Neural Search for Low Resource Scenarios

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

The current Hype

Few-Shot Learning!

Zero-Shot Learning!

State-of-the-art with just 20 examples!

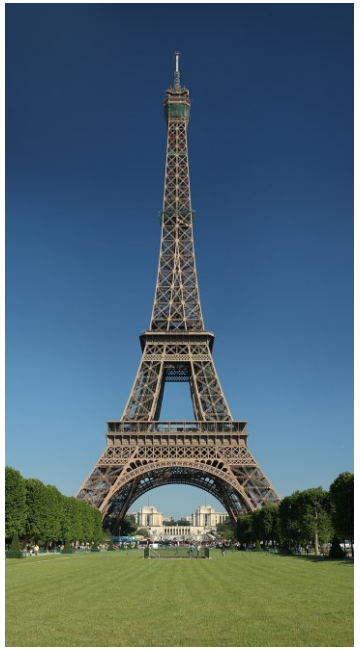
20 seconds training for state-of-the-art sentence embeddings!

Is Low-Resource Training Possible?

- Can we learn good models from just few hundred examples?

- If you want a generic, widely applicable model:
  - => No 😔

Example: Image-Text Matching



A

B

C

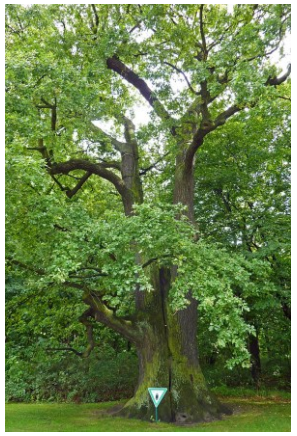
Which image shows the Eiffel tower?

Example: Image-Text Matching

A

B

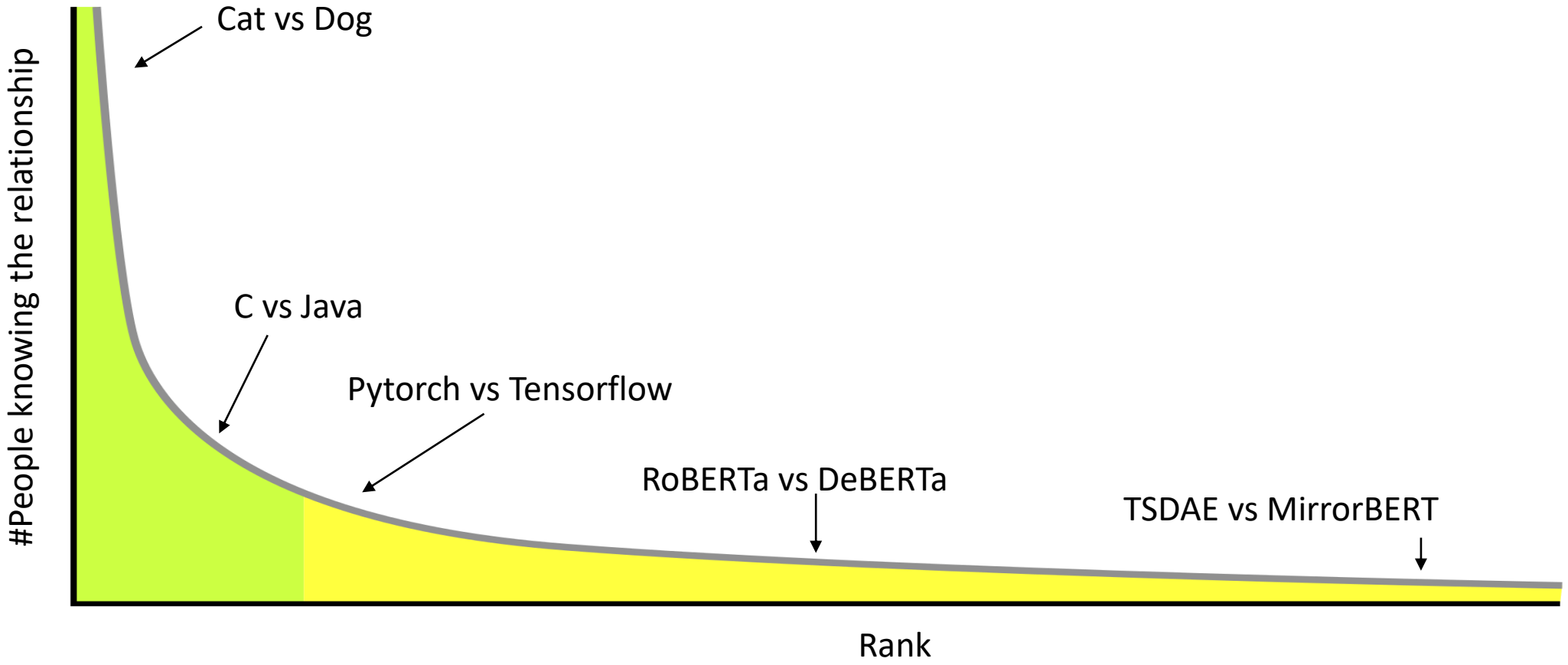
C



D

Which image shows the oak tree in Dötlingen
(small village in Germany I grew up)

The Long Tail of Semantic Relatedness



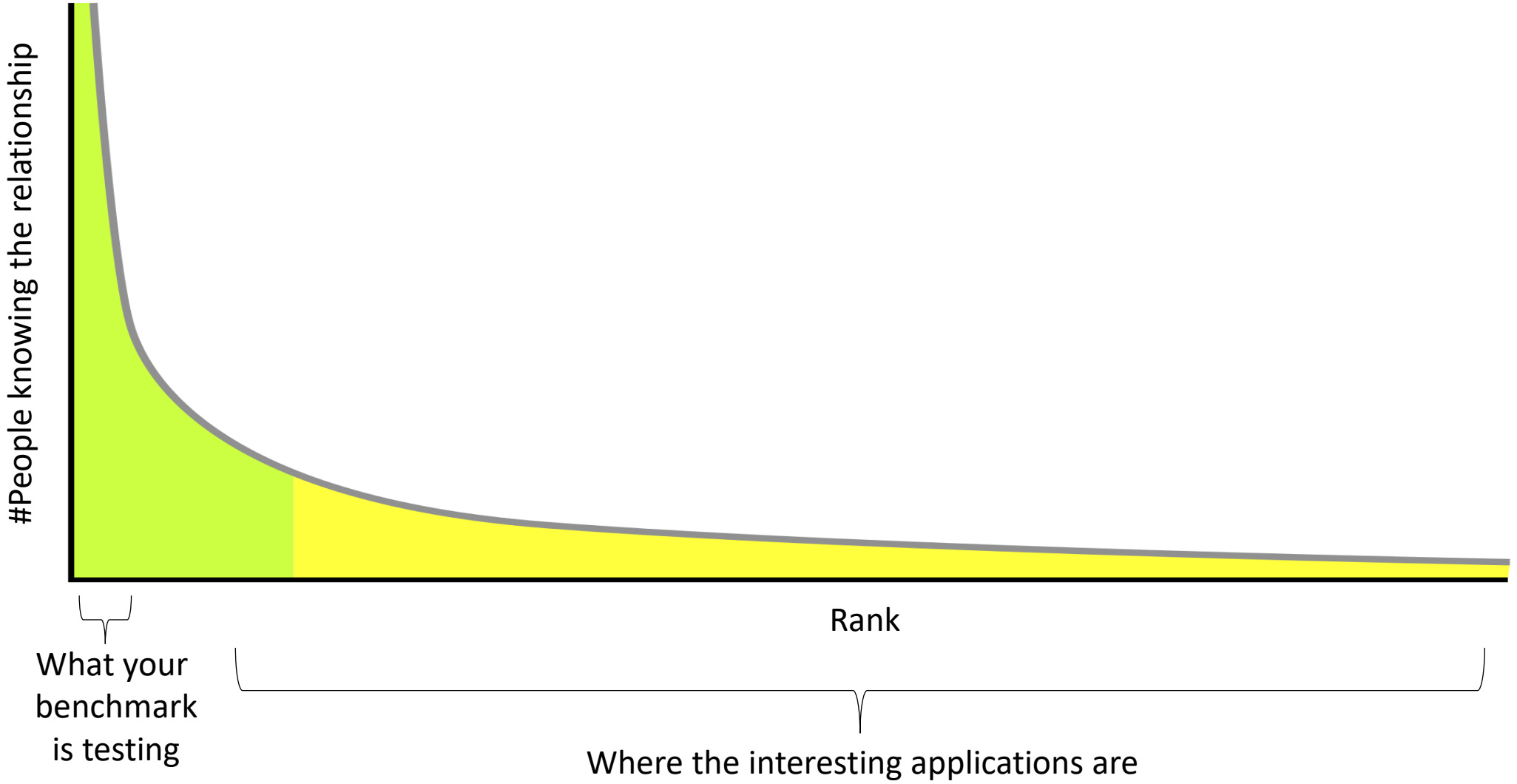
You need data!

