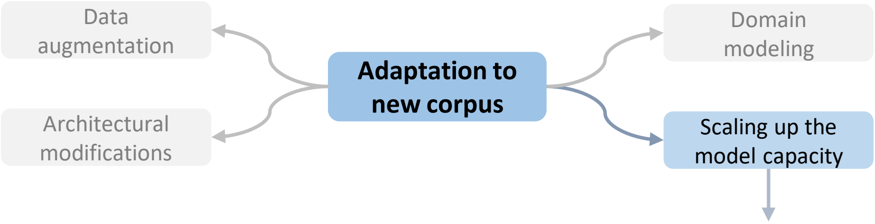
- For pre-training, the dual encoder is initialized from the T5 models and train on question-answer pairs collected from the Web

Adaptation to new corpus: Scaling up the model capacity



- For pre-training, the dual encoder is initialized from the T5 models and train on question-answer pairs collected from the Web
- For fine-tuning, the aim is to adapt the model to retrieval using a high-quality search corpus

Review scaling up the model capacity



Data augmentation

Architectural modifications

Adaptation to new corpus

Domain modeling

Scaling up the model capacity

GTR

(Ni et al. 2022)

Review scaling up the model capacity

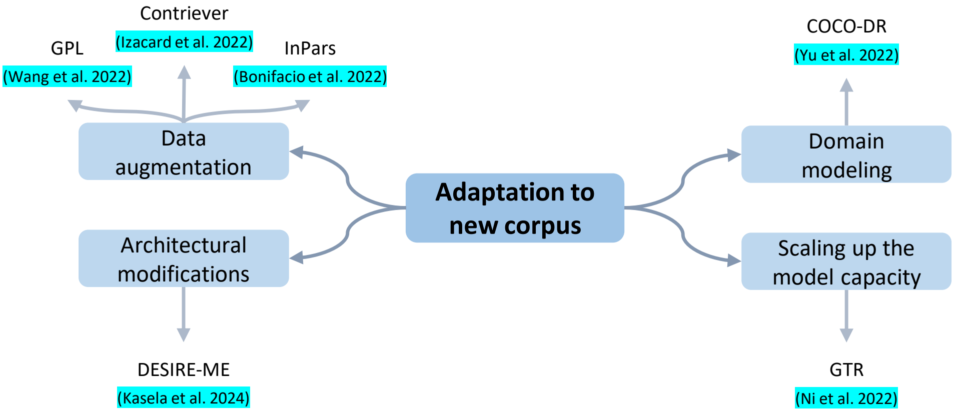
😃 Simple: Straightforward to improve OOD robustness

Review scaling up the model capacity

😀 Simple: Straightforward to improve OOD robustness

😦 Costly: High training overhead and requires more training data than before

Adaptation to new corpus



Adaptation to new corpus

Data augmentation

GPL (Wang et al. 2022)

Contriever (Izacard et al. 2022)

InPars (Bonifacio et al. 2022)

Architectural modifications

DESIRE-ME (Kasela et al. 2024)

Domain modeling

COCO-DR (Yu et al. 2022)

Scaling up the model capacity

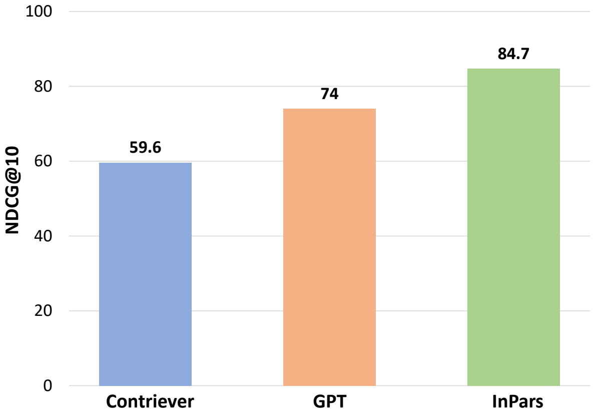
GTR (Ni et al. 2022)

Evaluation

Key idea: Evaluate the average ranking performance across different domains

- **NDCG** evaluates the quality of ranking results by measuring the gain of a document based on its position in the ranked list
- **MRR** evaluates the performance of a ranking result by calculating the average of the reciprocal ranks of the first relevant document answer
- **HIT** evaluates the proportion of times a relevant document is found within a set of top-N ranking results
- **AP** evaluates the average performance of the ranking performance metrics, overall new domains

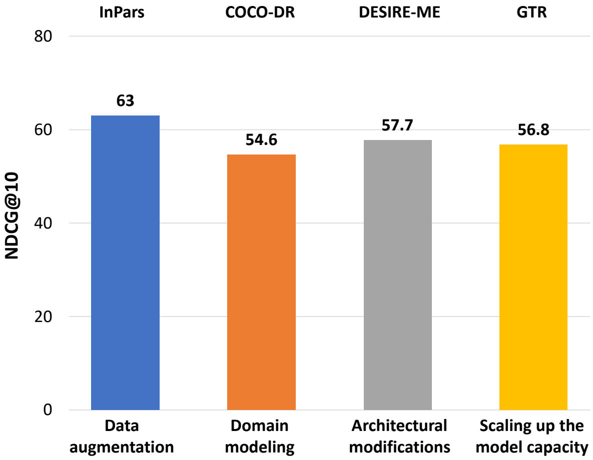
Comparison between data augmentation methods



Data source: [Bonifacio et al., 2022, Izacard et al., 2021]

- Original corpus: MS MARCO
- New corpus: TREC-COVID
- Observations: Effectiveness of relevance supervised signals: heuristic $<$ cross-coder judgment $<$ LLMs generation

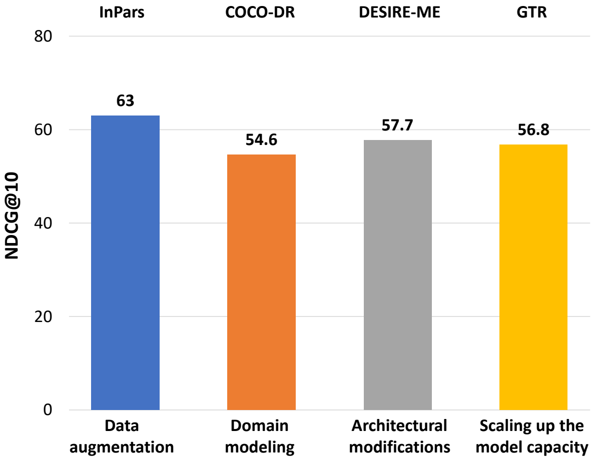
Comparison between adaptation to new corpus methods



- Original corpus: MS MARCO
- New corpus: NQ
- Observations: With the help of LLMs, data augmentation becomes the most effective method

Data source: [Bonifacio et al., 2022, Kasela et al., 2024, Ni et al., 2022]

Comparison between adaptation to new corpus methods



Data source: [Bonifacio et al., 2022, Kasela et al., 2024, Ni et al., 2022]

- Original corpus: MS MARCO
- New corpus: NQ
- Observations: Improvements from increasing model capacity or extending the model structure may be limited

Takeaway

For adaptation to new corpus:

Takeaway

For adaptation to new corpus:

- High-quality data and an appropriate modeling approach are key to the problem

Takeaway

For adaptation to new corpus:

- High-quality data and an appropriate modeling approach are key to the problem
- LLMs can play a variety of roles in it

Takeaway

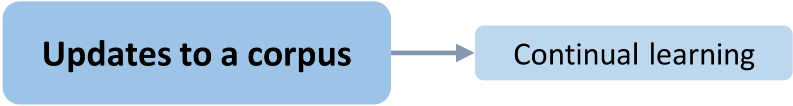
For adaptation to new corpus:

