Domain Adaptation for Dense Information Retrieval without Labeled Data

Nils Reimers
HuggingFace
Creator of Sentence-Transformers (www.SBERT.net)

Neural Search – Why all the Hype?

- Real example on (Simple) Wikipedia (170k documents)
- Query: `What is the capital of the United States?`
- Top-3 Hits

Lexical Search (BM25)

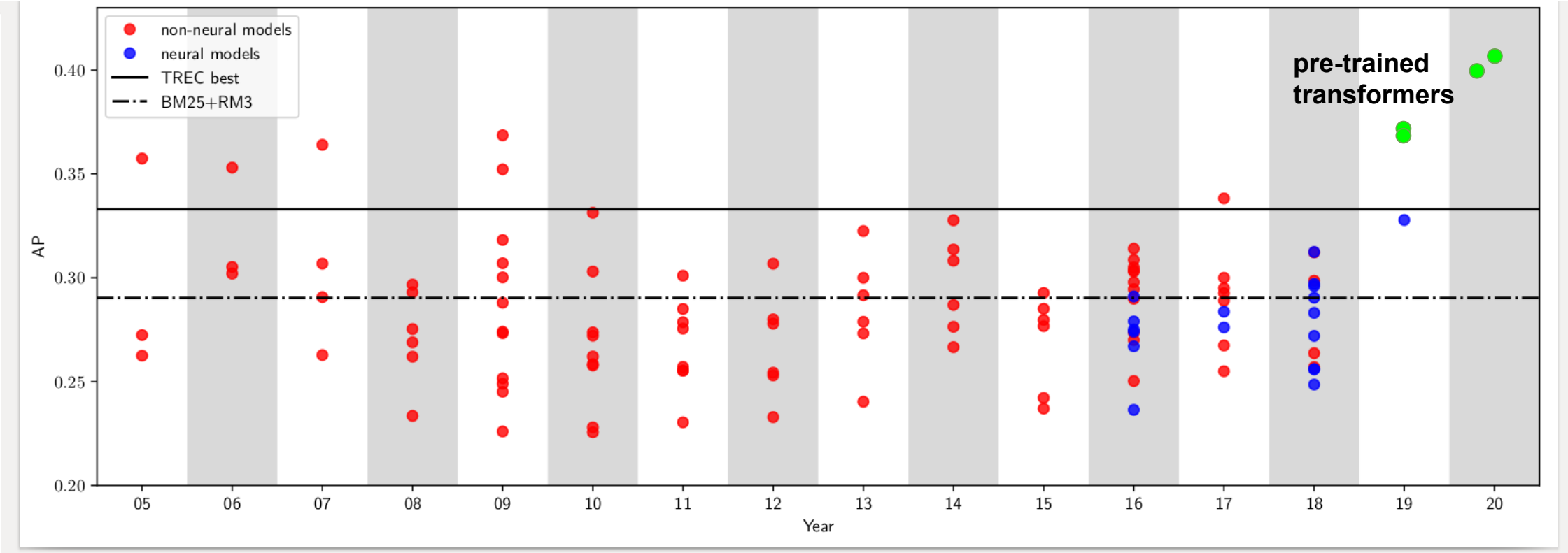
- **Capital** punishment (the death penalty) has existed in the **United States** […]
- Ohio is one of the 50 **states** in the **United States**. Its **capital** is Columbus. […]
- Nevada is one of the **United States' states**. Its **capital** […]

Neural Search

- Washington, D.C. […] is the **capital of the United States**. […]
- A capital city (or capital town or just capital) is a city or town, […]
- The United States **Capitol** is the building where the United States Congress meets […]

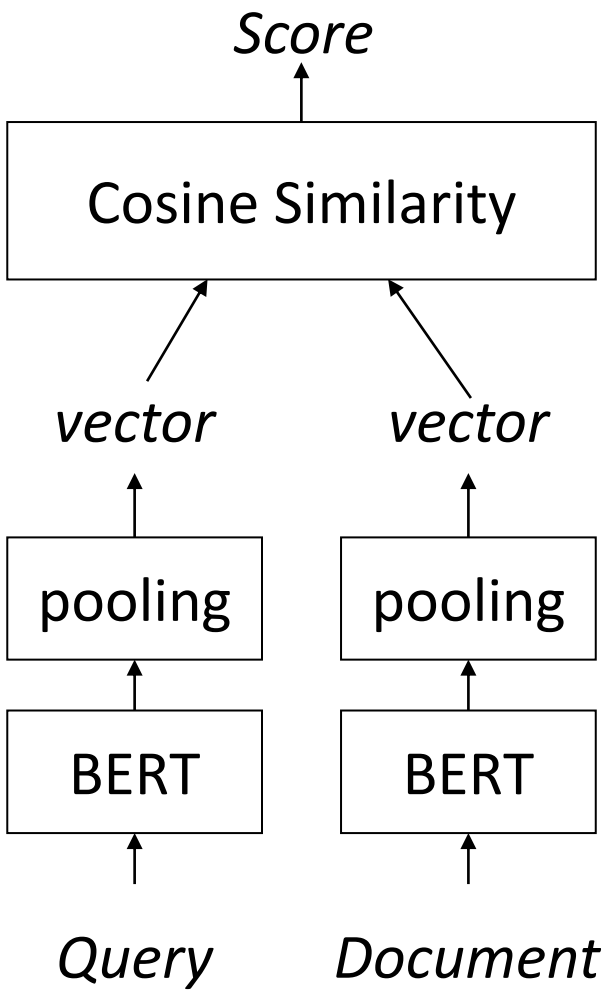
Neural Search – Why all the Hype?