- You must have seen an example at least once
  - Never seen the oak tree in Dötlingen?
  - => No chance to identify it among other oak trees

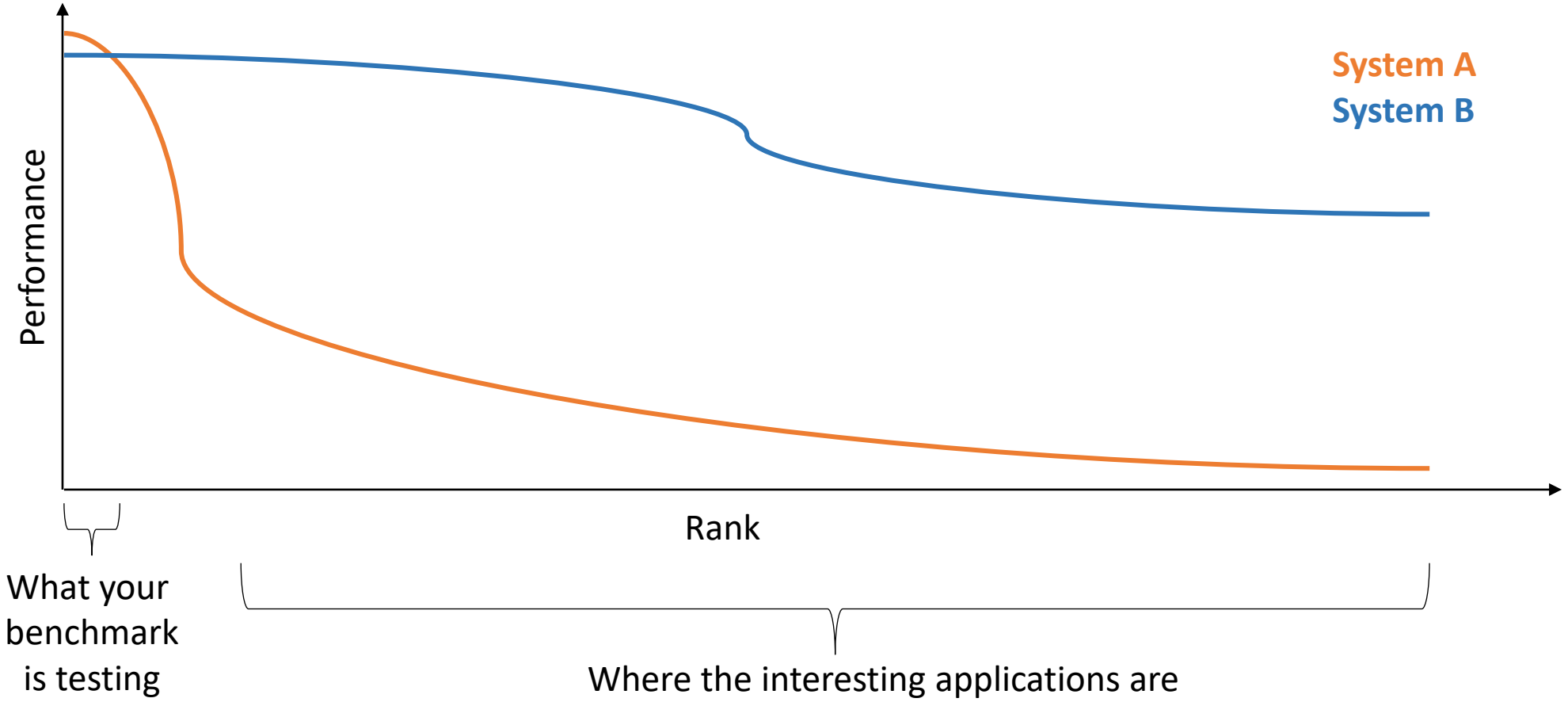
- Our world is so diverse
  - Animals, trees, flowers, leaves, fish
  - Monuments, buildings, statues, natural monuments
  - Food, People, items, cloths
  - Movies, drawings, art

- For a foundation model: A massive dataset is needed
  - Even when it would be sufficient to see each example only once

Be careful with your benchmarks!



Be careful with your benchmarks!

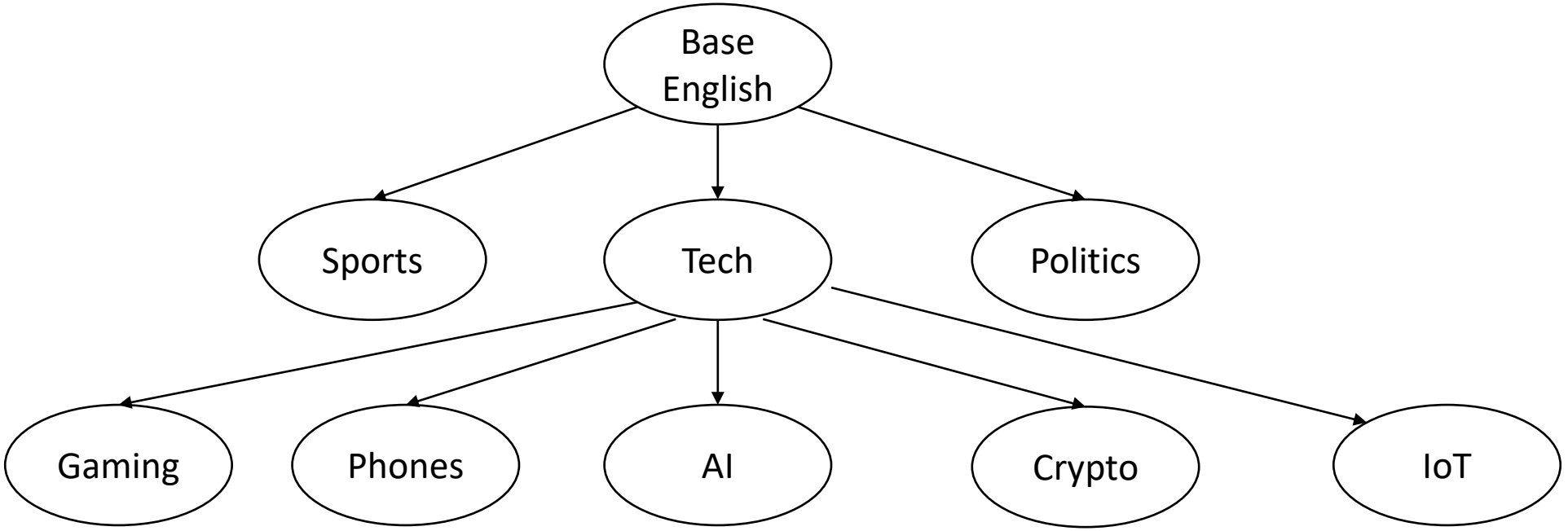


- Academics will optimize for the benchmark => will optimize for System A

We need better benchmarks

- We optimize on what we can measure
- Many (commonly accepted) benchmarks are extremely bad
  - E.g. STS-datasets to measure quality of sentence embedding models
- Most benchmarks are on the short head 😔
  - Easier to get data
  - Easier to annotate
  - You can use cheap student / crowd annotators
  - E.g. annotate relevant hits for “covid-19 symptoms” vs. “impact of PCV2-specific lymphocytes on CD3+ positive T-cells”
- We need diverse long-tail benchmarks!
  - Large performance differences on the long tail
  - The value for many applications / users is in the long tail
- Benchmarks need to evolve
  - Stop overfitting on the same 20-year old benchmark

The Challenge of Long-Tail Benchmarking



- Number of topics grows exponentially in the long tail
- Required expertise for annotators grows
- How many topics is you benchmark checking?