- High-quality data and an appropriate modeling approach are key to the problem
- LLMs can play a variety of roles in it
- There is a trade-off between efficiency and effectiveness

Classification of updates to a corpus

Updates to a corpus

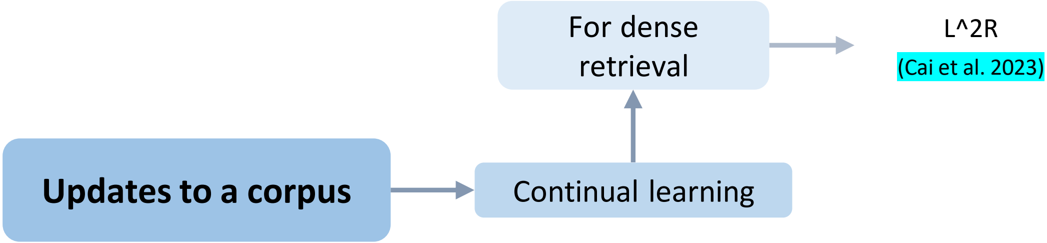
Classification of updates to a corpus



Updates to a corpus

Continual learning

Classification of updates to a corpus



Updates to a corpus

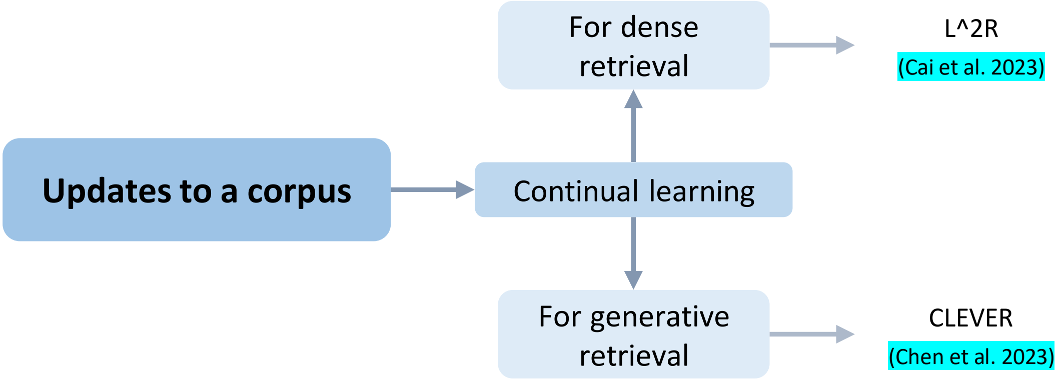
Continual learning

For dense retrieval

L^2R

(Cai et al. 2023)

Classification of updates to a corpus



Updates to a corpus

Continual learning

For dense retrieval

L^2R

(Cai et al. 2023)

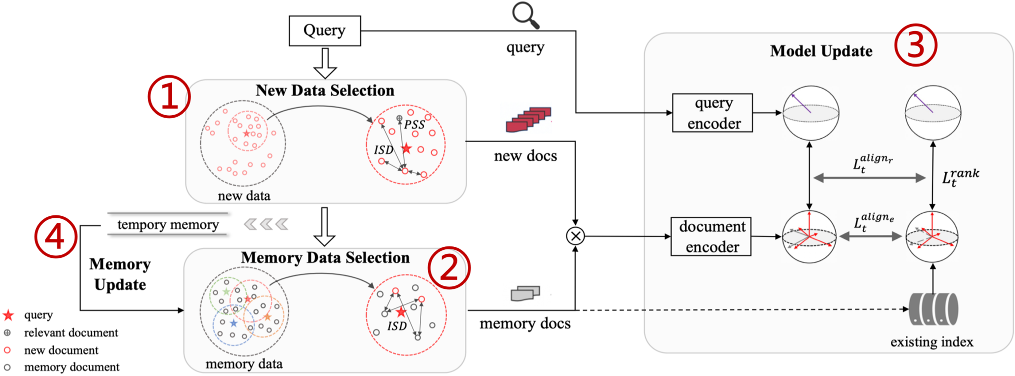
For generative retrieval

CLEVER

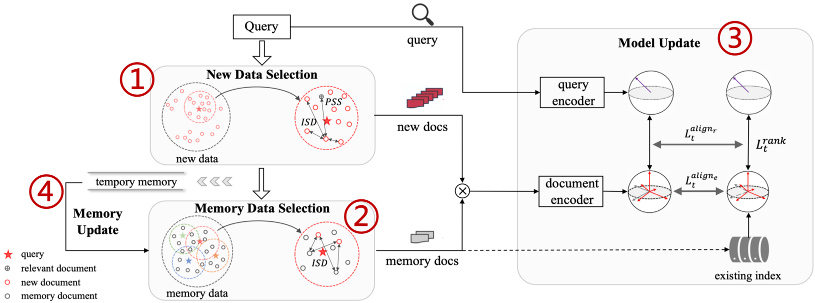
(Chen et al. 2023)

Updates to a corpus: Dense retrieval

L^{2}R employs a replay mechanism that maintains an external memory for storing a subset of historical documents for replay [Cai et al., 2023]



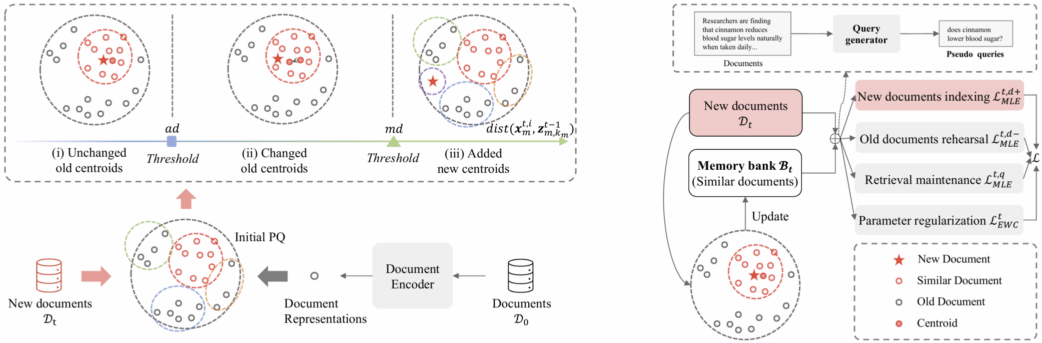
Updates to a corpus: Dense retrieval



- Expanding new knowledge
- Resolving catastrophic forgetting
- Updating the model based on selected new-old samples
- Updating memory based on new data

Updates to a corpus: Generative retrieval

CLEVER incrementally indexes new documents while supporting the ability to query both newly encountered documents and previously learned documents [Chen et al., 2023]

