Fine-tuning Embedding Models

Where and how to fine-tune models for dense retrieval

Hello! 👋

James Briggs

Staff Developer Advocate Pinecone (we're hiring)

Educator

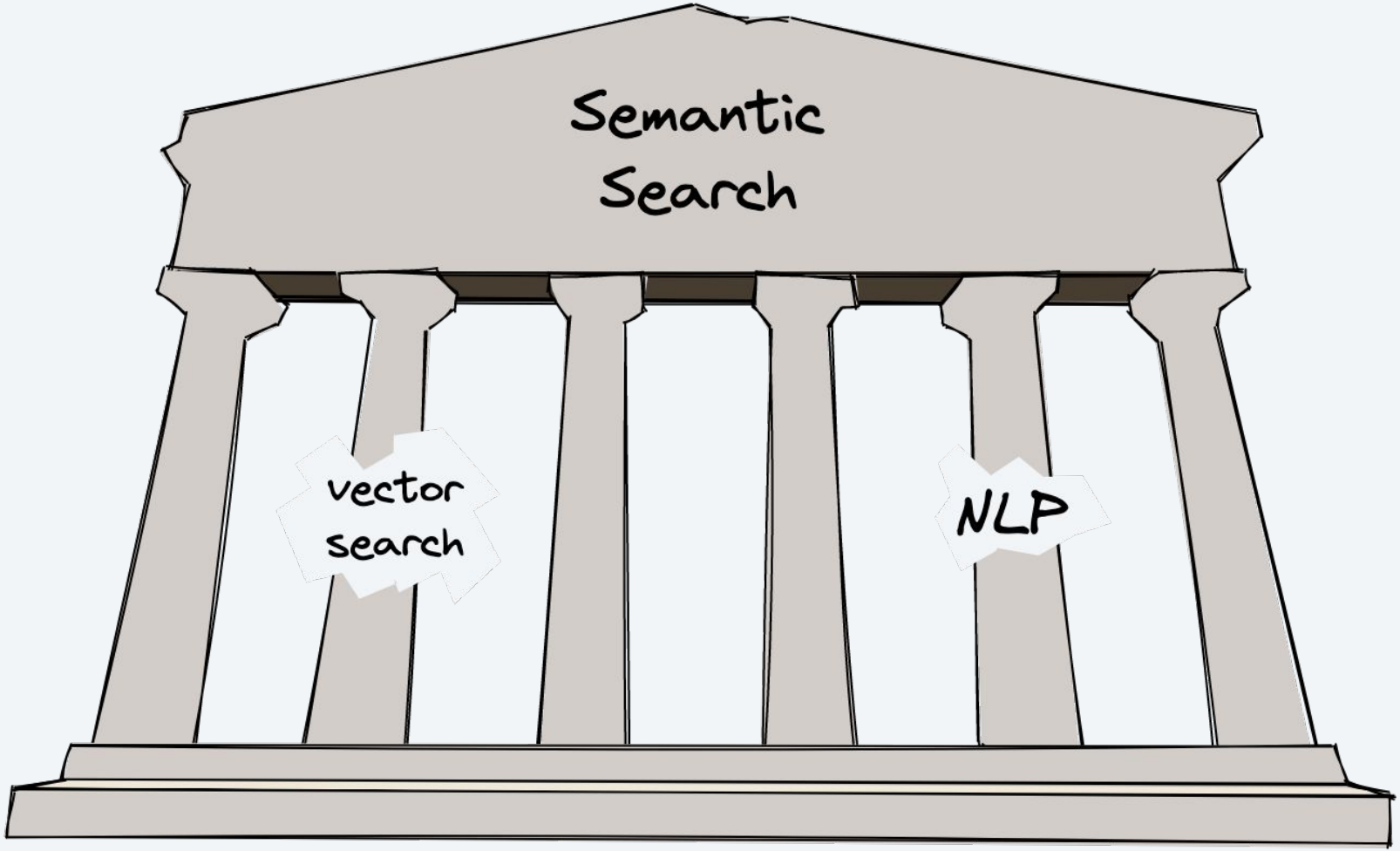
Director Aurelio Consulting aio

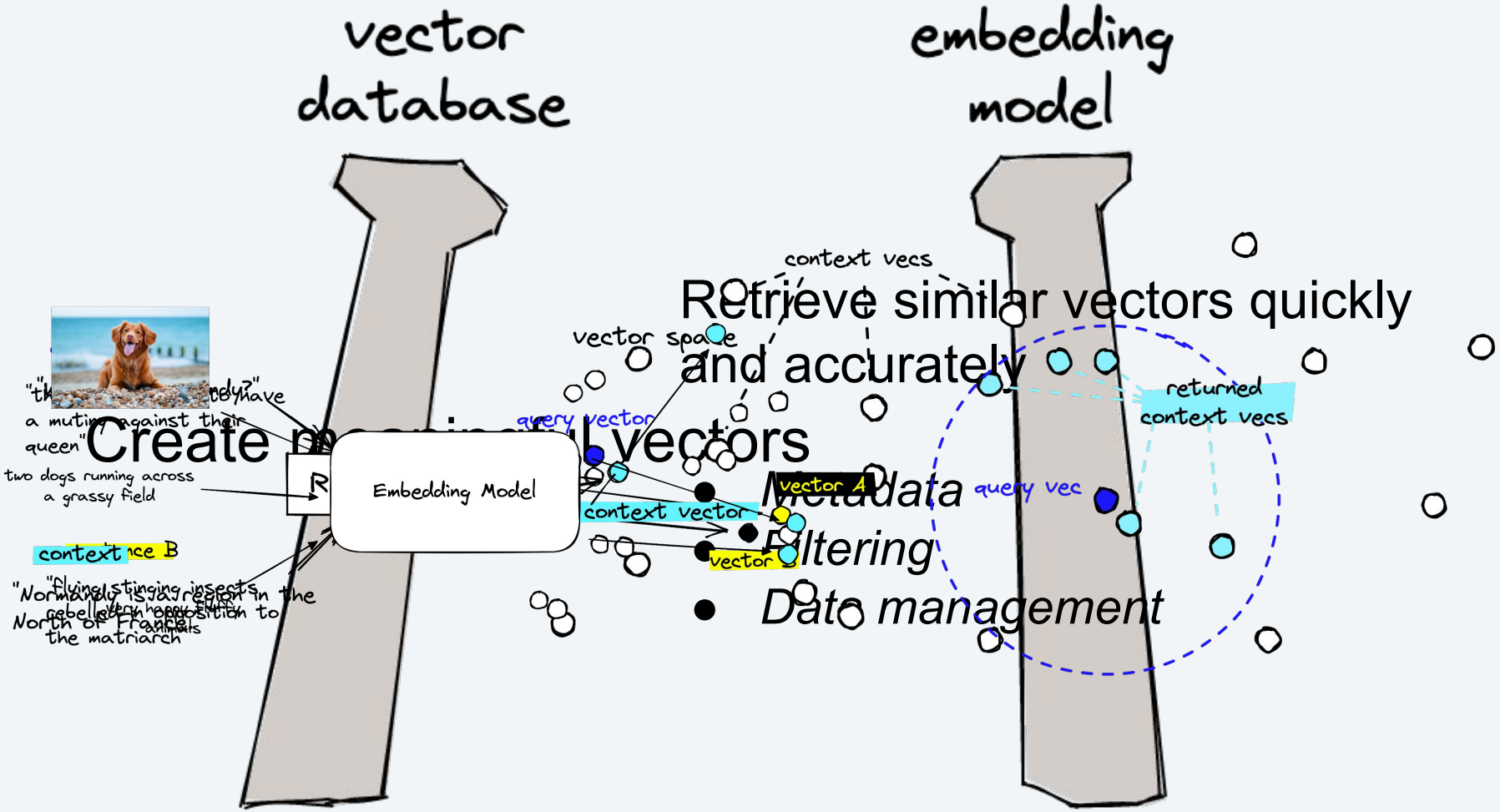
Freelance ML Engineering and Advisory

Other

Code: https://github.com/jamescalam/train-sentence-transformers

Ebook: https://www.pinecone.io/learn/nlp/


Semantic
Search
vector
search
NLP


vector database
embedding model
Retrieve similar vectors quickly and accurately
Create meaningful vectors
Metadata Filtering
Data management
context vecs
vector space
query vector
returned context vecs
query vec
vector A
context vector
vector B
Embedding Model
two dogs running across a grassy field
context
"Normandy is a region in the North of France"
"flying stinging insects"