Training Procedure vs. Training Data

Training Set	**Training Procedure**	**MS MARCO Retrieval Performance**	**Out-of-domain Retrieval Performance**
MS MARCO (500k pairs)	Simple Method	68.3	38.8
MS MARCO (500k pairs)	Extremely Sophisticated Method (TAS-B)	70.4	44.0

Training Procedure vs. Training Data

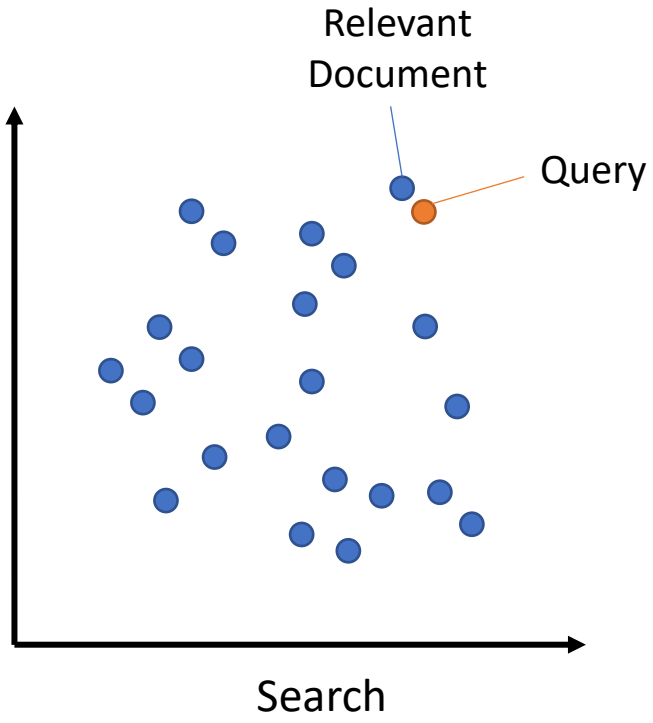
Training Set	**Training Procedure**	**MS MARCO Retrieval Performance**	**Out-of-domain Retrieval Performance**
MS MARCO (500k pairs)	Simple Method	68.3	38.8
MS MARCO (500k pairs)	Extremely Sophisticated Method (TAS-B)	70.4	44.0
Multiple sources (200M pairs) (5 days of work)	Extremely Simple Method	70.9	54.3

Summary – Part I

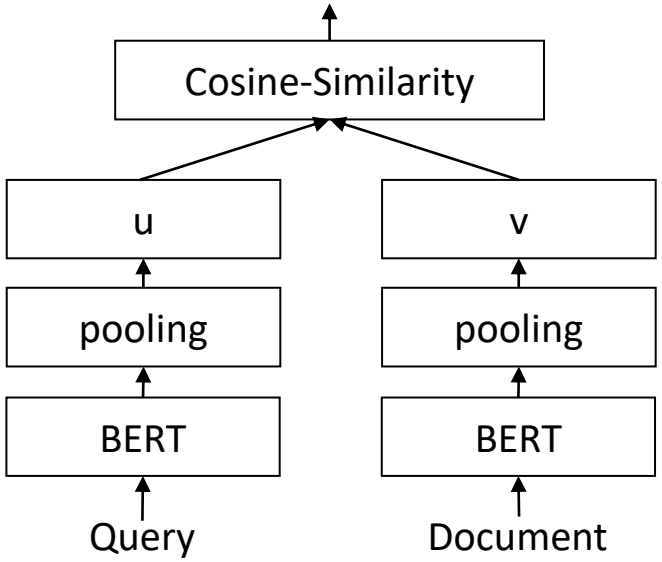
- True low resource learning is not realistic
  - World has too many facets
  - You have to see something at least once
- Important research questions:
  - How to learn unsupervised?
  - How to exploit structure in our data (like title & body, text & image)?
  - Data efficiency: How can we learn a concept from a single instance?
- We need better benchmarks
  - Must evolve with train set & models
  - Must check the long tail
- We need more & better datasets
  - Improving datasets improves models often more than tuning the architecture
  - Especially non-English datasets are needed