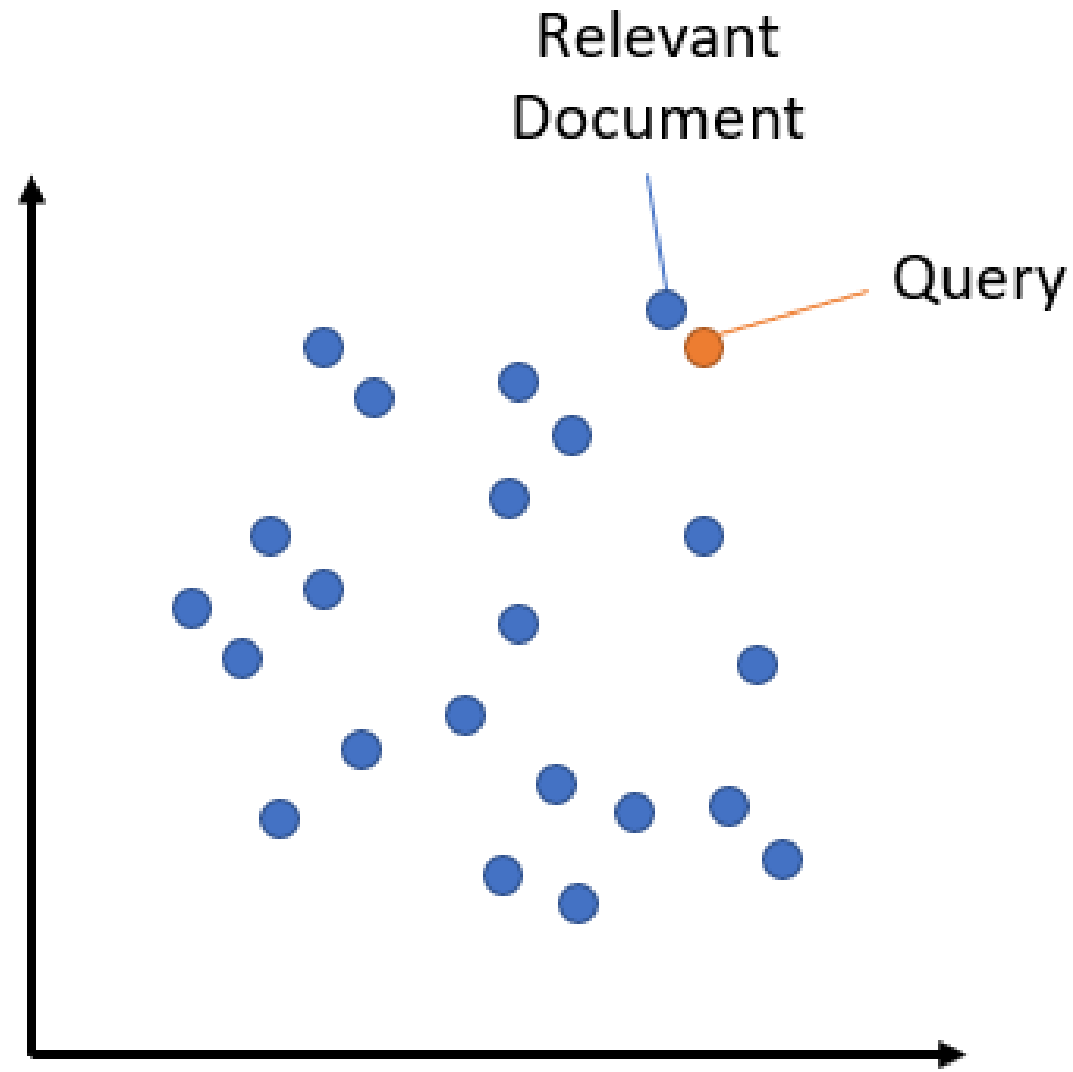
TREC Robust 04



Source: *Yang et al. SIGIR 2019. Critically Examining the Neural Hype...*

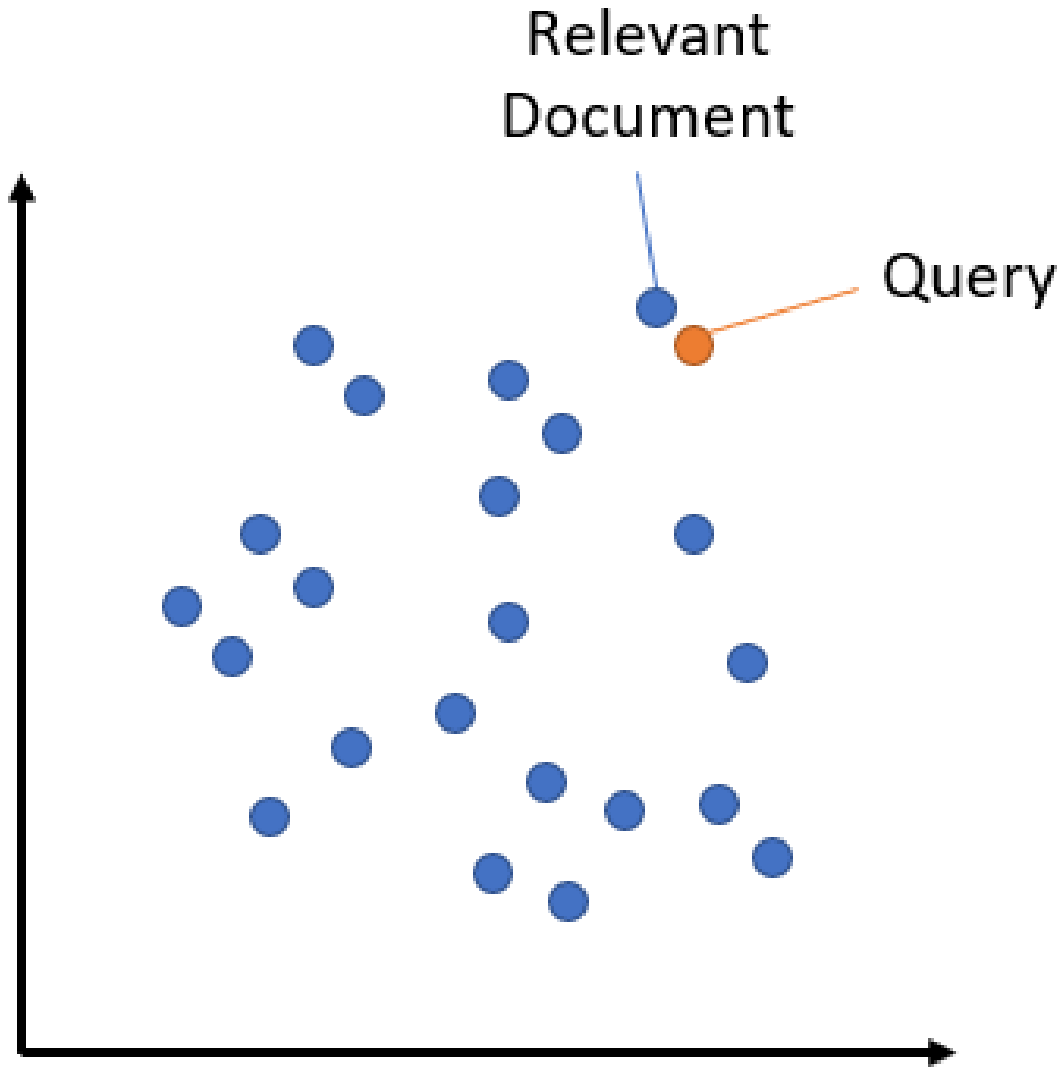
Neural Search – Bi-Encoders



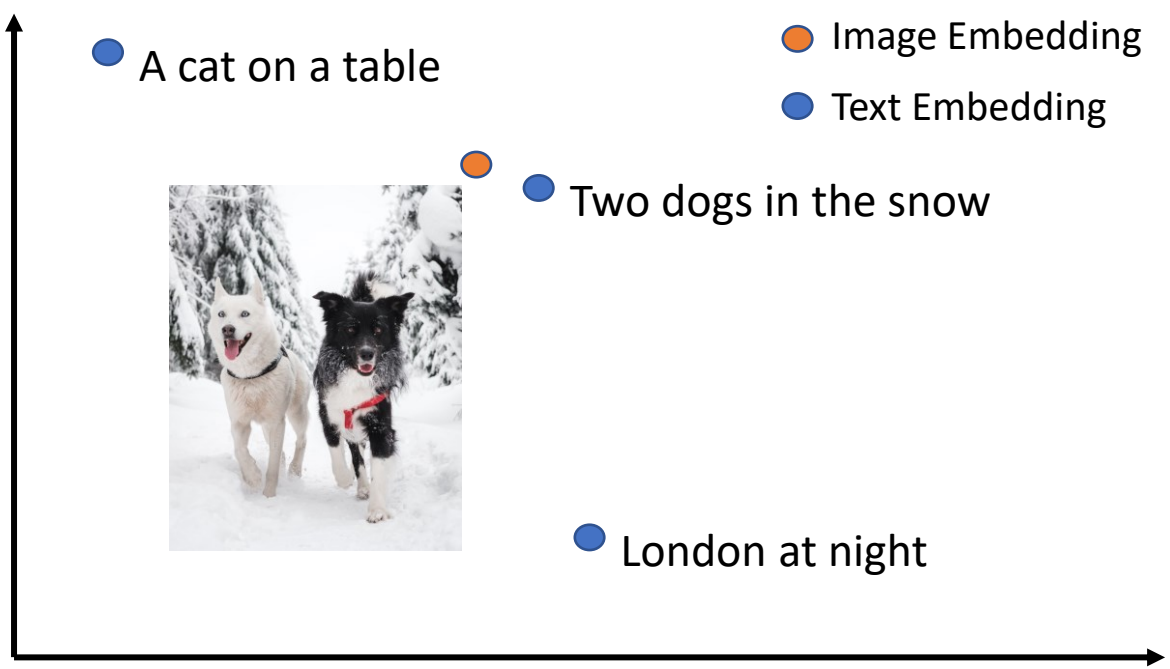


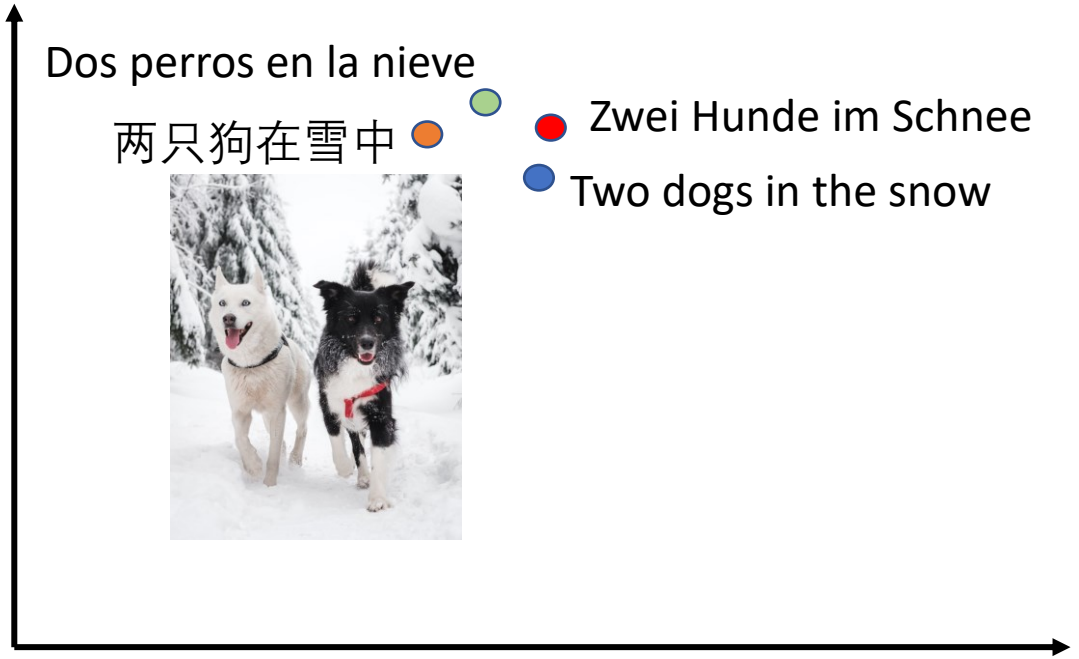
Neural Search – Bi-Encoders

- Can overcome the lexical gap
  - US vs USA vs United States
- Respects the word order
  - Visa from Germany to Canada
  - Visa from Canada to Germany
- Knows about related terms
  - “spearman correlation numpy” finds the entry: “spearman correlation SciPy”