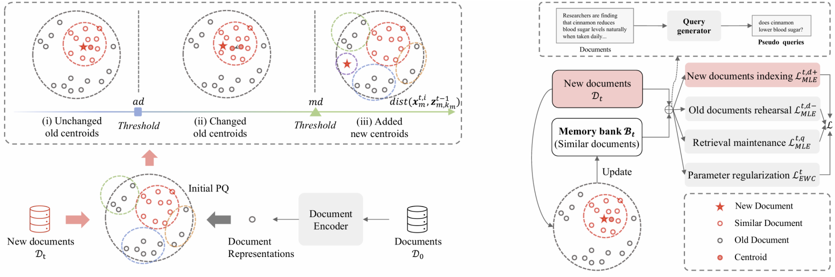


ad
md
$dist(\boldsymbol{x}_m^{t,i}, \boldsymbol{z}_{m,k_m}^{t-1})$
(i) Unchanged old centroids
Threshold
(ii) Changed old centroids
Threshold
(iii) Added new centroids
Initial PQ
New documents $\mathcal{D}_t$
Document Representations
Document Encoder
Documents $\mathcal{D}_0$
Researchers are finding that cinnamon reduces blood sugar levels naturally when taken daily...
Documents
Query generator
does cinnamon lower blood sugar?
Pseudo queries
New documents $\mathcal{D}_t$
Memory bank $\mathcal{B}_t$ (Similar documents)
Update
New documents indexing $\mathcal{L}_{MLE}^{t,d+}$
Old documents rehearsal $\mathcal{L}_{MLE}^{t,d-}$
Retrieval maintenance $\mathcal{L}_{MLE}^{t,q}$
Parameter regularization $\mathcal{L}_{EWC}^{t}$
$\mathcal{L}$
New Document
Similar Document
Old Document
Centroid

(a) Incremental product quantization (IPQ)

(b) Overall training objective

Updates to a corpus: Generative retrieval

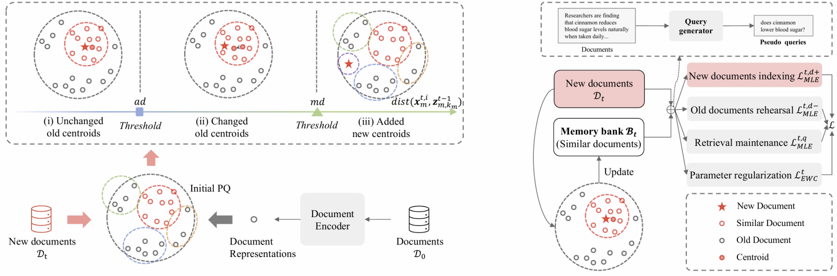

ad
md
$dist(x_m^{t,i}, z_{m,k_m}^{t-1})$
(i) Unchanged old centroids
Threshold
(ii) Changed old centroids
Threshold
(iii) Added new centroids
Initial PQ
New documents $\mathcal{D}_t$
Document Representations
Document Encoder
Documents $\mathcal{D}_0$
Researchers are finding that cinnamon reduces blood sugar levels naturally when taken daily...
Documents
Query generator
does cinnamon lower blood sugar?
Pseudo queries
New documents $\mathcal{D}_t$
Memory bank $\mathcal{B}_t$ (Similar documents)
Update
New documents indexing $\mathcal{L}_{MLE}^{t,d+}$
Old documents rehearsal $\mathcal{L}_{MLE}^{t,d-}$
Retrieval maintenance $\mathcal{L}_{MLE}^{t,q}$
Parameter regularization $\mathcal{L}_{EWC}^{t}$
$\mathcal{L}$
New Document
Similar Document
Old Document
Centroid

(a) Incremental product quantization (IPQ)

(b) Overall training objective

- Encoding new documents into docids by updating a subset of quantization centroids

Updates to a corpus: Generative retrieval

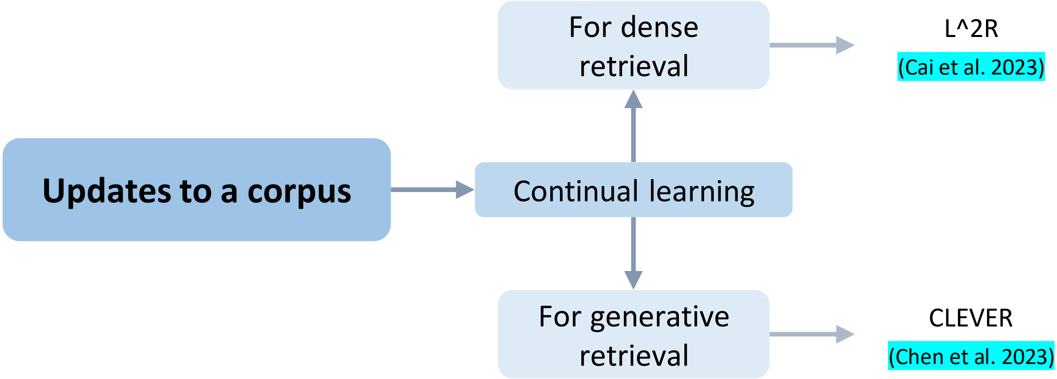

ad
md
$dist(x_m^{t,i}, z_{m,k_m}^{t-1})$
(i) Unchanged old centroids
Threshold
(ii) Changed old centroids
Threshold
(iii) Added new centroids
Initial PQ
New documents $\mathcal{D}_t$
Document Representations
Document Encoder
Documents $\mathcal{D}_0$
Researchers are finding that cinnamon reduces blood sugar levels naturally when taken daily...
Documents
Query generator
does cinnamon lower blood sugar?
Pseudo queries
New documents $\mathcal{D}_t$
Memory bank $\mathcal{B}_t$ (Similar documents)
Update
New documents indexing $\mathcal{L}_{MLE}^{t,d+}$
Old documents rehearsal $\mathcal{L}_{MLE}^{t,d-}$
Retrieval maintenance $\mathcal{L}_{MLE}^{t,q}$
Parameter regularization $\mathcal{L}_{EWC}^{t}$
$\mathcal{L}$
★ New Document
○ Similar Document
○ Old Document
● Centroid

(a) Incremental product quantization (IPQ)

(b) Overall training objective

- Encoding new documents into docids by updating a subset of quantization centroids
- Overall training objective for continual indexing while alleviating forgetting of the retrieval ability

Updates to a corpus



Review updates to a corpus methods

😃 Sustainable: Making neural IR models understand new documents as well as not forget old documents in dynamic scenarios

Review updates to a corpus methods

😀 Sustainable: Making neural IR models understand new documents as well as not forget old documents in dynamic scenarios

😟 Complex: Realization and fine-tuning requires experience

Specific evaluation for updates to a corpus

Key idea: Besides ranking metrics, we focus on the forgetting degree of the old corpus

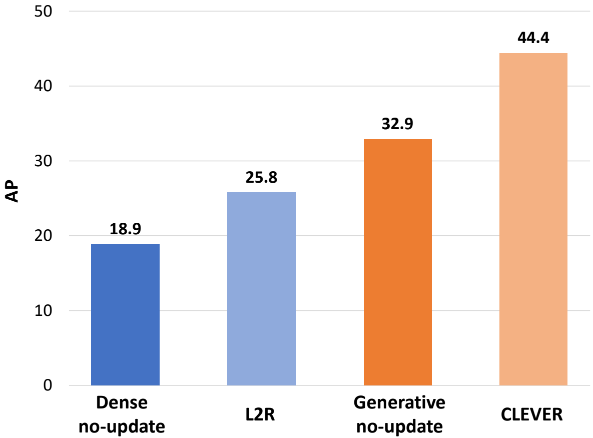
- **AP** evaluates the average performance over all sessions
- **Training time** evaluates the total time to learn new data while recalling old data
- **Forget**$_t$ evaluates how much the model forgets at session t:

$$\mathrm{Forget_t} = \frac{1}{t}\sum_{j=0}^{t-1} \max_{l \in \{0,\ldots,t-1\}} \left(p_{l,j} - p_{t,j}\right).$$

- **FWT** evaluates how well the model transfers knowledge from one session to future sessions:

$$\mathrm{FWT} = \frac{\sum_{i=1}^{j-1} \sum_{j=2}^{T} p_{i,j}}{\frac{T(T-1)}{2}}.$$