Part II: Domain-Adaptation for Text Embeddings and Neural Search

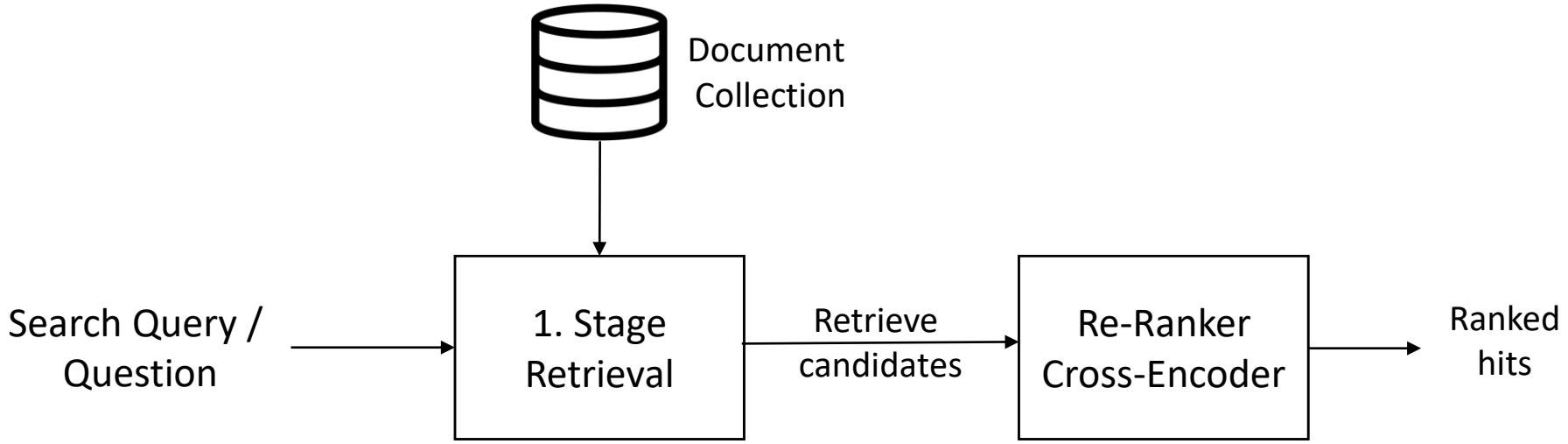
Neural Search – Using Dense Vectors

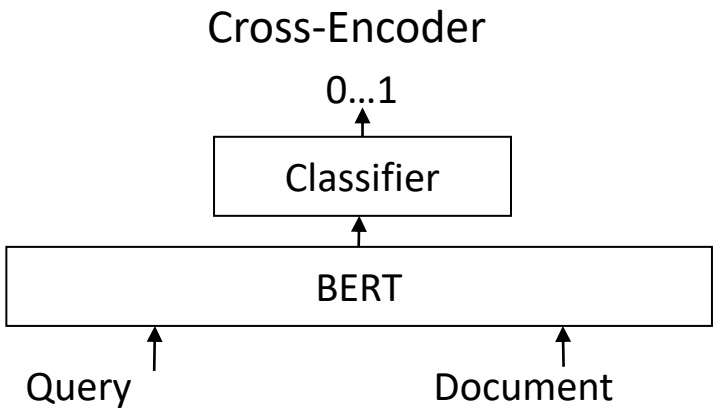


Bi-Encoder

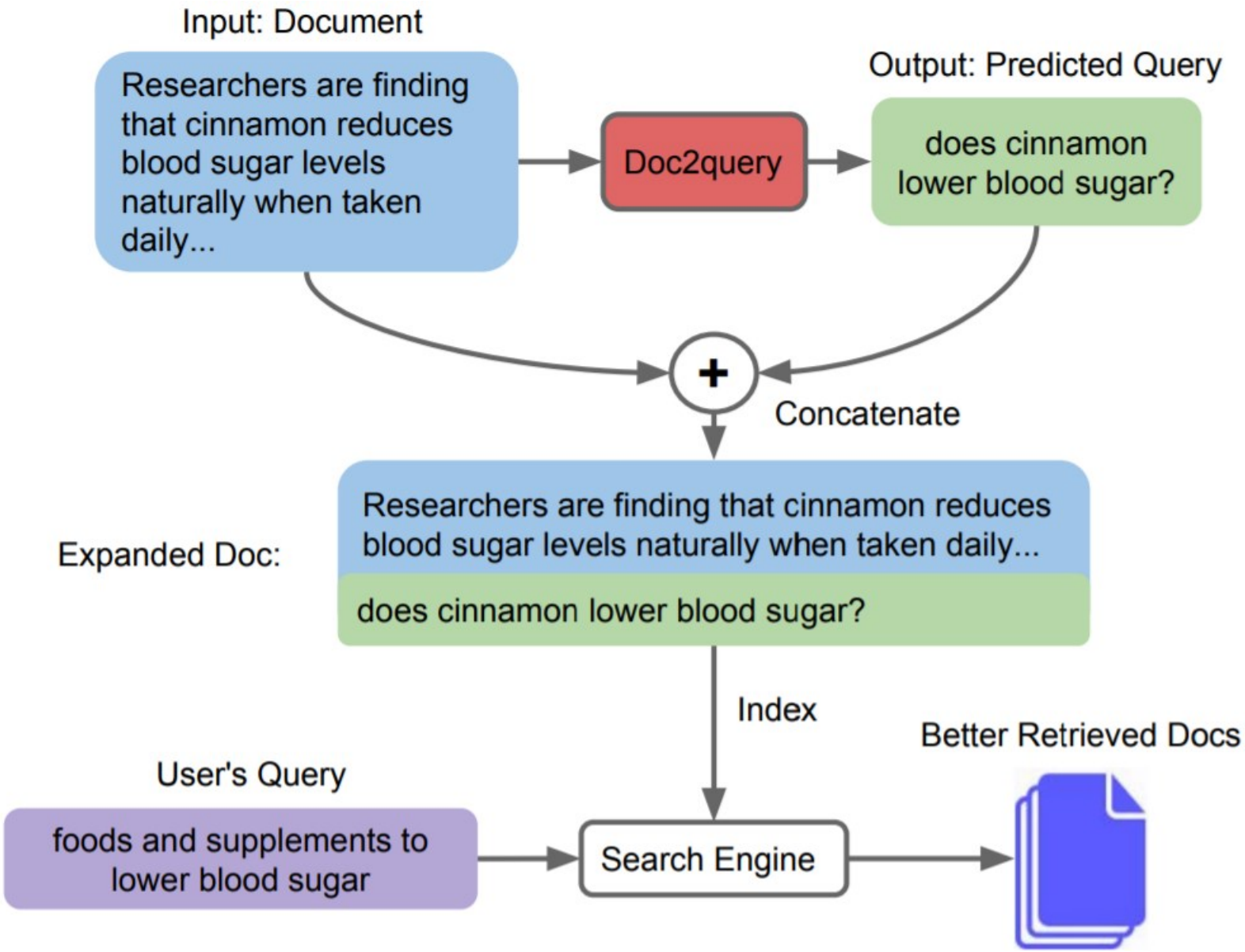


Neural Re-Rankers



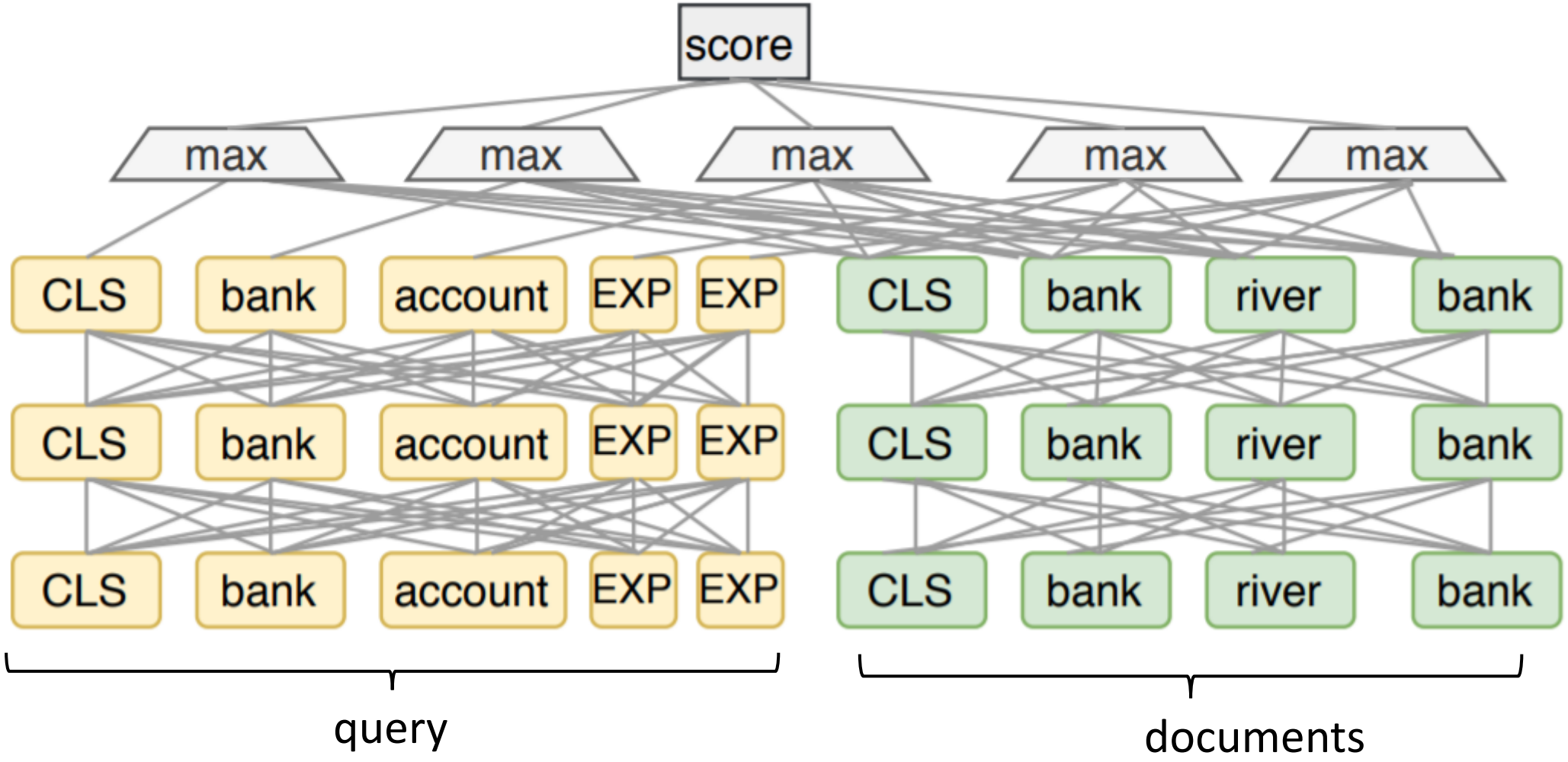


doc2Query



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

ColBERT



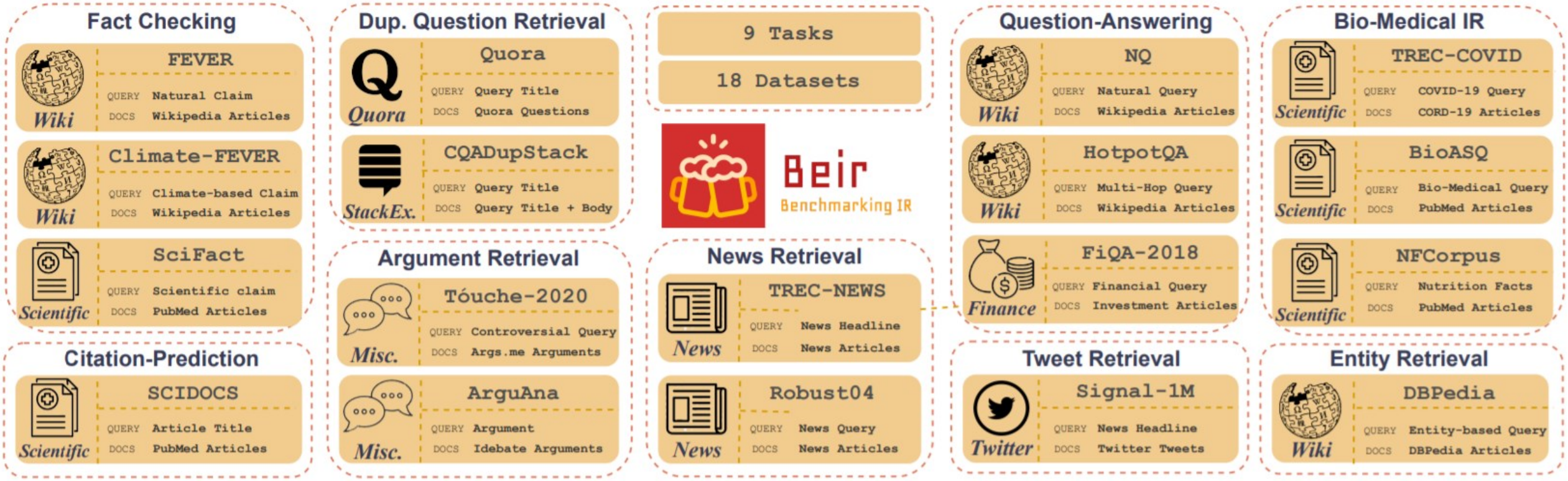
Img: https://arxiv.org/pdf/2104.07186.pdf

Neural Retrieval Is Great?

- Large amount of training data needed (100k+ pairs)
  - Only suitable for big companies 😢
  - Only suitable for already popular products 😢
  - New companies, academics, niche use cases are left out 😢

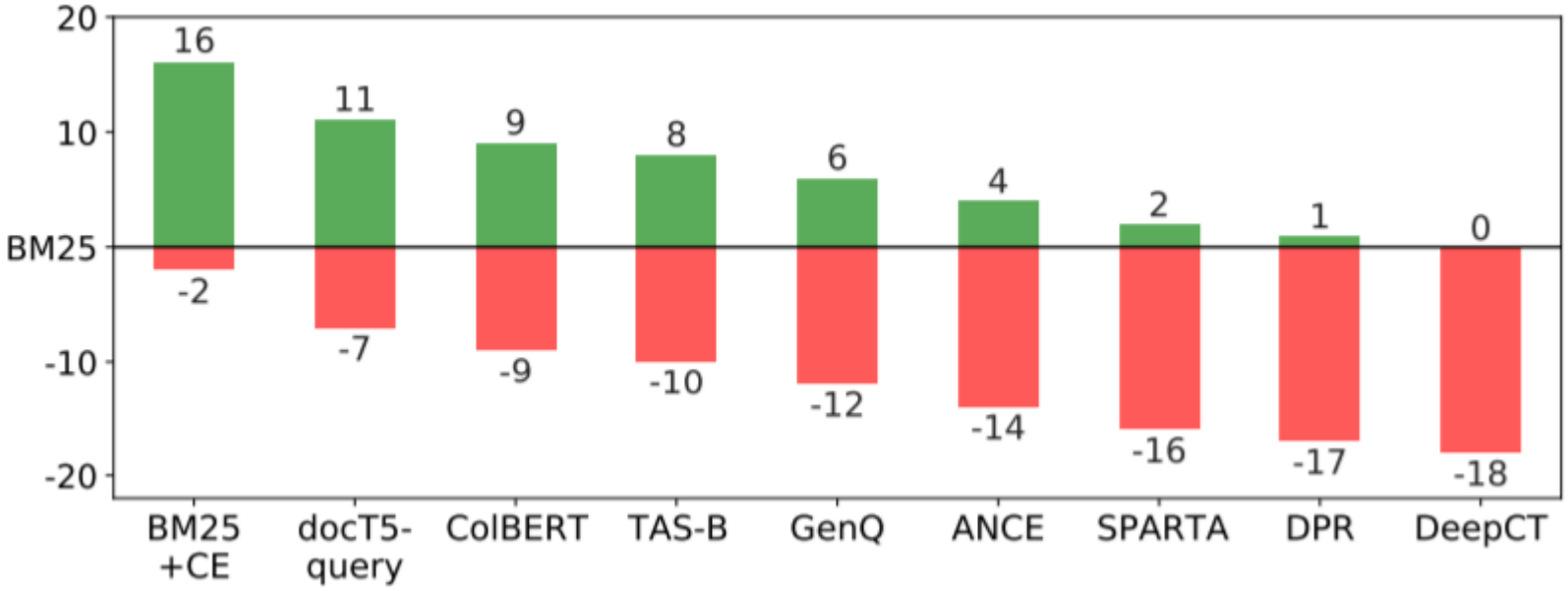
- Most domains / tasks / languages without large training sets
  - How well do models **generalize** to new domains / tasks?
  - How to improve performance from **unlabeled data**?

BEIR – Benchmarking IR



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

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

IR Benchmarking is difficult