Multi-Modal & Multi-Lingual Search





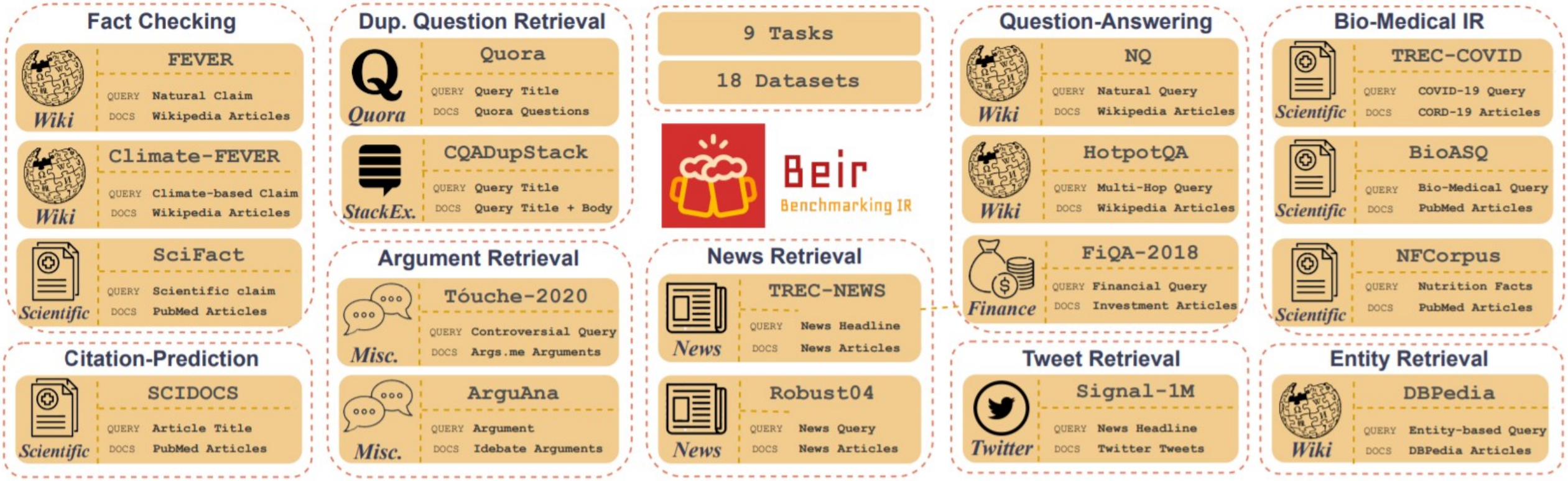
Bi-Encoders became popular

- Who is using it?
  - Baidu (https://arxiv.org/abs/2106.03373)
  - Facebook (https://arxiv.org/abs/2006.11632)
  - Netflix
  - Spotify
  - Amazon
  - …
- Especially useful for exploring
  - “Find movies similar to The Matrix”

Bi-Encoders and the Curse of the Unknowns

- How do Bi-Encoders handle unknown words?
  - Not seen during pre-training
  - Not seen during fine-tuning
- Our world is in constant change
- Where to put these words in a vector space?
  - XLNet
  - Clexchain
  - Forwrd
  - 0xc004f213
- How to know
  - Corona Virus ⇔ COVID-19 ⇔ SARS-Cov-2

BEIR – Benchmarking IR



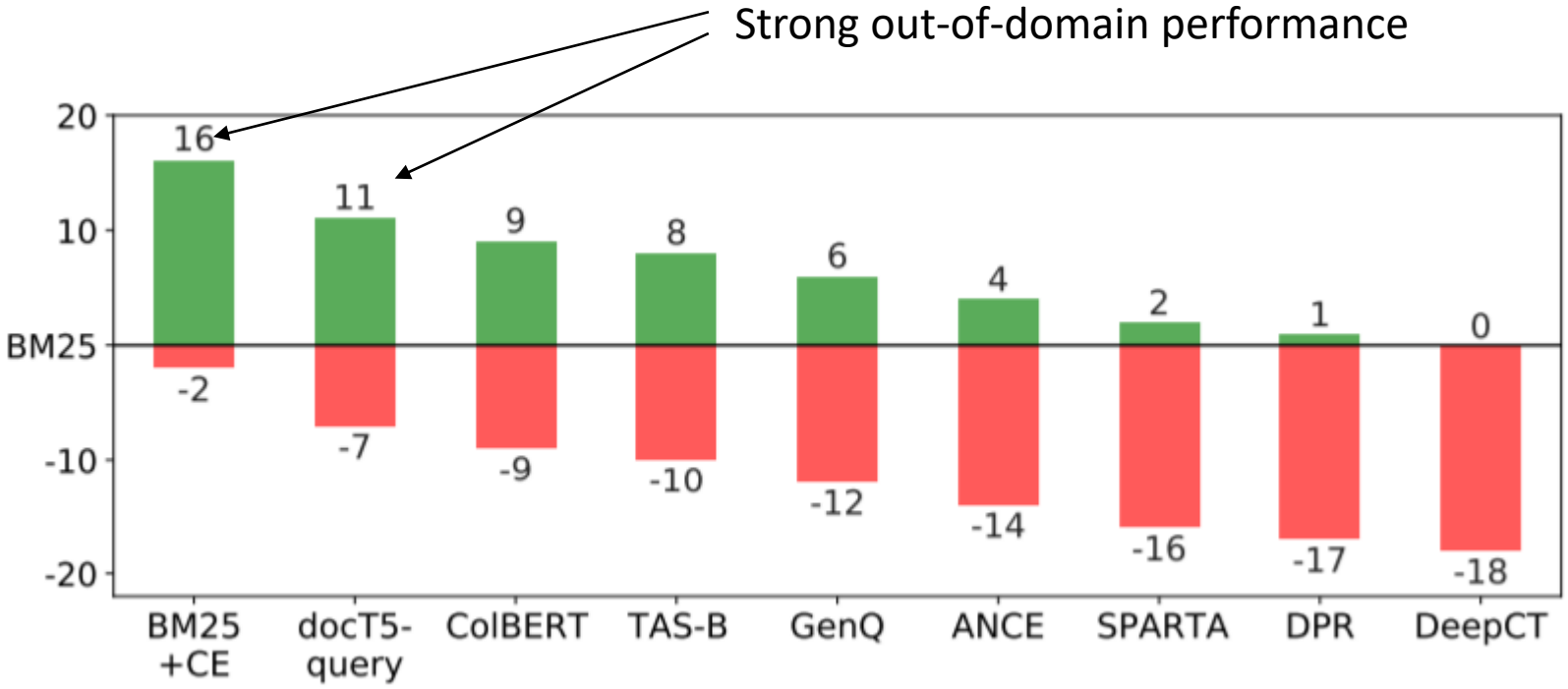
BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models, https://arxiv.org/abs/2104.08663

Bi-Encoders vs Lexical Search

Dataset	BM25	Dense Model (TAS-B)	Difference
In-Domain	22.8	40.8	+18.0
BioASQ	46.5	38.3	-8.2
SCIDOCS	15.9	14.9	-1.0

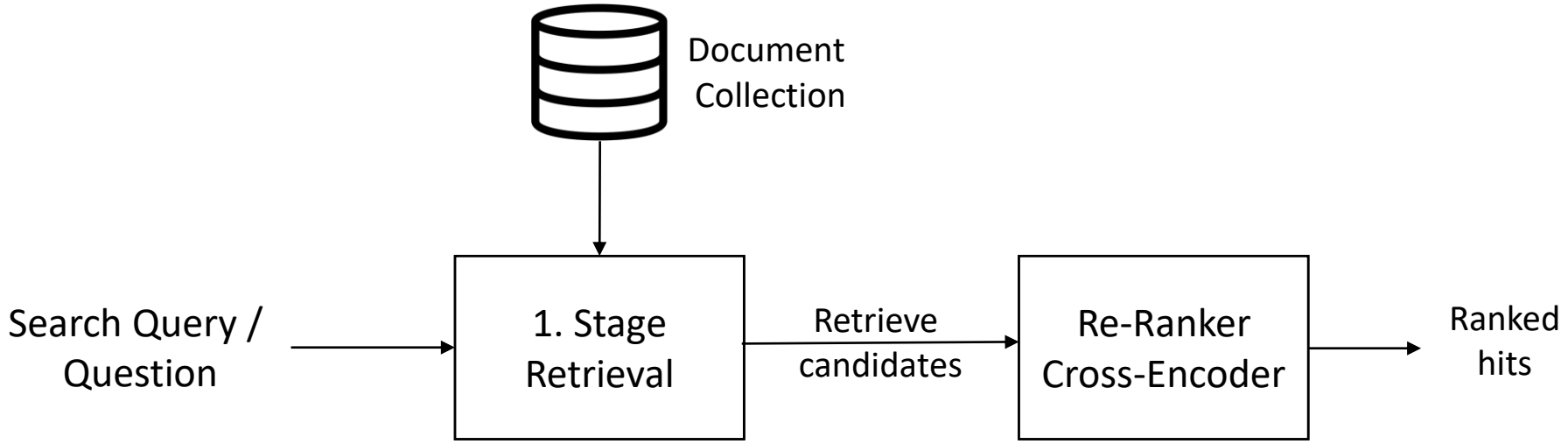
- BM25 was better on 10 / 18 datasets

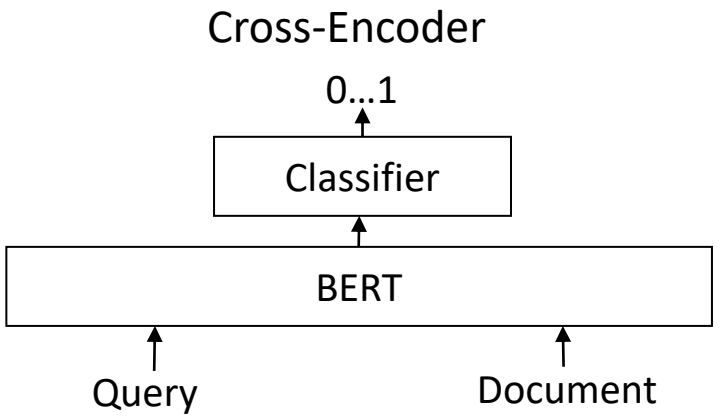
Do Models Generalize?