There are 1000s of *pretrained* embedding models

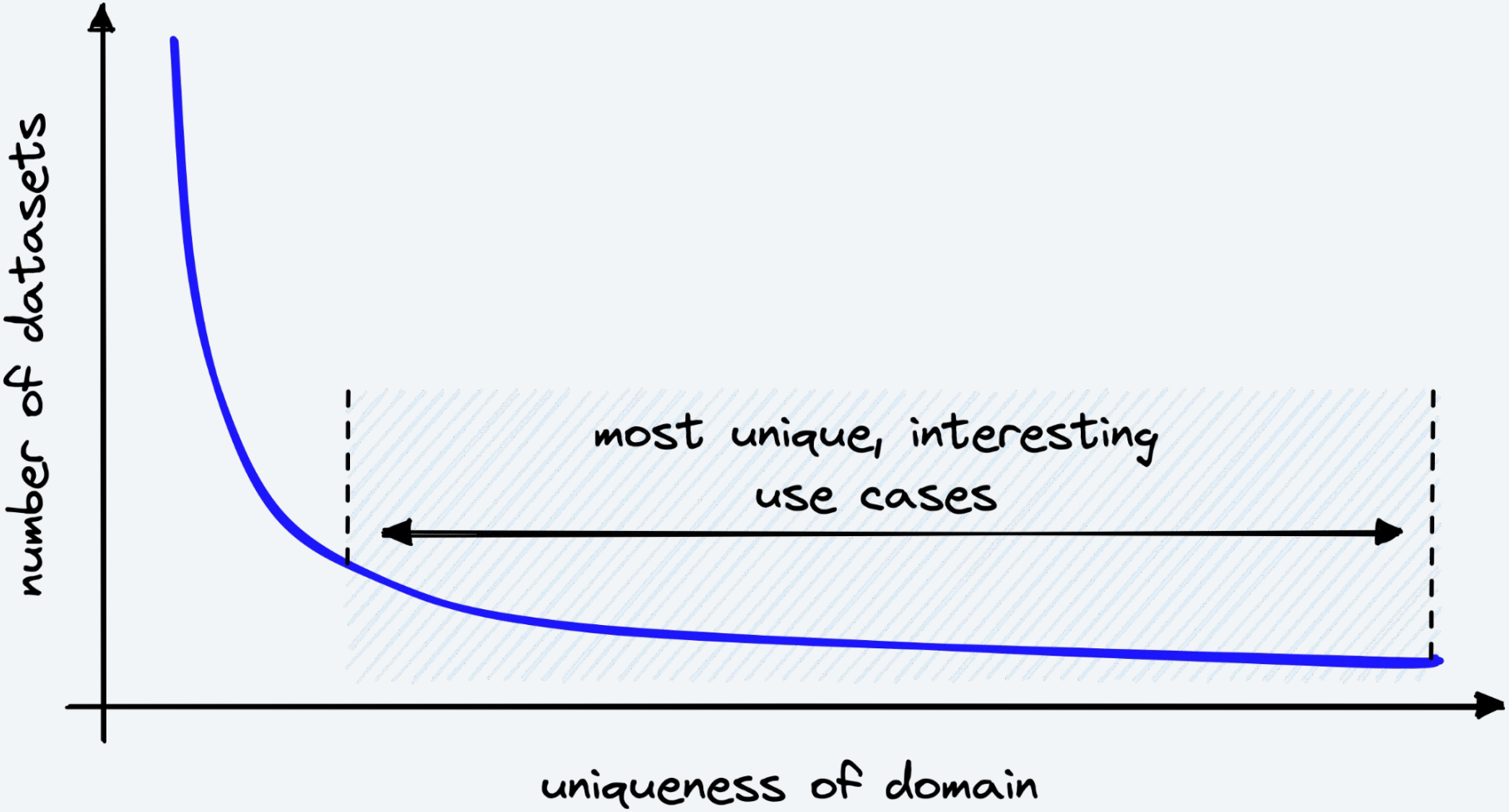
huggingface.co/models

pytorch.org/hub

api.openai.com/

os.cohere.ai/ co:here

Why fine-tune anything?