- Only tiny fraction of (query, document) pairs annotated
- Other pairs: Assumed to be irrelevant
- Lexical models used to create annotation pool
- How many un-annotated docs are systems retrieving?

Model (→)	**BM25**	**DeepCT**	**SPARTA**	**docT5query**	**DPR**	**ANCE**	**TAS-B**	**ColBERT**	**BM25+CE**
Hole@10 (in %)	6.4%	19.4%	12.4%	2.8%	30.6%	14.4%	31.8%	12.4%	1.6%
nDCG@10 performances before and after manual annotation on TREC-COVID [65]									
Original (w/ holes)	0.656	0.406	0.538	0.713	0.332	0.654	0.481	0.677	**0.757**
Annotated (w/o holes)	0.668	0.472	0.624	0.714	0.445	0.735	0.555	0.735	**0.760**

Number of top-10 hits retrieved by systems that were not annotated on TREC-COVID-19 dataset

There is no Free Lunch

DBPedia [21] (1 Million)			Retrieval Latency		Index
Rank	Model	Dim.	GPU	CPU	Size
(1)	BM25+CE	–	450ms	6100ms	0.4GB
(2)	ColBERT	128	350ms	–	20GB
(3)	docT5query	–	–	30ms	0.4GB
(4)	BM25	–	–	20ms	0.4GB
(5)	TAS-B	768	14ms	125ms	3GB
(6)	GenQ	768	14ms	125ms	3GB
(7)	ANCE	768	20ms	275ms	3GB
(8)	SPARTA	2000	–	20ms	12GB
(9)	DeepCT	–	–	25ms	0.4GB
(10)	DPR	768	19ms	230ms	3GB