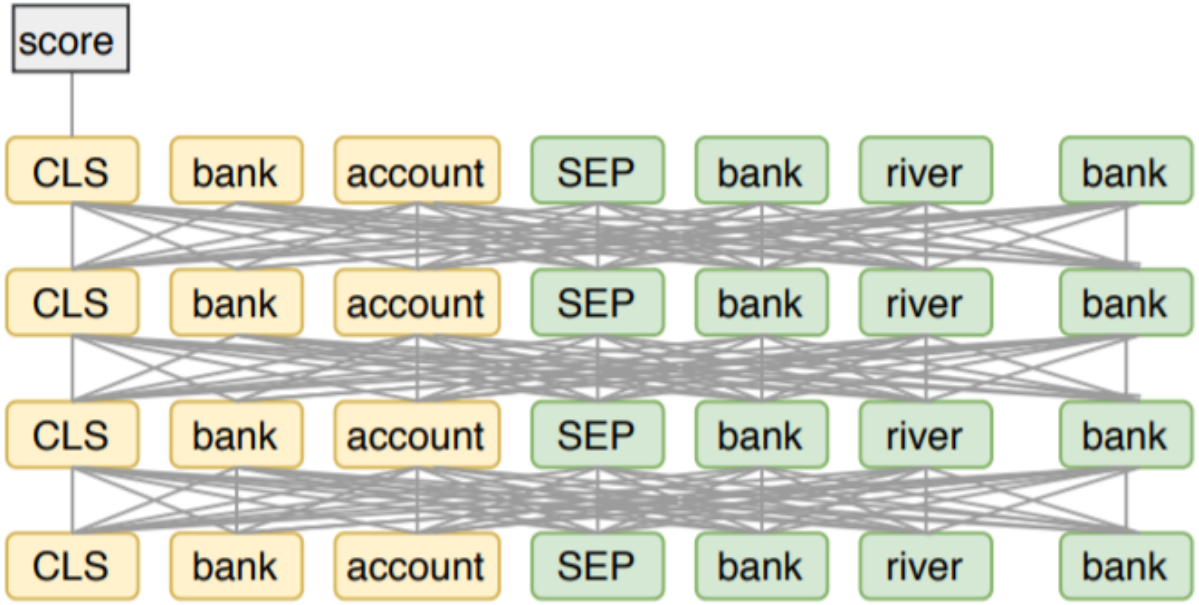
- BM25 lexical search a strong baseline
- BM25 + CrossEncoder re-ranking perform the best
- Embedding models (TAS-B, ANCE, DPR) with issues for unknown domains

Cross-Encoders

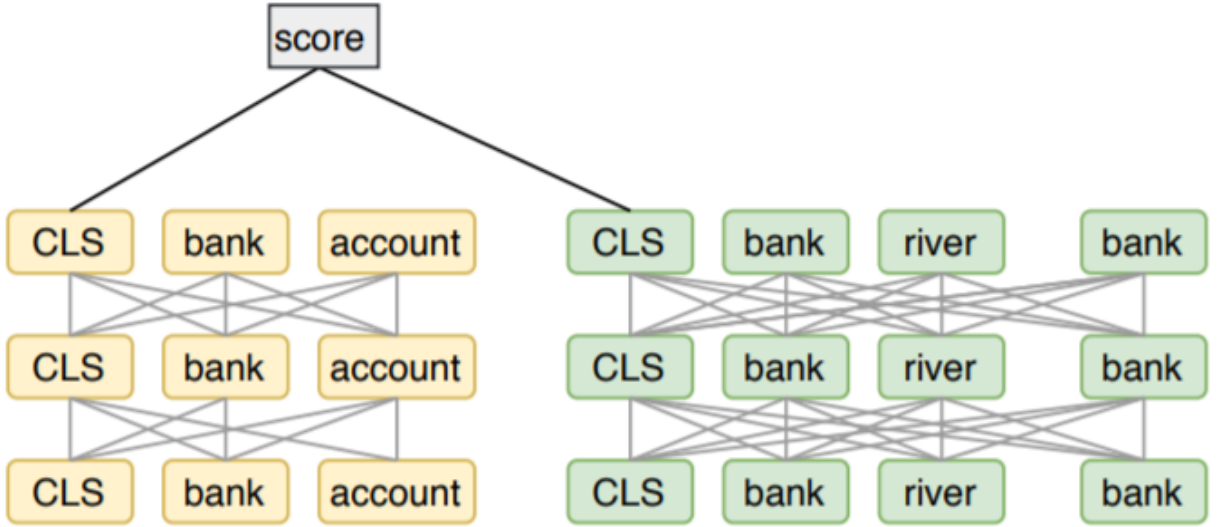




Cross-Encoders vs Bi-Encoders



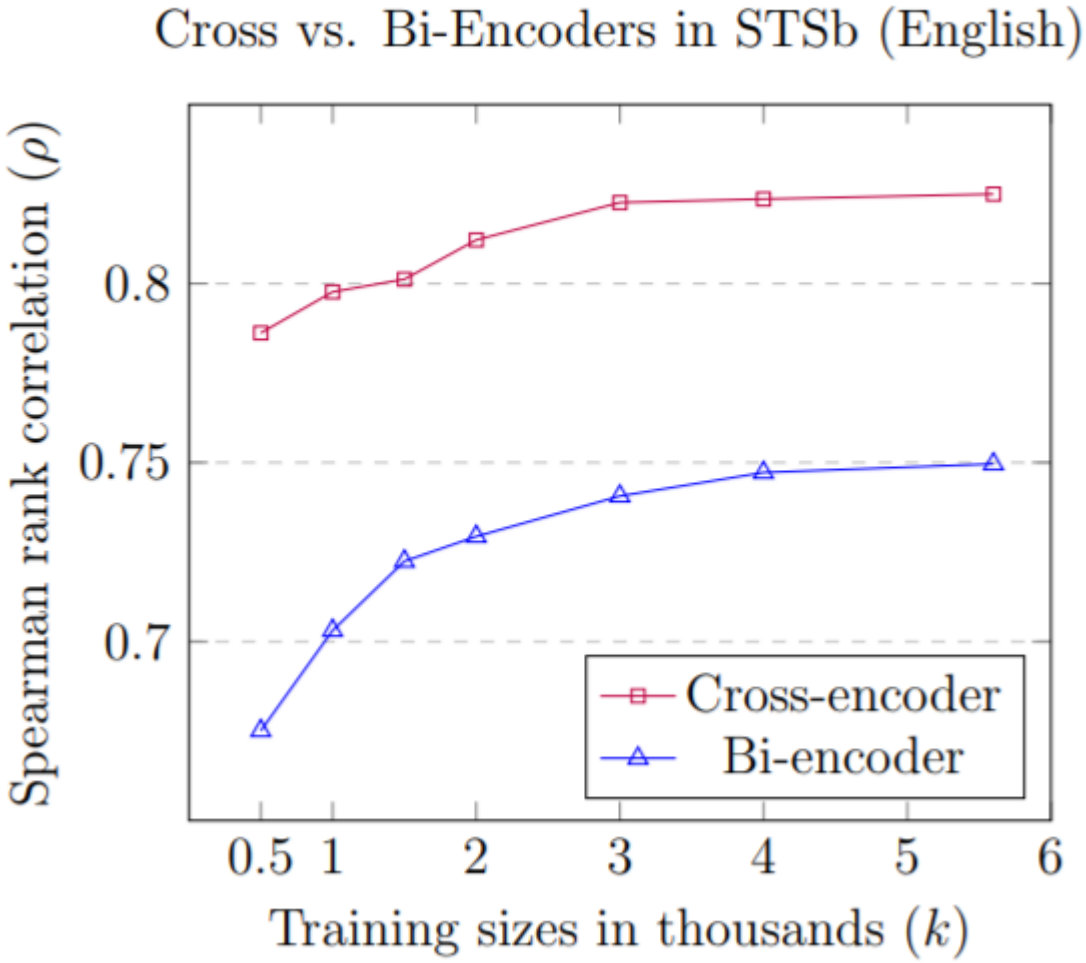
(a) Cross-Attention Model (e.g., BERT reranker)