Comparison between updates to a corpus methods



- Dataset of dense retrieval: LL-MultiCPR
- Dataset of generative retrieval: CDI-MS
- Ranking metric: MRR@10
- Observations: Continual learning can effectively improve the performance of dense retrieval and generative retrieval in dynamic senario

Data source: [Cai et al., 2023, Chen et al., 2023]

Takeaway

For updates to a corpus:

Takeaway

For updates to a corpus:

- Understanding of new data and recall of old data need to be balanced

Takeaway

For updates to a corpus:

- Understanding of new data and recall of old data need to be balanced
- Effective selection of old data can help understand new data

Takeaway

For updates to a corpus:

- Understanding of new data and recall of old data need to be balanced
- Effective selection of old data can help understand new data
- Maintaining a well-structured memory is important

OOD generalizability on unseen queries: Benchmarks

Query variation datasets are designed to contain sets of queries that aim for the same information need but are expressed in various ways

OOD generalizability on unseen queries: Benchmarks

Query variation datasets are designed to contain sets of queries that aim for the same information need but are expressed in various ways

They can include paraphrased queries, queries with typos, order-swapped queries, and queries without stop words

Original query	who wrote most of the declaration of independence
Misspelling	who wr**eit** most of the declaration of independence
Naturality	~~who~~ wrote most ~~of the~~ declaration ~~of~~ independence
Order	who **declaration** most of the **wrote** of independence
Paraphrasing	who **authored** most of the declaration of independence

OOD generalizability on unseen queries: Benchmarks

Unseen query type datasets consist of queries that are not represented in the training data, either by virtue of their topic or the nature of the information being sought

OOD generalizability on unseen queries: Benchmarks

Unseen query type datasets consist of queries that are not represented in the training data, either by virtue of their topic or the nature of the information being sought

For example, the MS MARCO dataset contains 5 types of queries, i.e., location, numeric, person, description, and entity:

Query type	Percentage
Description	53.12%
Numeric	26.12%
Entity	8.81%
Location	6.17%
Person	5.78%

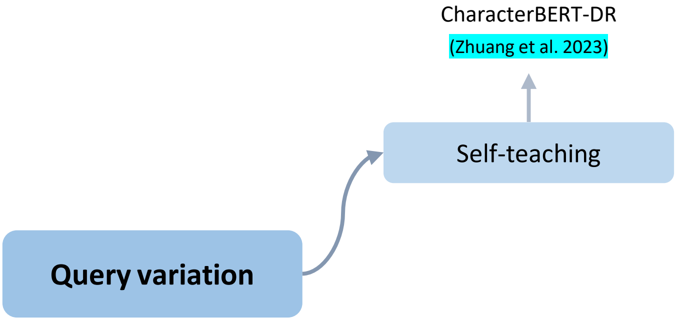
OOD generalizability on unseen queries: Benchmarks

Type	Dataset	#Q_{eval}
Query variation	DL-Typo [Zhuang and Zuccon, 2022]	60
	noisy-MS MARCO [Campos et al., 2023]	5.6k
	rewrite-MS MARCO [Campos et al., 2023]	5.6k
	noisy-NQ [Campos et al., 2023]	2k
	noisy-TQA [Campos et al., 2023]	3k
	noisy-ORCAS [Campos et al., 2023]	20k
	variations-ANTIQUE [Penha et al., 2022]	2k
	variations-TREC19 [Penha et al., 2022]	430
	[Zhuang and Zuccon, 2021]	41k
Unseen query type	MS MARCO [Nguyen et al., 2016]	15k
	L4 [Surdeanu et al., 2008]	10k

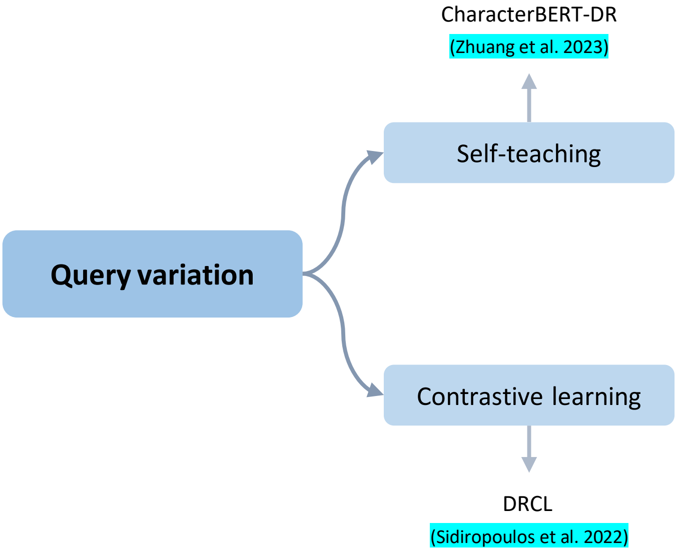
Classification of query variation

Query variation

Classification of query variation



Classification of query variation



CharacterBERT-DR

(Zhuang et al. 2023)

Self-teaching

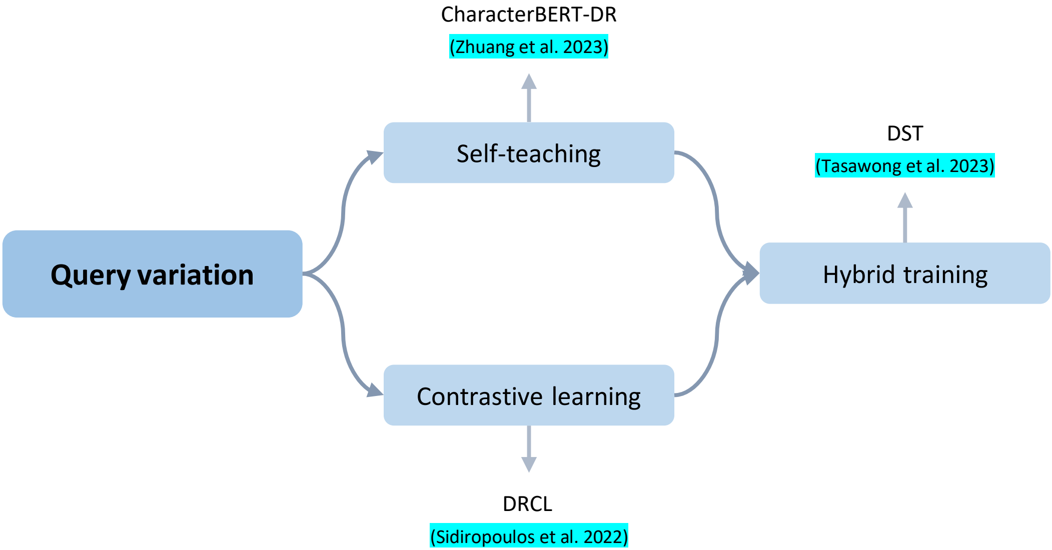
Query variation

Contrastive learning

DRCL

(Sidiropoulos et al. 2022)

Classification of query variation



Query variation

Self-teaching

CharacterBERT-DR

(Zhuang et al. 2023)