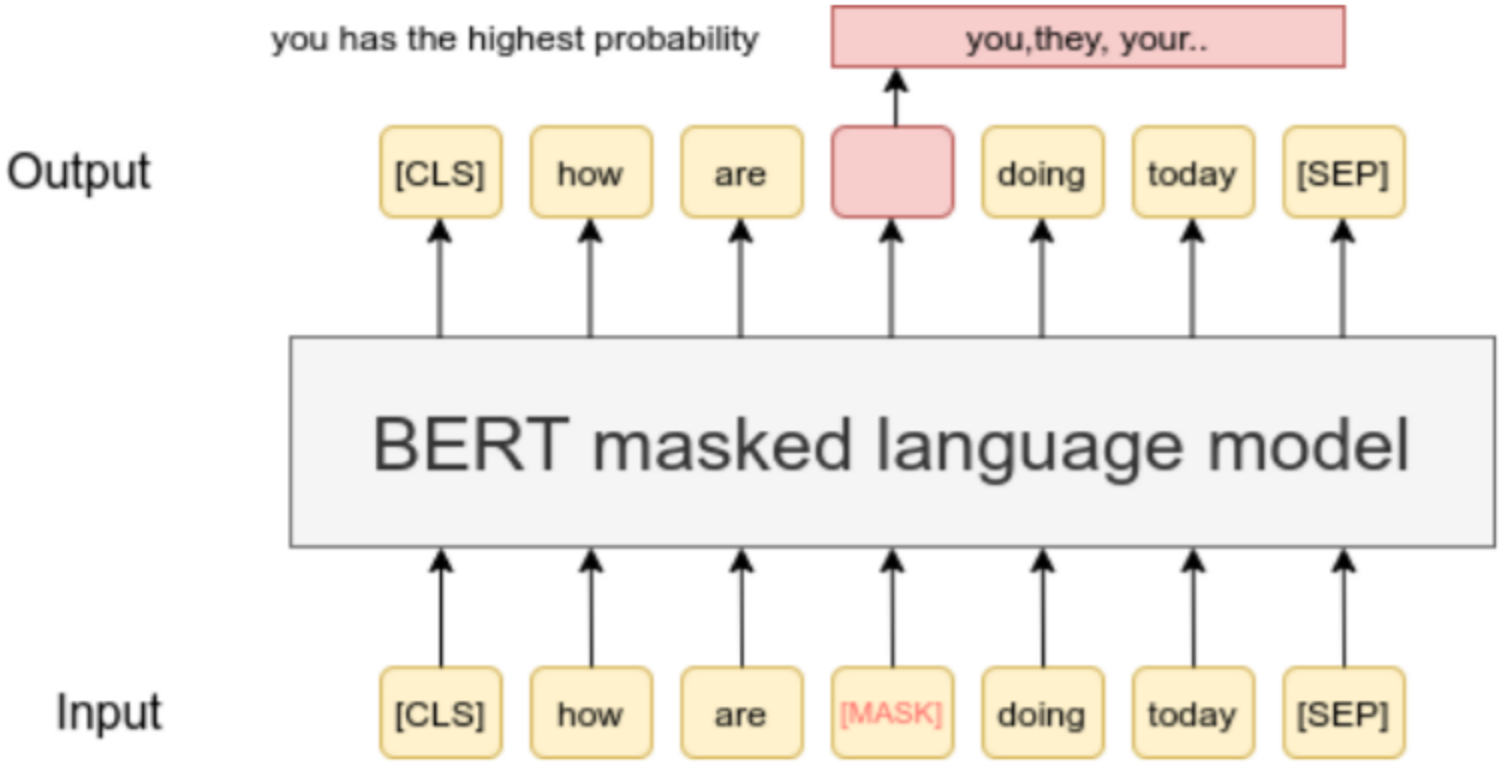
Table 3: Estimated average retrieval latency and index sizes for a single query in DBPedia [21]. Ranked from best to worst on zero-shot BEIR. Lower the latency or memory is desired.

- Strong models:
  - CE: Slow at inference
  - ColBERT: Slow at infernece + large memory overhead
  - docT5query: Extremly slow indexing
- Dense Embedding Models:
  - Efficient
  - Issues with out-of-domain

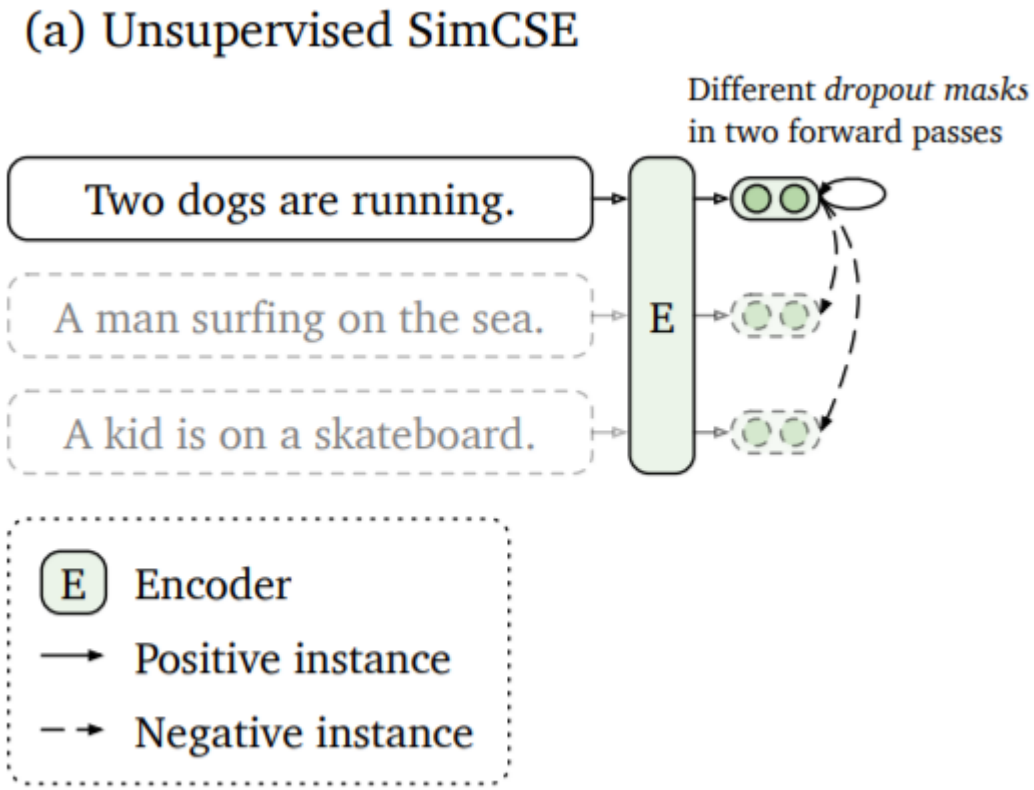
(Single Vector) Embedding Models

- Work well on their training domain
- Fast & efficient
- Has issues on out-of-domain data
  - New word – not see during training
  - Where to locate it in vector space?
- How to perform domain adaption without labeled data?

Masked Language Model (MLM)



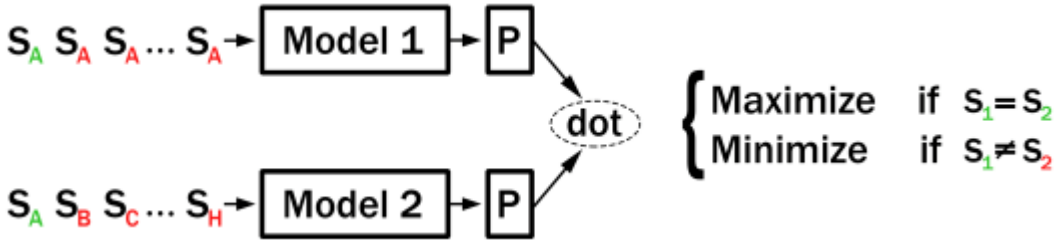
Mirror-BERT / SimCSE



- Usage of MultipleNegativeRankingLoss
- Input pairs:
  (sent1, sent1)
  (sent2, sent2)
  ….
- Due to dropout: slightly different embeddings for f(sent1) and f(sent1)