(b) Dense Retrievers

Img: https://arxiv.org/pdf/2104.07186.pdf

Cross-Encoders vs Bi-Encoders



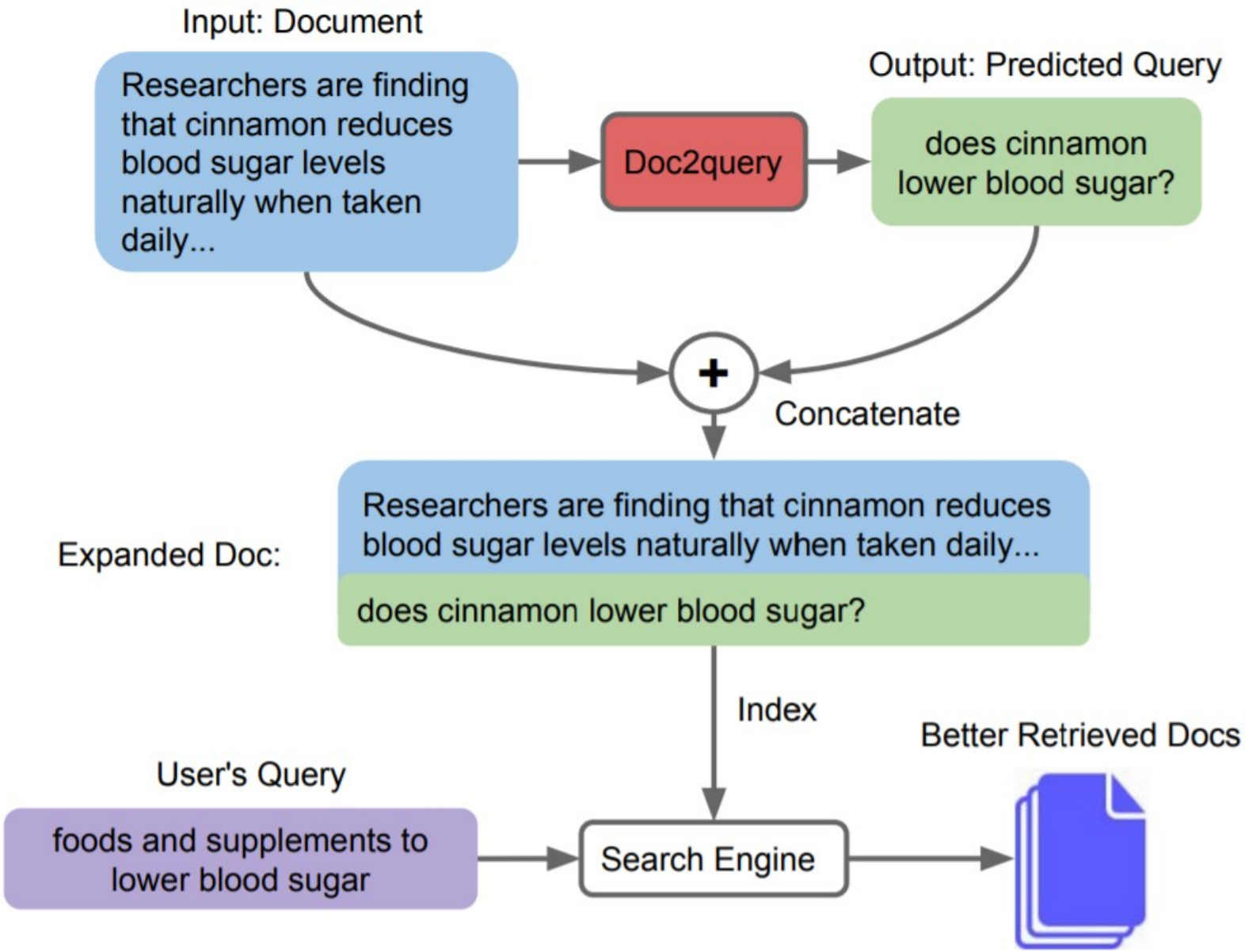
Img: https://arxiv.org/pdf/2010.08240.pdf

Cross-Encoders vs Bi-Encoders

Dataset	BM25	Dense Model (TAS-B)	BM25 + CE
In-Domain	22.8	40.8	41.3
BioASQ	46.5	38.3	52.3
CQADupStack	29.9	31.4	37.0
TREC-COVID	65.5	48.1	75.7
SCIDOCS	15.9	14.9	16.6

- BM25 + CE on average 13.8 points better than dense

doc2Query



Img: https://arxiv.org/pdf/1904.08375.pdf

Why not using Cross-Encoders / doc2query?

- Cross-Encoders are slow (even small ones)
  - E.g. query has 10 tokens, docs have 240 tokens, re-rank 100 docs
  - Bi-Encoders: Compute embedding for query (e.g. 10ms)
  - Cross-Encoder: Re-rank 100 x 250 token docs
    - Forward pass for 250 tokens takes ~25*25 = 625 times longer
    - Overall 62,500 times longer to get results

- Doc2query is slow at indexing
  - Generates 40 query per passage
  - Question generation is extremely slow
  - Costs to generate queries for 8M docs: $750
  - Computing dense embeddings: $1

Source: https://arxiv.org/pdf/2108.08513.pdf
https://medium.com/@nils_reimers/openai-gpt-3-text-embeddings-really-a-new-state-of-the-art-in-dense-text-embeddings-6571fe3ec9d9

How to Adapt Bi-Encoders to New Domains?

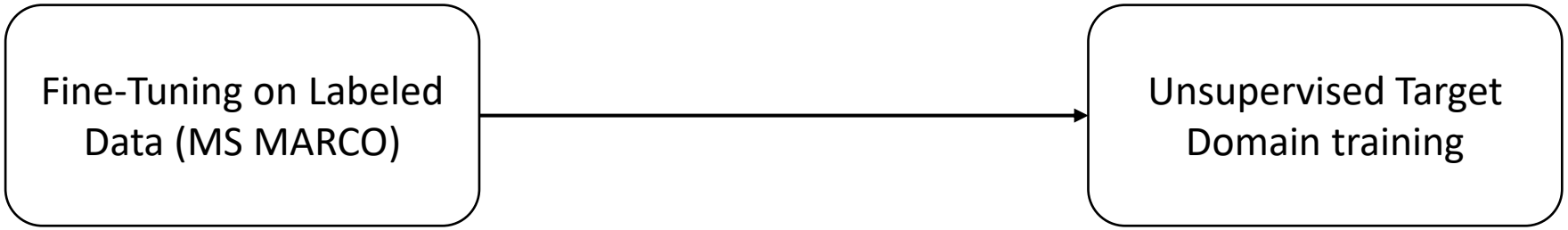
Goal: No need for labeled data

1. Adaptive Pre-Training

2. GPL - Generative Pseudo Labeling (https://arxiv.org/abs/2112.07577)

Adaptive Pre-Training

What we want:

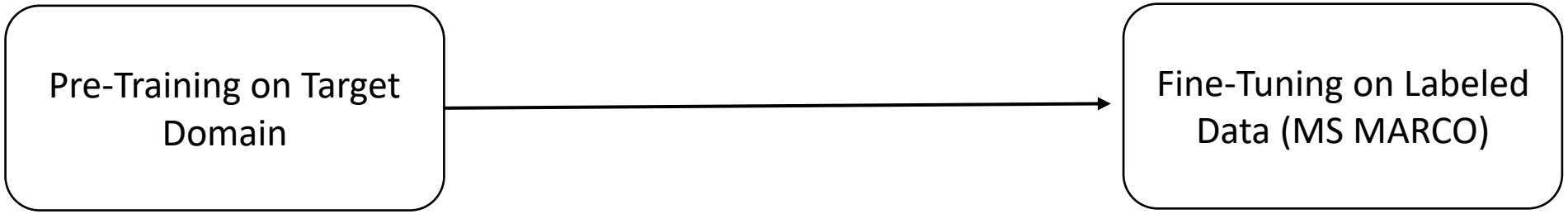


Sadly doesn't work well

- Improves performance only in some cases
- Improvement not so large

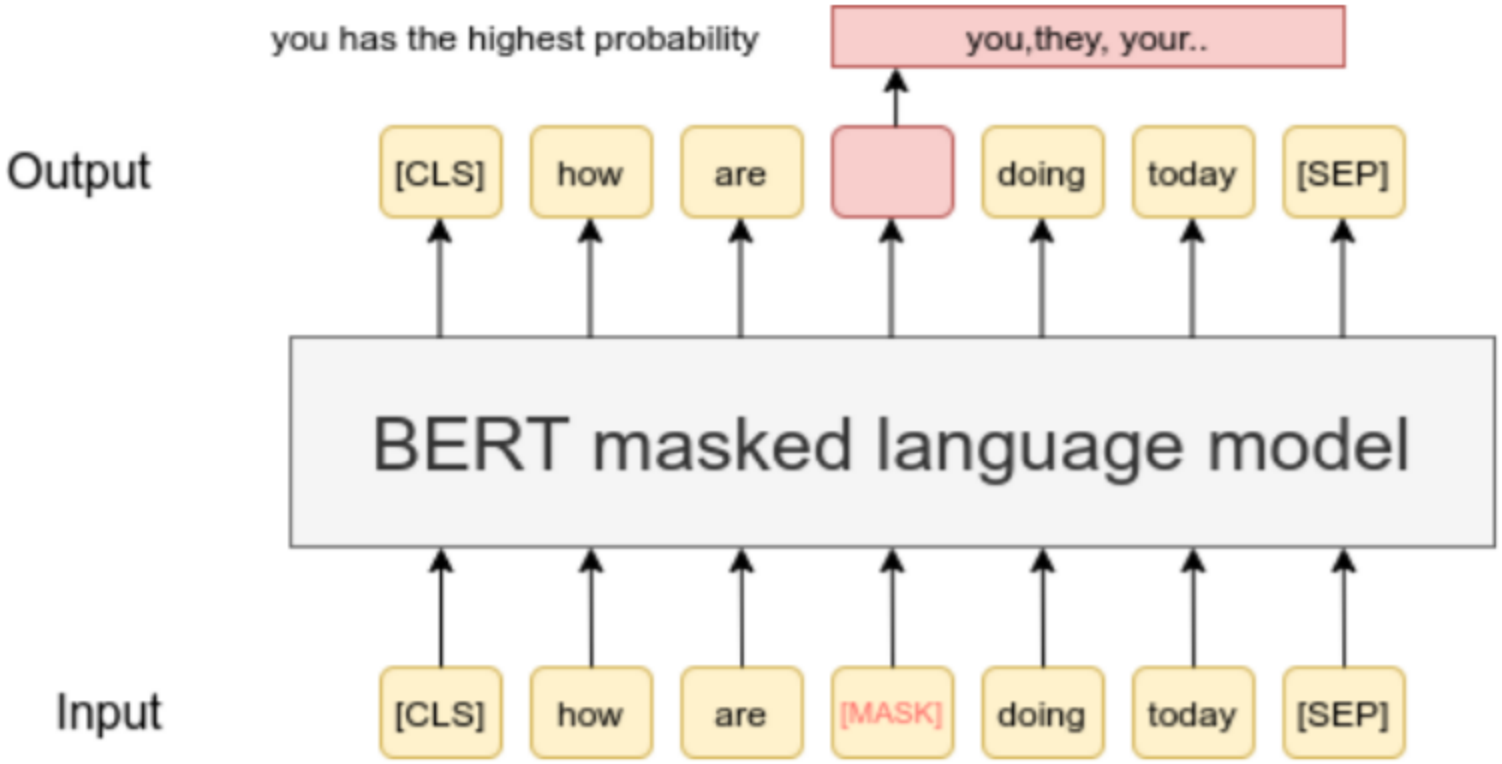
https://arxiv.org/pdf/2104.06979.pdf

Adaptive Pre-Training

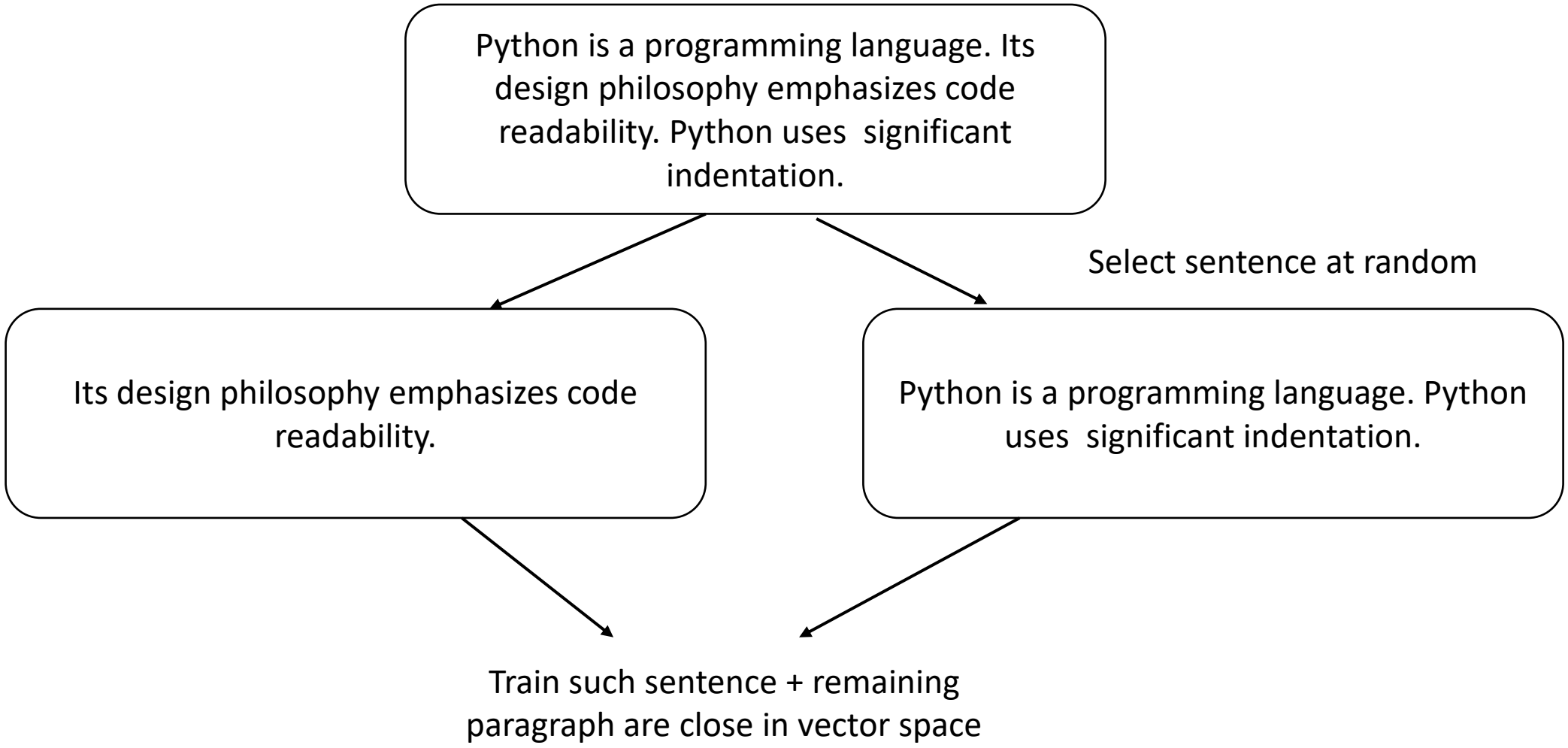


Methods for Pre-Training	**Does it work?**
Masked Language Modeling (MLM)	Yes
TSDAE	Yes
Inverse Cloze Task (ICT)	Yes
SimCSE	No – weaker than base model
Contrastive Tension (CT)	No – weaker than base model
Condenser (CD)	No – weaker than base model

Masked Language Model (MLM)

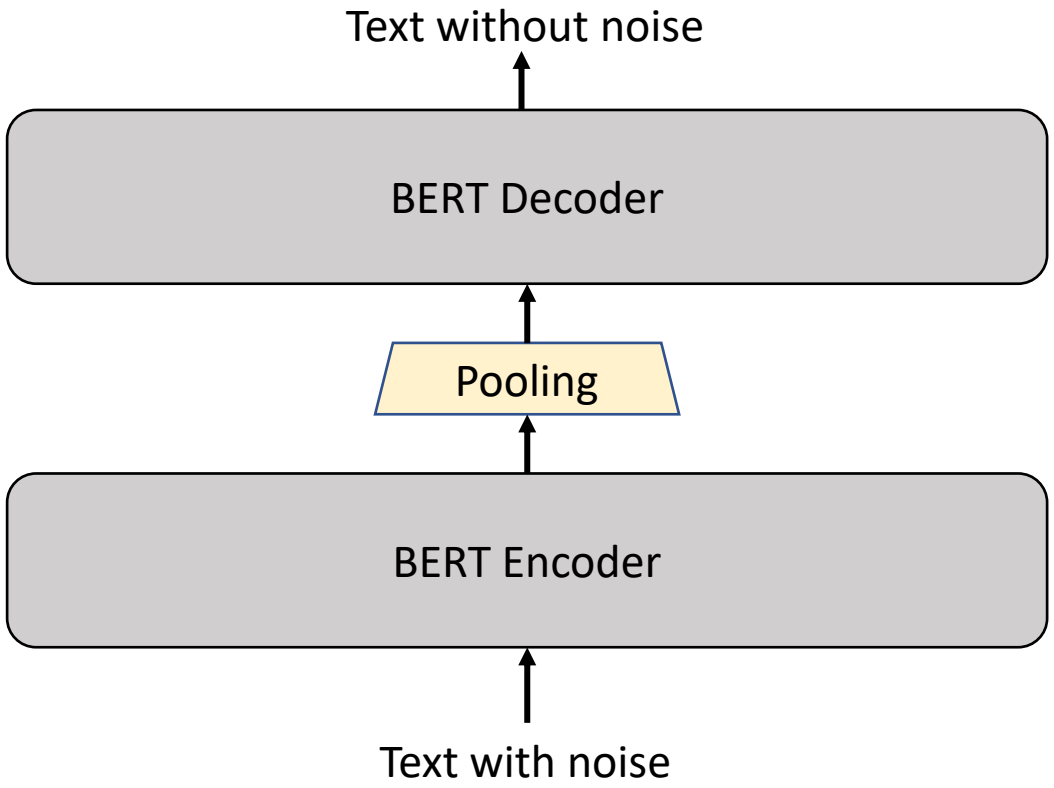


Inverse Cloze Task (ICT)



https://aclanthology.org/P19-1612.pdf

TSDAE



- Delete randomly words in the text
- Pass through the encoder
- Apply pooling to get fixed-sized text embedding
- Decoder must reconstruct text without noise from this text embedding