Contrastive learning

DRCL

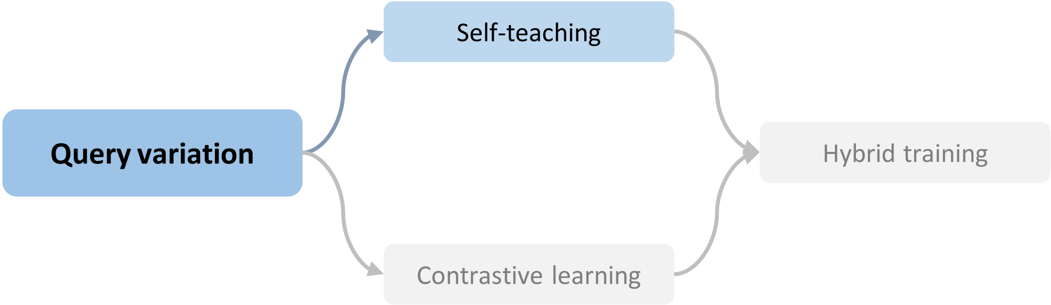
(Sidiropoulos et al. 2022)

Hybrid training

DST

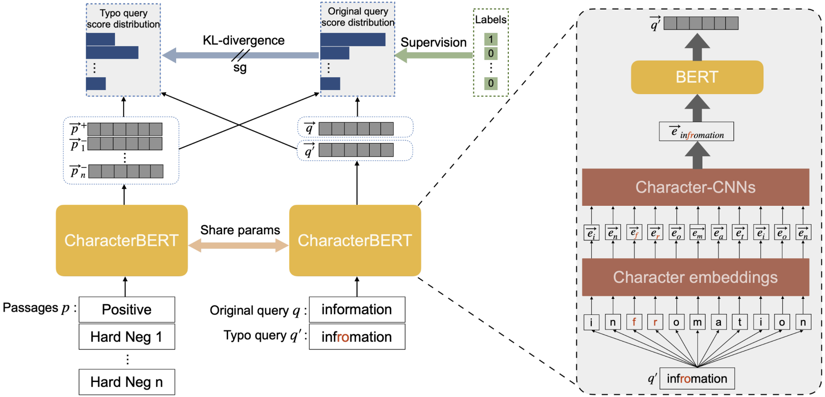
(Tasawong et al. 2023)

Query variation: Self-teaching

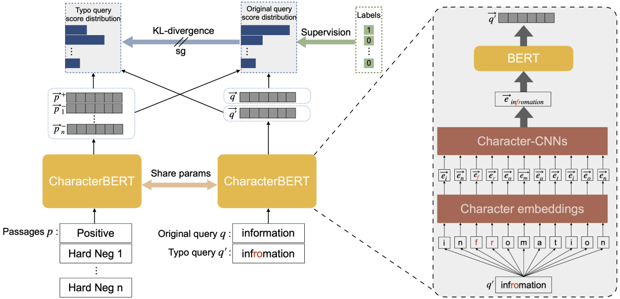


Query variation: Self-teaching

CharacterBERT-DR uses CharacterBERT with a self-teaching training method, that distills knowledge from queries without typos into queries with typos [Zhuang and Zuccon, 2022]

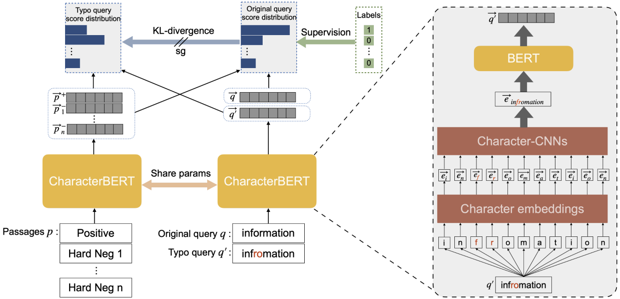


Query variation: Self-teaching



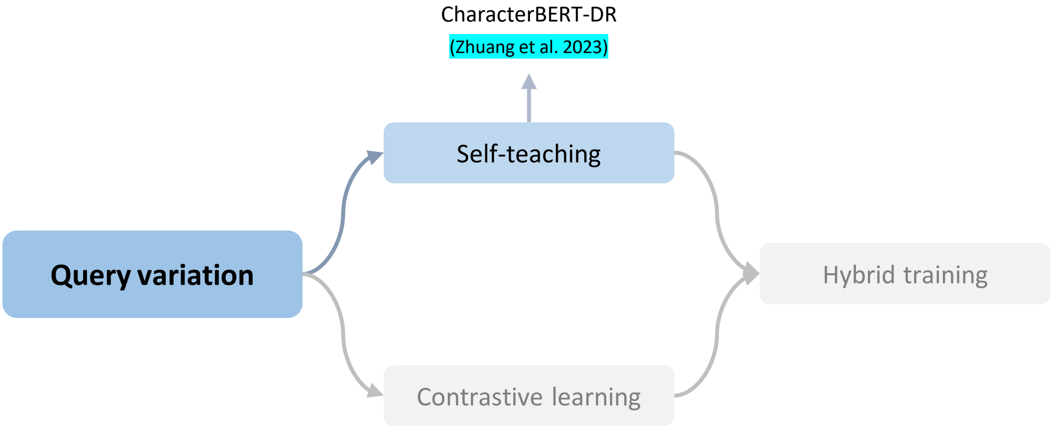
- Modify the [CLS] token embedding output from CharacterBERT to encode both queries and passages

Query variation: Self-teaching



- Modify the [CLS] token embedding output from CharacterBERT to encode both queries and passages
- Use self-teaching to minimise the difference between the score distribution obtained from the query with typos and the corresponding clean query

Review self-teaching


CharacterBERT-DR

(Zhuang et al. 2023)

Query variation

Self-teaching

Contrastive learning

Hybrid training

Review self-teaching

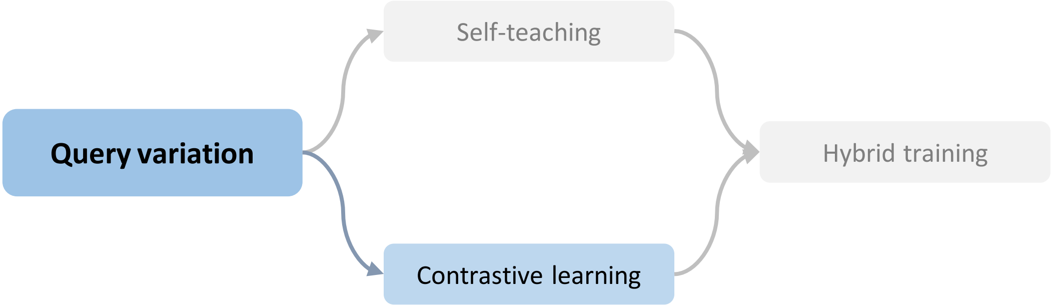
😃 Simple: Easy to implement

Review self-teaching

😃 Simple: Easy to implement

😟 Data-starved: Models may not be adequately trained when typo data is limited

Query variation: Contrastive learning



Query variation: Contrastive learning

DRCL improves robustness under query variations by combining data augmentation with contrastive learning [Sidiropoulos and Kanoulas, 2022]

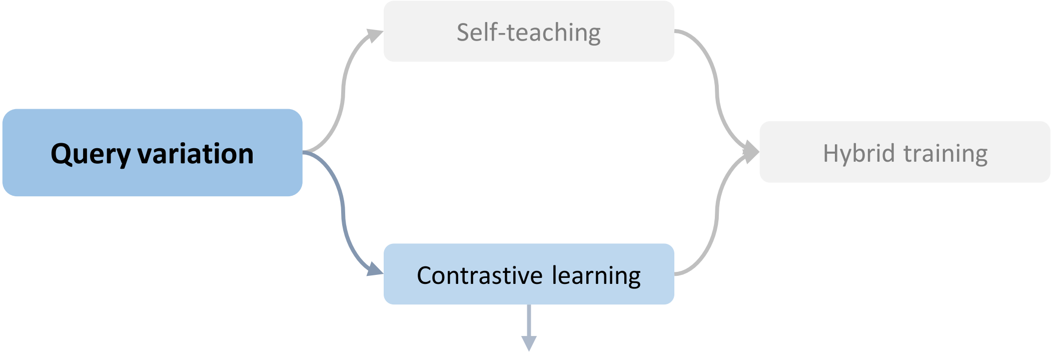
- **Data augmentation:** On training time, each original correctly query is randomly used itself or variations