https://arxiv.org/abs/2104.08027 / https://arxiv.org/abs/2104.08821

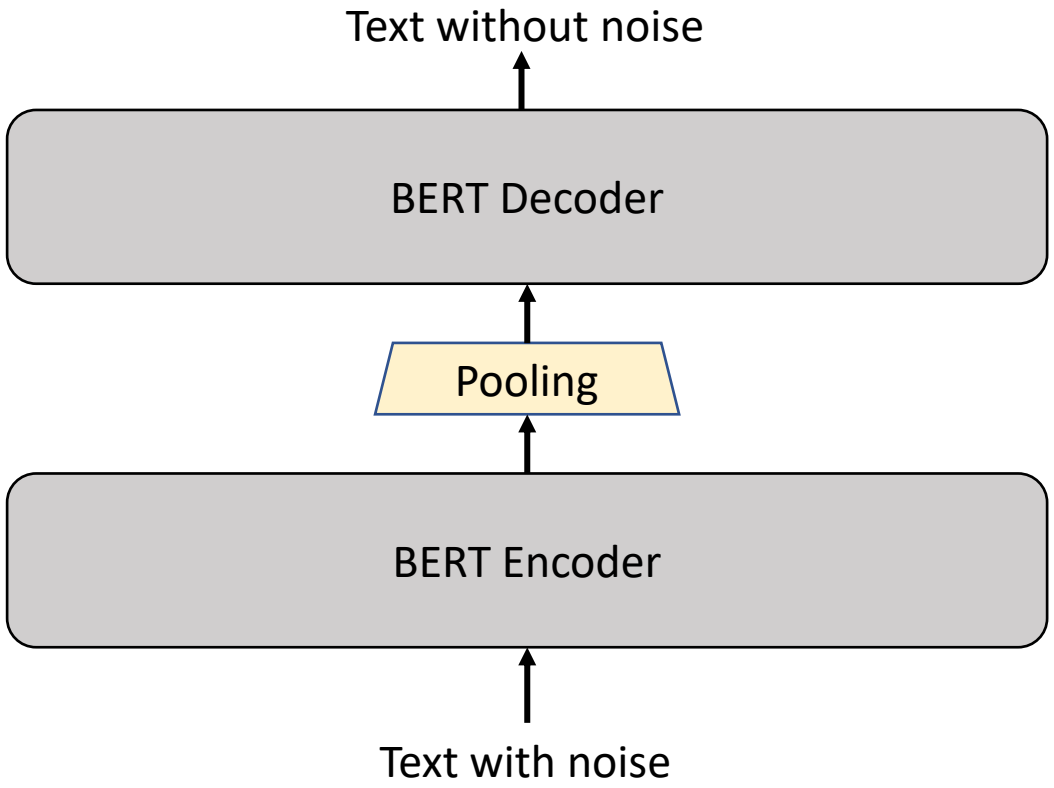
Contrastive Tension (CT)



- Initialize with two identical models
- Pass pairs with identical and with different sentences
- Maximize dot-score for identical sentences
- Minimize dot-score for different sentences

https://openreview.net/pdf?id=Ov_sMNau-PF

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

Issues in the Evaluation

- So far unsupervised methods evaluated on STS data

- Extremely bad way to evaluate unsupervised methods on STS datasets
  - Performance has near zero correlation to performance on real-world task
  - Simple sentences without domain specific knowledge
  - Unrealistic label distribution

- In TSDAE: Evaluation on domain specific datasets
  - AskUbuntu, StackExchange, Twitter, Scientific Publications

Evaluation

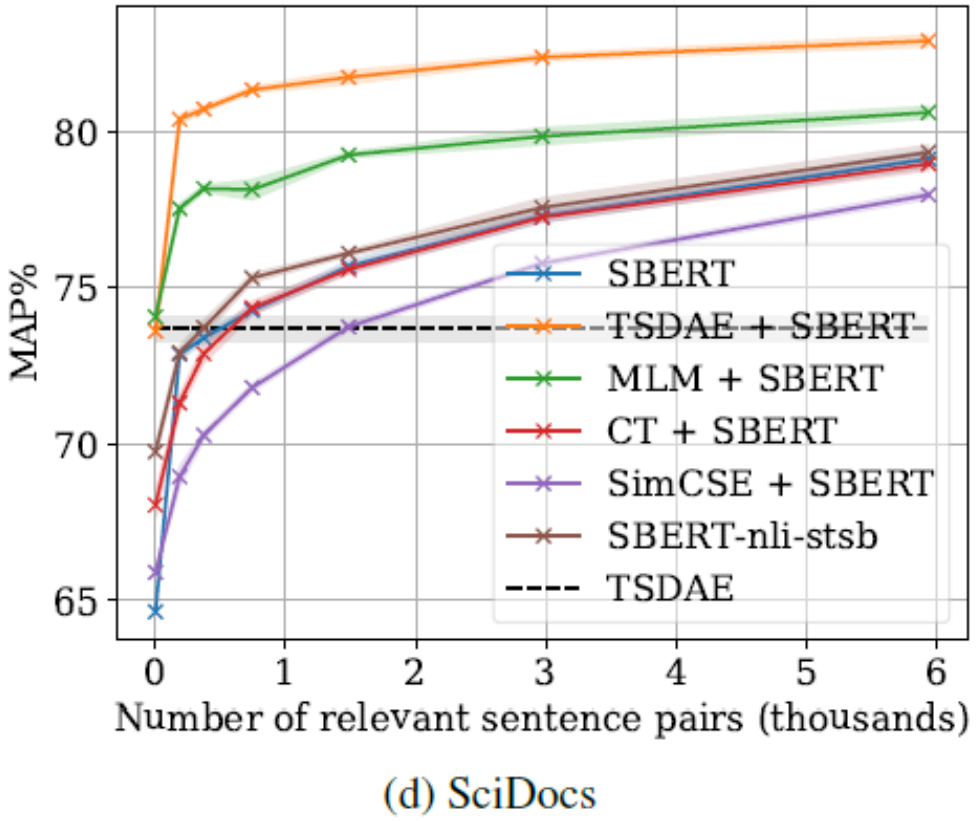
Method	**Avg. over 4 datasets**
TSDAE	55.2
MLM	52.9
CT	52.4
SimCSE	50.6
Out-of-the-box model	
SBERT on NLI+STSb	52.3

How good are unsupervised methods?

	AskUbuntu	Twitter Paraphrases	StackExchange	SciDocs
Unsupervised in-domain TSDAE on bert-base	55.6	74.1	36.2	**74.5**
Supervised out-of-domain mpnet + NLI + STSb	56.0	**78.9**	35.7	71.4
Supervised out-of-domain distilbert + MS MARCO	**56.1**	74.6	**40.3**	70.8

- Supervised pre-trained models hard to beat
- Diversity of pre-training dataset critical
  - Large, diverse dataset => great results across tasks

Unsupervised Method for Pre-Training



(d) SciDocs

- Train unsupervised on large corpus from your domain
- Train supervised with some labels from your domain
- SimCSE / CT: Not helpful
- TSDAE / MLM: Big improvement