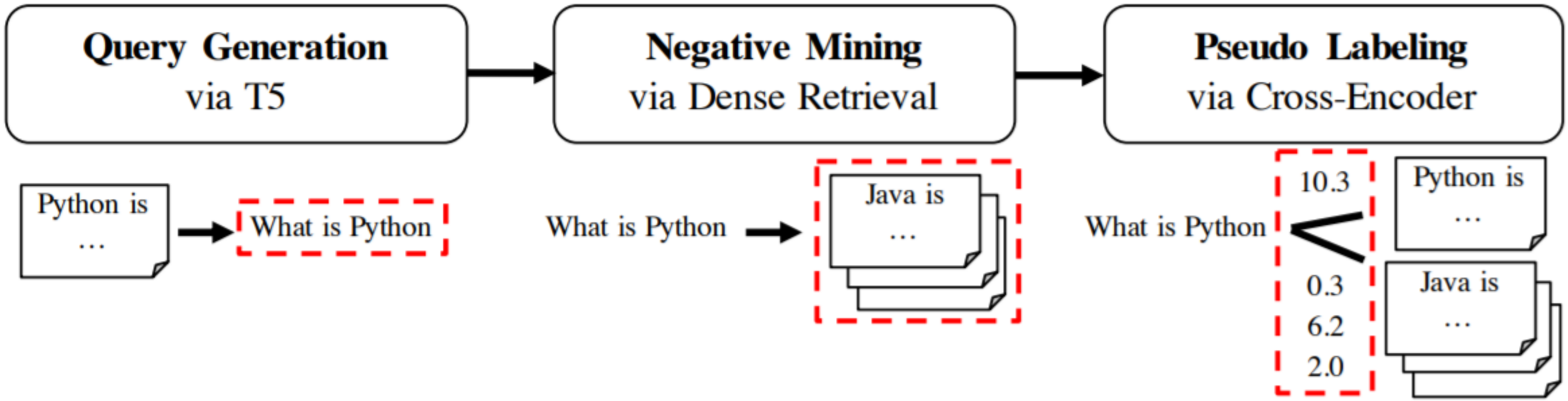
https://arxiv.org/abs/2104.06979

Adaptive Pre-Training - Results

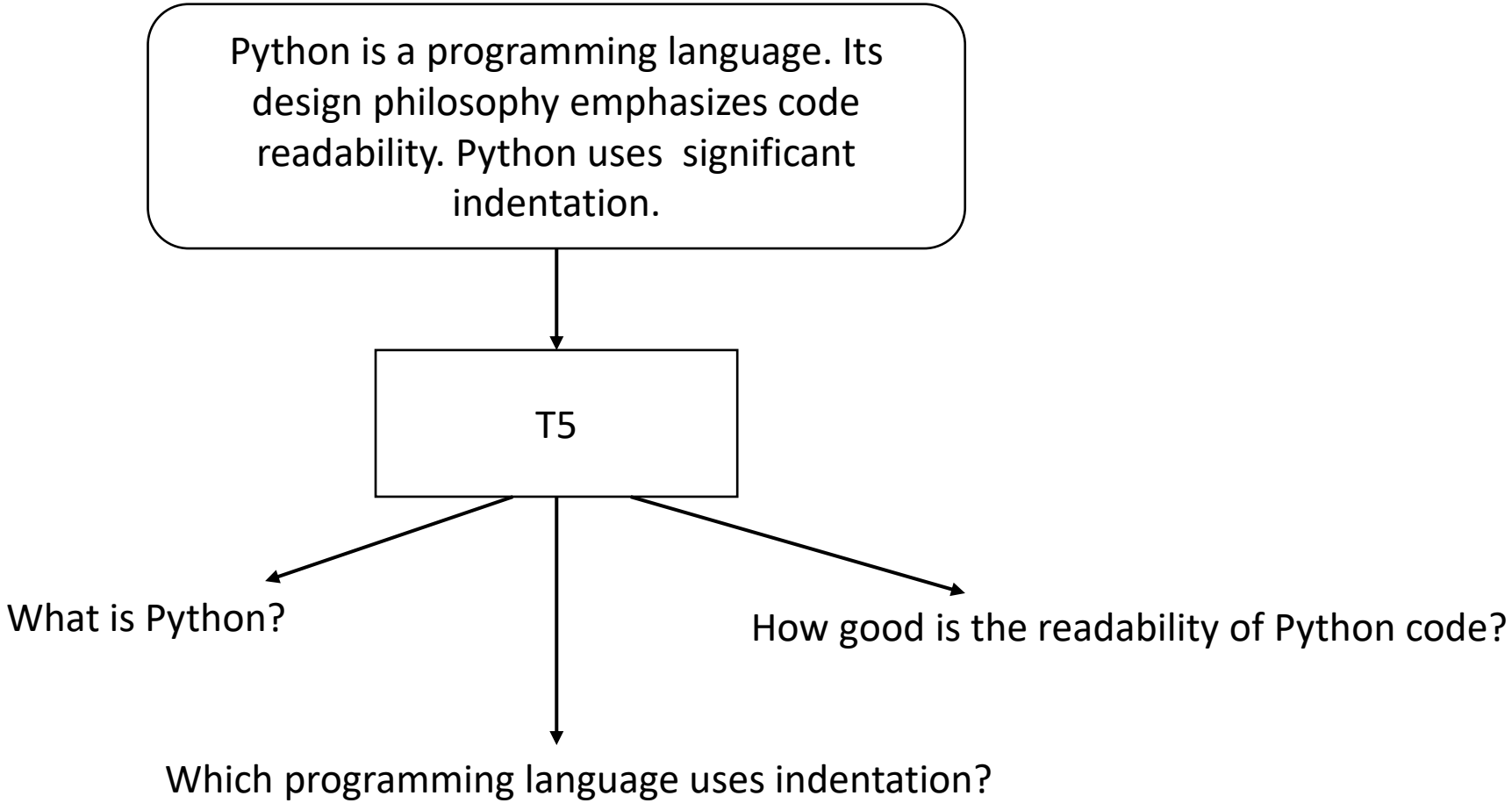
Models	4 Sentence Tasks	6 Dense IR Tasks
Out-of-the-box	52.3	45.2
Source -> Target		
TSDAE	54.2	-
MLM	51.1	-
Target -> Source		
TSDAE	56.5	49.2
MLM	55.9	46.7
ICT	-	46.5
SimCSE	52.4	45.0
CD	-	44.7
CT	53.0	44.0