Query variation: Contrastive learning

DRCL improves robustness under query variations by combining data augmentation with contrastive learning [Sidiropoulos and Kanoulas, 2022]

- **Data augmentation:** On training time, each original correctly query is randomly used itself or variations
- **Contrastive learning:** Comparing the similarity between a query and its typoed variations and other distinct queries

Review contrastive learning



DRCL

(Sidiropoulos et al. 2022)

Review contrastive learning

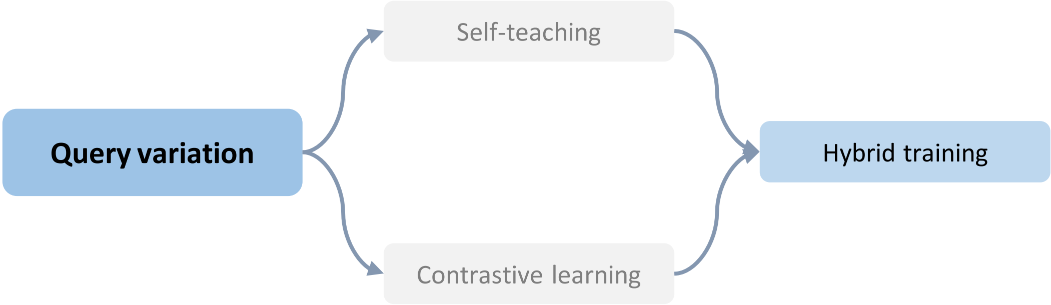
😃 Data-rich: Models can be fully trained

Review contrastive learning

😃 Data-rich: Models can be fully trained

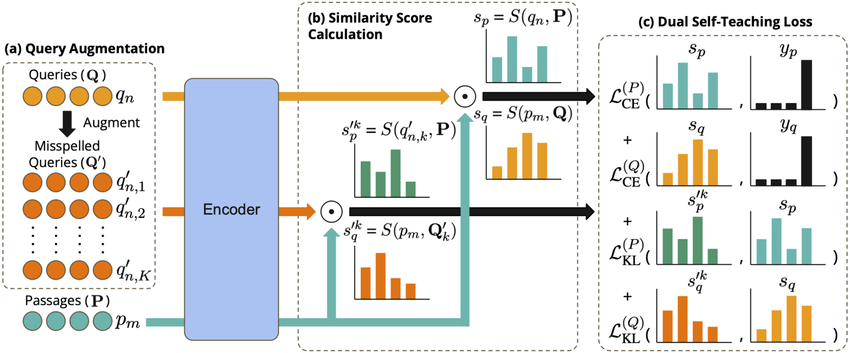
😟 Costly: Need to construct large amounts of training data

Query variation: Hybrid training



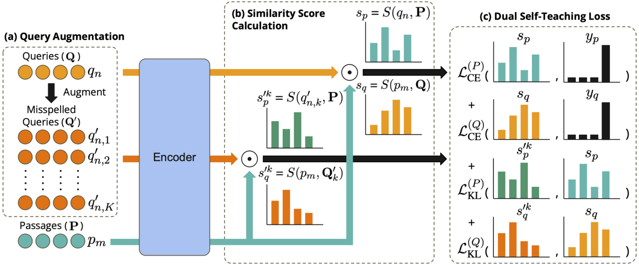
Query variation: Hybrid training

DST adopts the idea of contrastive learning and self-teaching to learn robust representations [Tasawong et al., 2023]



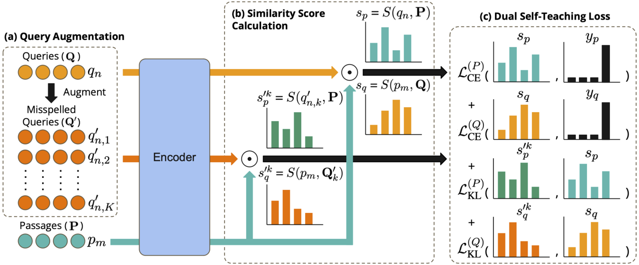
(a) Query Augmentation
Queries ($\mathbf{Q}$)
q_n
Augment
Misspelled Queries ($\mathbf{Q}'$)
$q'_{n,1}$
$q'_{n,2}$
$q'_{n,K}$
Passages ($\mathbf{P}$)
p_m
Encoder
(b) Similarity Score Calculation
$s_p = S(q_n, \mathbf{P})$
$s_q = S(p_m, \mathbf{Q})$
$s'^k_p = S(q'_{n,k}, \mathbf{P})$
$s'^k_q = S(p_m, \mathbf{Q}'_k)$
(c) Dual Self-Teaching Loss
$\mathcal{L}^{(P)}_{\text{CE}}(s_p, y_p)$
$+ \mathcal{L}^{(Q)}_{\text{CE}}(s_q, y_q)$
$+ \mathcal{L}^{(P)}_{\text{KL}}(s'^k_p, s_p)$
$+ \mathcal{L}^{(Q)}_{\text{KL}}(s'^k_q, s_q)$

Query variation: Hybrid training


(a) Query Augmentation
Queries ($\mathbf{Q}$)
q_n
Augment
Misspelled Queries ($\mathbf{Q}'$)
$q'_{n,1}$
$q'_{n,2}$
$q'_{n,K}$
Passages ($\mathbf{P}$)
p_m
Encoder
(b) Similarity Score Calculation
$s_p = S(q_n, \mathbf{P})$
$s_q = S(p_m, \mathbf{Q})$
$s_p'^k = S(q'_{n,k}, \mathbf{P})$
$s_q'^k = S(p_m, \mathbf{Q}'_k)$
(c) Dual Self-Teaching Loss
$\mathcal{L}_{\mathrm{CE}}^{(P)}(s_p, y_p)$
$+ \mathcal{L}_{\mathrm{CE}}^{(Q)}(s_q, y_q)$
$+ \mathcal{L}_{\mathrm{KL}}^{(P)}(s_p'^k, s_p)$
$+ \mathcal{L}_{\mathrm{KL}}^{(Q)}(s_q'^k, s_q)$

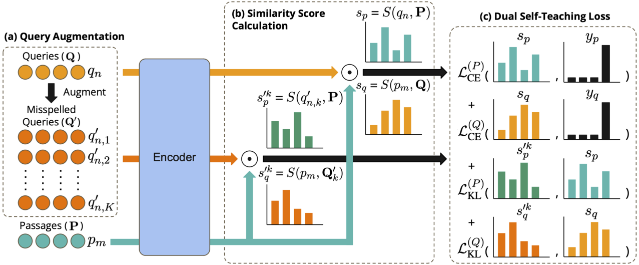
- **Alignment:** align queries with their corresponding passages