TSDAE – Domain Adaptation Technique

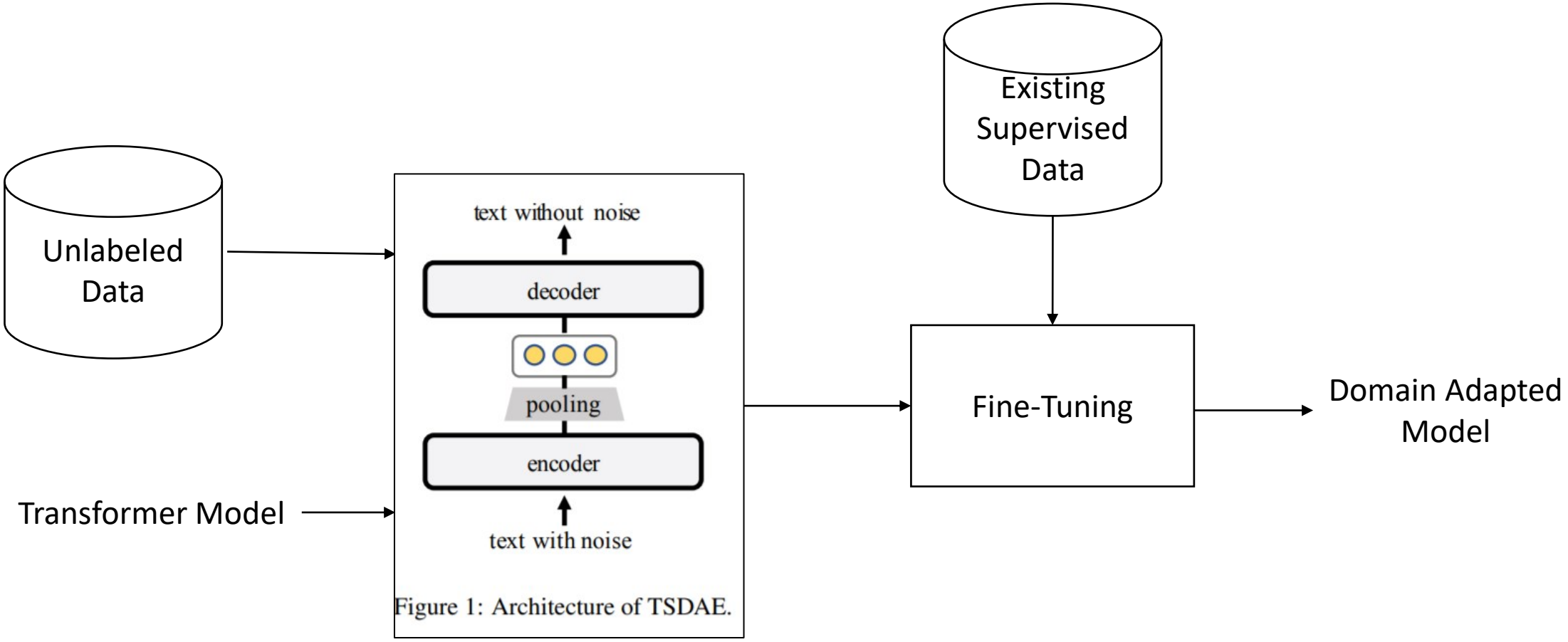


Figure 1: Architecture of TSDAE.

TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning, https://arxiv.org/abs/2104.06979

Domain Adaptation

Method	**Unsupervised**	**NLI+STS -> Unsupervised**	**Unsupervised -> NLI+STS**
TSDAE	55.2	54.2	**56.5**
MLM	52.9	51.1	**55.9**
CT	52.4	52.9	**53.0**
SimCSE	50.6	51.2	**52.4**
Baseline (NLI+STS)	52.3		

- First train unsupervised on your domain
- Then train supervised on available training data from other domains

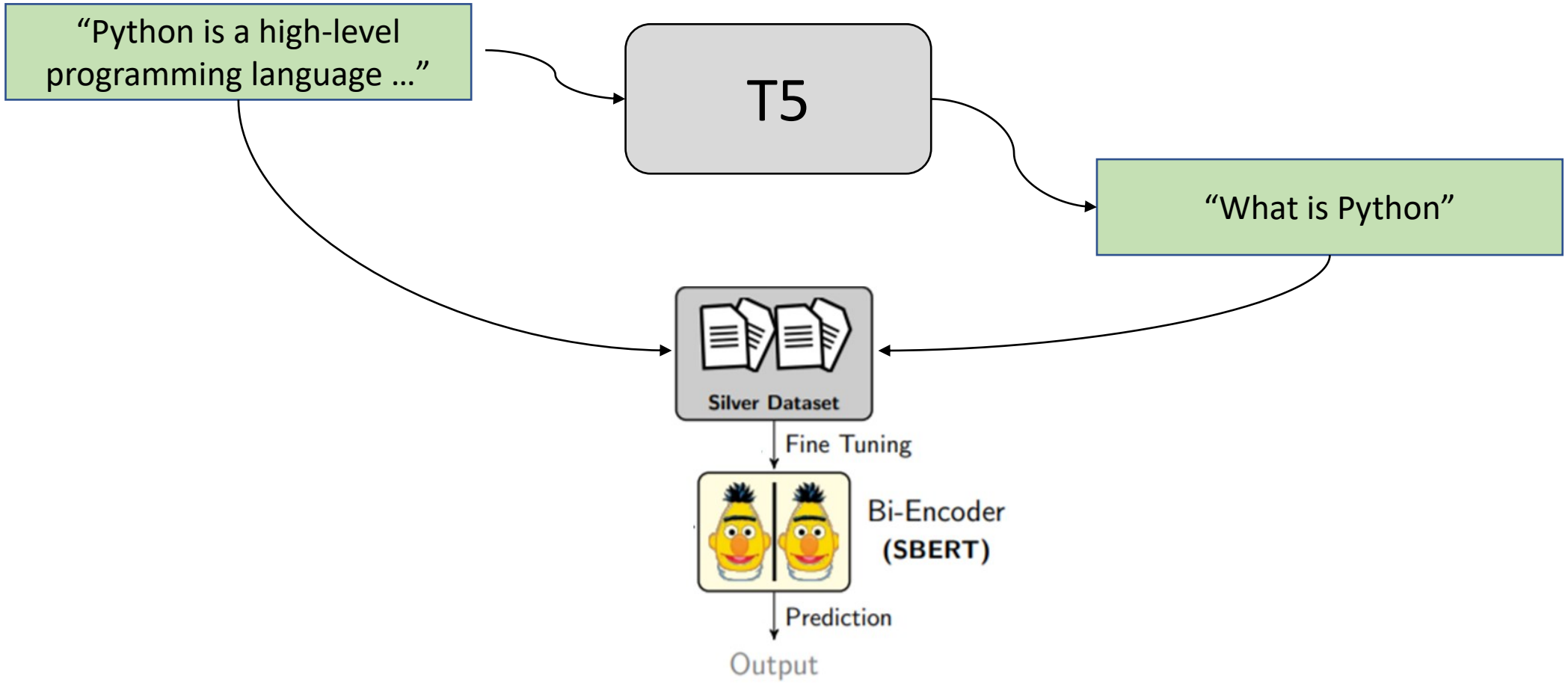
Domain Adaptation – Semantic Search

Pre-training method	Avg. on 6 tasks (NCDG@10)
TSDAE	48.7
Inverse-Cloze-Task (ICT)	46.2
MLM	45.8
CT	44.5
SimCSE	44.4
No Pre-Training	44.0
ConDensor (CD)	43.9

- For given query / question, find relevant docs
- Pre-training on specialized domain
- Supervised training on MS MARCO

Upcoming work by Kexin Wang

Upcoming work: GPL



Upcoming work by Kexin Wang

GPL for Domain Adaptation

- Generate queries for docs in your domain
- Fine-tune bi-encoder
- Improves performances 4 – 10 points

Model	**Avg. on 6 tasks (NCDG@10)**
MS MARCO	44.0
TSDAE (domain specific) + MS MARCO	48.7
New method: GPL	
GPL	50.4
TSDAE (domain specific) + GPL	51.5

Upcoming work by Kexin Wang

Summary – Part II

- Unsupervised Sentence Embeddings
  - Does not work that well (yet)
  - Worse performance than out-of-the-box models which were trained on diverse data
  - Quickly converge (MirrorBERT: “within 20 seconds”)
    - Does not matter if you have 1k or 1B sentences from your domain
    - does not really learn anything new about your domain
- Domain Adaption
  - First unsupervised training, then supervised training
  - TSDAE > ICT > MLM > CT > SimCSE/MirrorBERT > CD
  - Unclear how to adapt an existing model to a new domain
- Query Generation with GPL promising (upcoming work)
  - Strong improvements

Conclusions

- Work on your data - Better datasets needed
- Better benchmarks needed
  - Many papers overfit with narrow evaluation on the short head
  - Diverse, long-tail evaluation needed as model evolve
- Unsupervised Learning is the future
  - So far great results with supervised data (e.g. query & relevant doc)
  - Supervised training: Hard to scale to many domains / new domains etc.
- How can we learn the most from as little structure as possible?
  - Annotated data (query, relevant doc)
  - Mined data (Q&A pages from websites)
  - Website Title & Body
  - Individual documents
  - Ind. paragraphs
  - Ind. sentences