https://arxiv.org/abs/2104.06979 / https://arxiv.org/abs/2112.07577

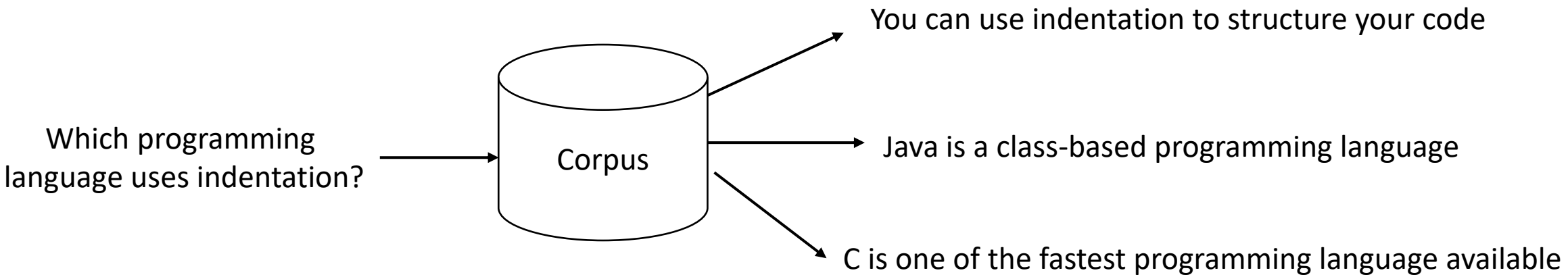
GPL – Generative Pseudo Labeling



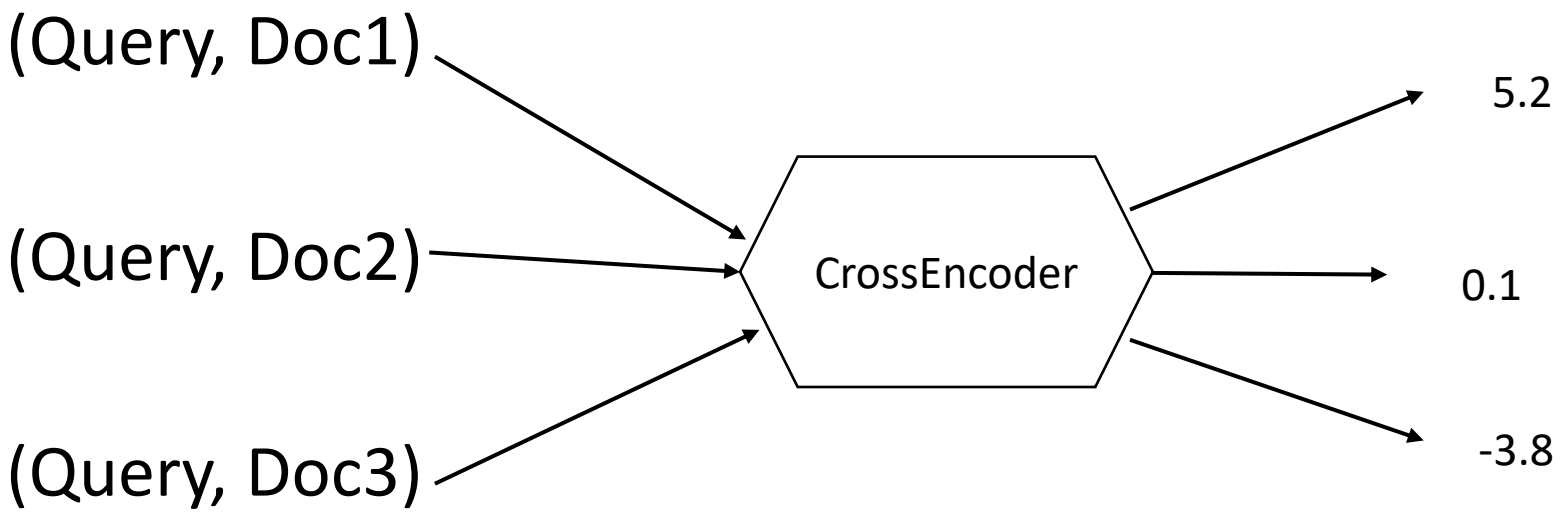
Step 1: Generate Queries



Step 2: Mine Negatives



Step 3: Score Pairs with CrossEncoder



Why do we need the CrossEncoder?

👉**Query** asks for **definition** of “futures contract”

👉**Easy** negatives: Mention “futures contract” only

👉**False** negative

👉**Hard** negative: Give partial definition

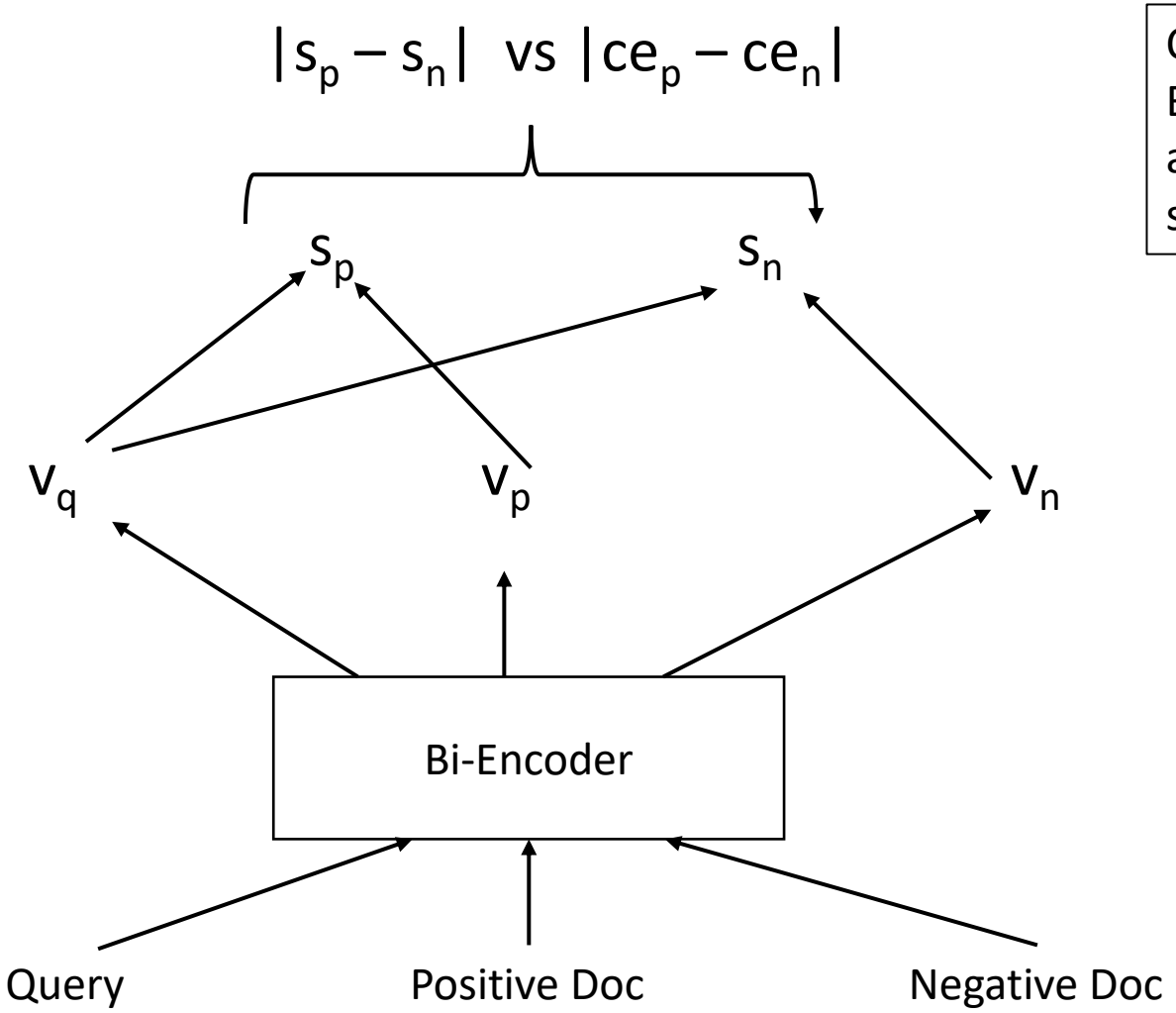
Item	Text	GPL	QGen
Query	what is **futures contract**	–	–
Positive	**Futures contracts** are a member of a larger class of financial assets called derivatives ...	10.3	1
Negative 1	... Anyway in this one example the s&p 500 **futures contract** has an ”initial margin” of $19,250, meaning ...	2.0	0
Negative 2	... but the moment you exercise you must have $5,940 in a margin account to actually use the **futures contract** ...	0.3	0
Negative 3	... a **futures contract** is simply a contract that requires party A to buy a given amount of a commodity from party B at a specified price...	8.2	0
Negative 4	... A **futures contract** commits two parties to a buy/sell of the underlying securities, but ...	6.9	0

Train Bi-Encoder with MarginMSE-Loss

Compute Loss

Compute dot-scores

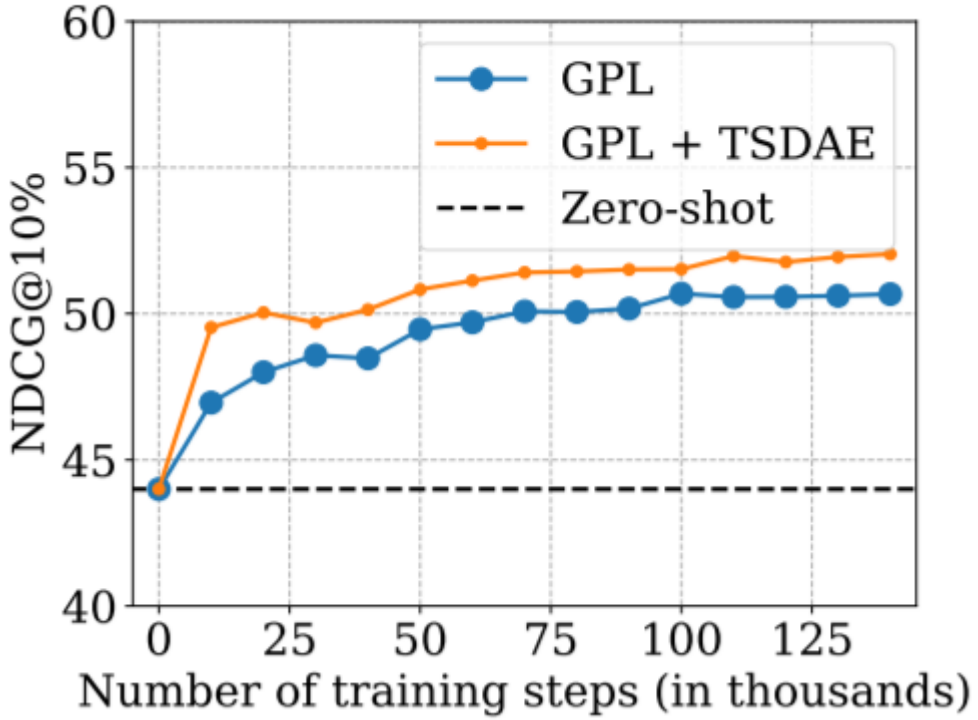
Compute Embeddings



CrossEncoder teaches BiEncoder how far vectors are supposed to be in vector space

Results

Models	6 Dense IR Tasks
Out-of-the-box	45.2
Target -> Source	
TSDAE	49.2
MLM	46.7
Generative Pseudo Labeling	
GPL	51.4
TSDAE+GPL	52.4



Conclusion

- Dense Models have issues with unknown words
  - Unclear how to represent them in a vector space
- Pre-training on target -> Fine-tuning on source domain can help
  - Does not solve this issue fully
  - High computational overhead
- Generative Pseudo Labeling
  - Nice performance increase without need of labeled data
  - Computational overhead still quite high (~1 day on V100 GPU)
- Efficient & continual domain adaptation still an open question mark
- Code:
  - https://domain-adaptation.SBERT.net
  - https://GPL.SBERT.net