Query variation: Hybrid training


(a) Query Augmentation
Queries ($\mathbf{Q}$)
q_n
Augment
Misspelled Queries ($\mathbf{Q}'$)
$q'_{n,1}$
$q'_{n,2}$
$q'_{n,K}$
Passages ($\mathbf{P}$)
p_m
Encoder
(b) Similarity Score Calculation
$s_p = S(q_n, \mathbf{P})$
$s_q = S(p_m, \mathbf{Q})$
$s'^k_p = S(q'_{n,k}, \mathbf{P})$
$s'^k_q = S(p_m, \mathbf{Q}'_k)$
(c) Dual Self-Teaching Loss
s_p y_p
$\mathcal{L}^{(P)}_{\text{CE}}($, $)$
+
s_q y_q
$\mathcal{L}^{(Q)}_{\text{CE}}($, $)$
+
s'^k_p s_p
$\mathcal{L}^{(P)}_{\text{KL}}($, $)$
+
s'^k_q s_q
$\mathcal{L}^{(Q)}_{\text{KL}}($, $)$

- **Alignment:** align queries with their corresponding passages
- **Robustness:** align misspelled queries with their pristine queries

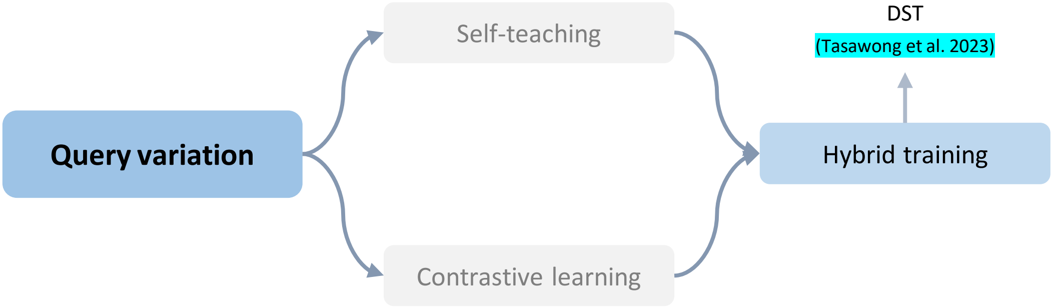
Query variation: Hybrid training



(a) Query Augmentation
Queries ($\mathbf{Q}$)
q_n
Augment
Misspelled Queries ($\mathbf{Q}'$)
$q'_{n,1}$
$q'_{n,2}$
$q'_{n,K}$
Passages ($\mathbf{P}$)
p_m
Encoder
(b) Similarity Score Calculation
$s_p = S(q_n, \mathbf{P})$
$s_q = S(p_m, \mathbf{Q})$
$s'^k_p = S(q'_{n,k}, \mathbf{P})$
$s'^k_q = S(p_m, \mathbf{Q}'_k)$
(c) Dual Self-Teaching Loss
$\mathcal{L}^{(P)}_{\text{CE}}(s_p, y_p)$
$+ \mathcal{L}^{(Q)}_{\text{CE}}(s_q, y_q)$
$+ \mathcal{L}^{(P)}_{\text{KL}}(s'^k_p, s_p)$
$+ \mathcal{L}^{(Q)}_{\text{KL}}(s'^k_q, s_q)$

- **Alignment:** align queries with their corresponding passages
- **Robustness:** align misspelled queries with their pristine queries
- **Contrast:** separate queries that refer to different passages and passages that correspond to different queries

Review hybrid training



Review hybrid training

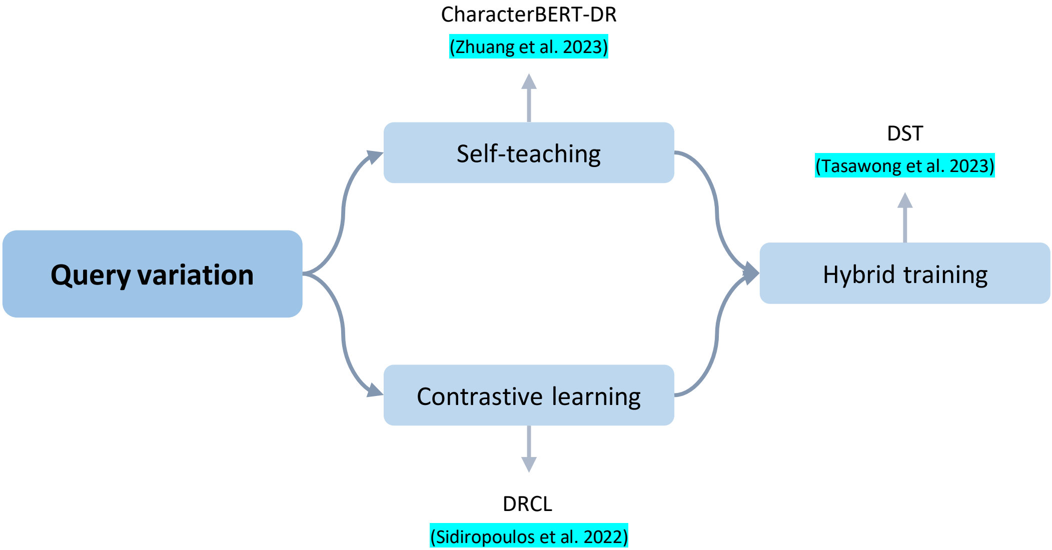
😀 Sufficient: Multiple training objectives guarantee model robustness to query variants

Review hybrid training

😃 Sufficient: Multiple training objectives guarantee model robustness to query variants

🤔 Empirical: The need to balance between different training objectives

Query variation

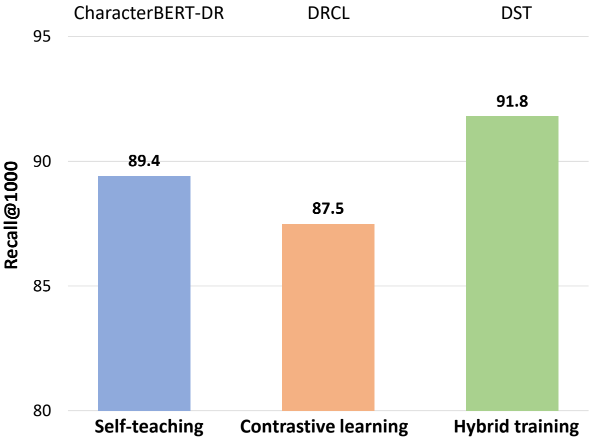


Evaluation

In addition to MRR and NDCG, the ranking performance under unseen queries is evaluated by other common metrics for query variation and unseen query type

- **Recall** measures the proportion of relevant documents that are successfully retrieved from the total amount of relevant documents available
- **MAP** quantifies the average precision of retrieval across different recall levels, effectively summarizing the precision at each point where a relevant document is retrieved

Comparison between query variation methods



- Dataset: MS MARCO (with typo)
- Observations: Self-teaching is made more effective by contrastive learning, and combining these two training methods allows for further model robustness improvements

Data source: [Sidiropoulos and Kanoulas, 2022, Tasawong et al., 2023, Zhuang and Zuccon, 2

Takeaway

For query variation:

Takeaway

For query variation:

- An appropriate backbone is the foundation

Takeaway

For query variation:

- An appropriate backbone is the foundation
- Alignment and contrast are key

Takeaway

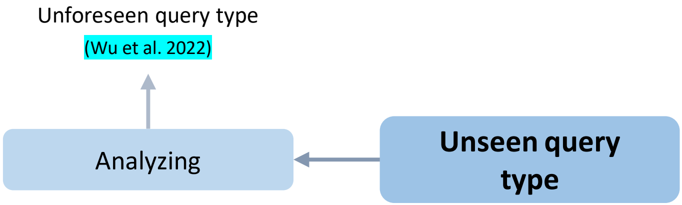
For query variation:

- An appropriate backbone is the foundation
- Alignment and contrast are key
- Integration of various training objectives is the icing on the cake

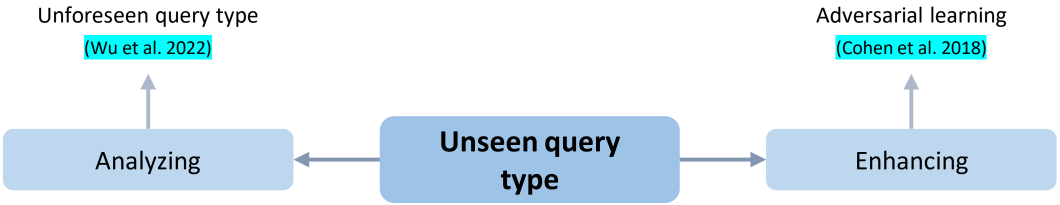
Classification of unseen query type

Unseen query type

Classification of unseen query type



Classification of unseen query type



Unforeseen query type

(Wu et al. 2022)

Analyzing

Unseen query type

Enhancing

Adversarial learning

(Cohen et al. 2018)

Specific metrics for unseen query type

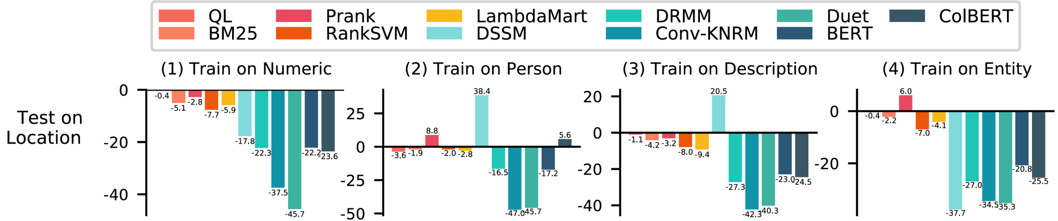
DR$_{OOD}$ evaluates the drop rate between the ranking performance on the original type of queries and the ranking performance on the unseen type of queries:

$$DR_{OOD} = \frac{p_{OOD} - p_{IID}}{p_{IID}},$$

where p_{IID} is the ranking performance on original type of queries and p_{OOD} is the ranking performance on unseen type of queries

Unseen query type: Analyzing

NRMs have poor performance on unseen query types

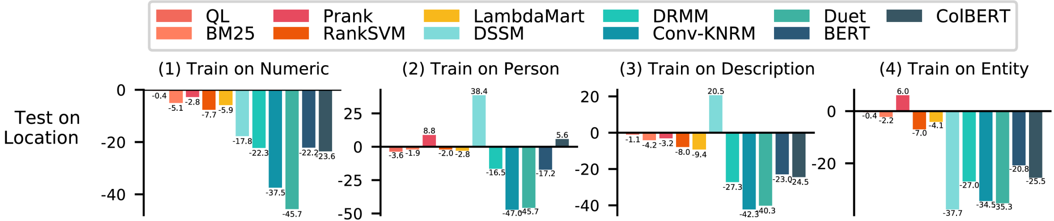


- NRMs with deep networks can fit seen query types well, at the cost of further loss in performance on the held-out OOD query types

Data source: [Wu et al., 2022]

Unseen query type: Analyzing

NRMs have poor performance on unseen query types



- NRMs with deep networks can fit seen query types well, at the cost of further loss in performance on the held-out OOD query types
- Pre-trained models have shown good robustness to OOD query types

Data source: [Wu et al., 2022]

Unseen query type: Enhancing

Cohen et al. study the effectiveness of adversarial learning as a cross-domain regularizer to deal with unseen query type [Cohen et al., 2018]

- Force the NRMs to learn domain-independent features that are useful to estimate relevance

Unseen query type: Enhancing

Cohen et al. study the effectiveness of adversarial learning as a cross-domain regularizer to deal with unseen query type [Cohen et al., 2018]

- Force the NRMs to learn domain-independent features that are useful to estimate relevance
- Shift the model parameters in the opposite direction to the domain specific spaces on the manifold

Unseen query type: Enhancing

Cohen et al. study the effectiveness of adversarial learning as a cross-domain regularizer to deal with unseen query type [Cohen et al., 2018]

- Force the NRMs to learn domain-independent features that are useful to estimate relevance
- Shift the model parameters in the opposite direction to the domain specific spaces on the manifold

Further work in this field is waiting to be explored . . .

References

References i

L. Bonifacio, H. Abonizio, M. Fadaee, and R. Nogueira. Inpars: Data augmentation for information retrieval using large language models. *arXiv preprint arXiv:2202.05144*, 2022.

Y. Cai, K. Bi, Y. Fan, J. Guo, W. Chen, and X. Cheng. L2r: Lifelong learning for first-stage retrieval with backward-compatible representations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 183–192, 2023.

D. Campos, C. Zhai, and A. Magnani. Noise-robust dense retrieval via contrastive alignment post training. *arXiv preprints arXiv:2304.03401*, 2023.

J. Chen, R. Zhang, J. Guo, M. de Rijke, W. Chen, Y. Fan, and X. Cheng. Continual learning for generative retrieval over dynamic corpora. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 306–315, 2023.

D. Cohen, B. Mitra, K. Hofmann, and W. B. Croft. Cross domain regularization for neural ranking models using adversarial learning. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1025–1028, 2018.

References ii

N. Craswell, B. Mitra, E. Yilmaz, D. Campos, and J. Lin. Ms marco: Benchmarking ranking models in the large-data regime. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1566–1576, 2021.

G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, and E. Grave. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021.

P. Kasela, G. Pasi, R. Perego, and N. Tonellotto. Desire-me: Domain-enhanced supervised information retrieval using mixture-of-experts. In *European Conference on Information Retrieval*, pages 111–125. Springer, 2024.

S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, and N. Goharian. Simplified data wrangling with ir_datasets. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2429–2436, 2021.

T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@NIPS*, 2016.

References iii

J. Ni, C. Qu, J. Lu, Z. Dai, G. H. Abrego, J. Ma, V. Zhao, Y. Luan, K. Hall, M.-W. Chang, et al. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, 2022.

G. Penha, A. Câmara, and C. Hauff. Evaluating the robustness of retrieval pipelines with query variation generators. In *European Conference on Information Retrieval*, pages 397–412. Springer, 2022.

G. Sidiropoulos and E. Kanoulas. Analysing the robustness of dual encoders for dense retrieval against misspellings. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2132–2136, 2022.

M. Surdeanu, M. Ciaramita, and H. Zaragoza. Learning to rank answers on large online qa collections. In *proceedings of ACL-08: HLT*, pages 719–727, 2008.

P. Tasawong, W. Ponwitayarat, P. Limkonchotiwat, C. Udomcharoenchaikit, E. Chuangsuwanich, and S. Nutanong. Typo-robust representation learning for dense retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1106–1115, 2023.

References iv

N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

K. Wang, N. Thakur, N. Reimers, and I. Gurevych. Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, 2022.

C. Wu, R. Zhang, J. Guo, Y. Fan, and X. Cheng. Are neural ranking models robust? *ACM Transactions on Information Systems*, 41(2):1–36, 2022.

Y. Yu, C. Xiong, S. Sun, C. Zhang, and A. Overwijk. Coco-dr: Combating distribution shifts in zero-shot dense retrieval with contrastive and distributionally robust learning. *arXiv preprint arXiv:2210.15212*, 2022.

S. Zhuang and G. Zuccon. Dealing with typos for bert-based passage retrieval and ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2836–2842, 2021.

References v

S. Zhuang and G. Zuccon. Characterbert and self-teaching for improving the robustness of dense retrievers on queries with typos. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1444–1454, 2022.