number of datasets
most unique, interesting
use cases
uniqueness of domain

• 3 days ago

if I wanted to do TSDAE and GPL together is there a specific order? Like pretrained using TSDAE first then fine tune with GPL

• 3 weeks ago

Hi James! You mentioned a project you did involving the Dhivehi and using TSDAE. Given it's hard to find a doc2query and cross-encoder model in Dhivehi, could you have used GPL as well?

1 month ago

Hi James, I'm looking to finetune a transformer model on my custom dataset for which I'm annotating documents by myself, My doubt is how many documents do I need to annotate for finetuning?

• 3 months ago

I like that approach with the 3-4 sentences. I made something similar but I grouped one minute worth of audio together. Overlapping sentences sounds like it can solve some of the issues I have seen. I actually have many questions: Hope that's okay
- How does this way differ from using Haystack QA and pipeline with Pinecone? Is this more accurate?
- If I wanted to train the QA model to better answer questions regarding a specific topic, say only include psychology videos TTS and then train the model to better understand psychological terms, how would I do that?

4 months ago

how to train sbert with a specific domain?

6 months ago

Awesome video James, how can we fine tune a model on our custom data, is there any way to train the model with out custom data. One more question which is a better approach this one or facebook/DPR model?

• 10 months ago

Hello, i have one question : is it possible to fine tune sbert with unsupervised data ? I have 2 job catalogs that i need to match by textual similarity but i dont have labels (for each job I only have it's title and description)

• 1 year ago

Great series! This is exactly what i have been looking for the project im working on. Thanks for making this. It would really great if you could share more on how to finetune the three models for new languages or share some codes we can explore. Thanks again.

09/21/2022

I hope to achieve good results, it is a challenge to train a model in Spanish, I don't know any dataset specialised in legal-QA, so I will make a fine tuning to a legal dataset in my language.

09/12/2022

@jamesbriggs do you have any examples of pretrain using mlm or tsdae on allmpnet models

08/31/2022

Has any one performed any form of evaluation of haystack? (How else can we improve it right?) I mean to check score and improve. I am trying to implement their evaluation system, but cant really figure it out.

09/11/2022

Hello 🙂
I have a question: When trying to fine-tune a pretrained BERT model to a new domain (Domain adaptation) should I perform the MLM task on few epochs or should I also apply the NSP task?

08/31/2022

On the github, there is literally the code to do semantic similarity for his top performaing models
https://github.com/Muennighoff/sgpt#asymmetric-semantic-search
There are 4 types tho and I don't understand what the difference is.
Cross-Encoder vs BiEncoder with Symmetric vs Asymmetric semantic search
I don't know which one to use for the best performance.

07/14/2022

Hello and thank you for all the help you are providing! I am working on semantic search for my internship/thesis and for the evaluation of different models, I have gathered around 100 similar sentence-pairs from the company data and I have been getting the cosine similarity for each pair with different models (sbert/use/elmo/simCSE etc. and non-contextual algorithms as well). Does it make sense to evaluate them this way, looking for example at the mean of the similarity scores or should I be looking into something else? I also want to try and finetune models on this list. The domain is healthtech (a mixture of medical/biomedical and digital health) and it's user-generated text (so the sentences are not purely technical) . I would really appreciate any ideas you may have!! Thank you!!

@jamesbriggs mpnet is generally a good choice though for ...

08/16/2022

Yeah, using mpnet only... Don't think model has any issue... Might be fine-tuning would help here

Do you have any references for tuning mpnet?

01/10/2022

hello. I have a QA model and want to do some re-ranking on answers. I produced multiple answers just like the qa pipeline in huggingface.

To re-rank them i'm thinking to use sentence bert with MSE loss on their F1 scores where the input is (question, answer_i), does this approach make sense?

09/15/2022

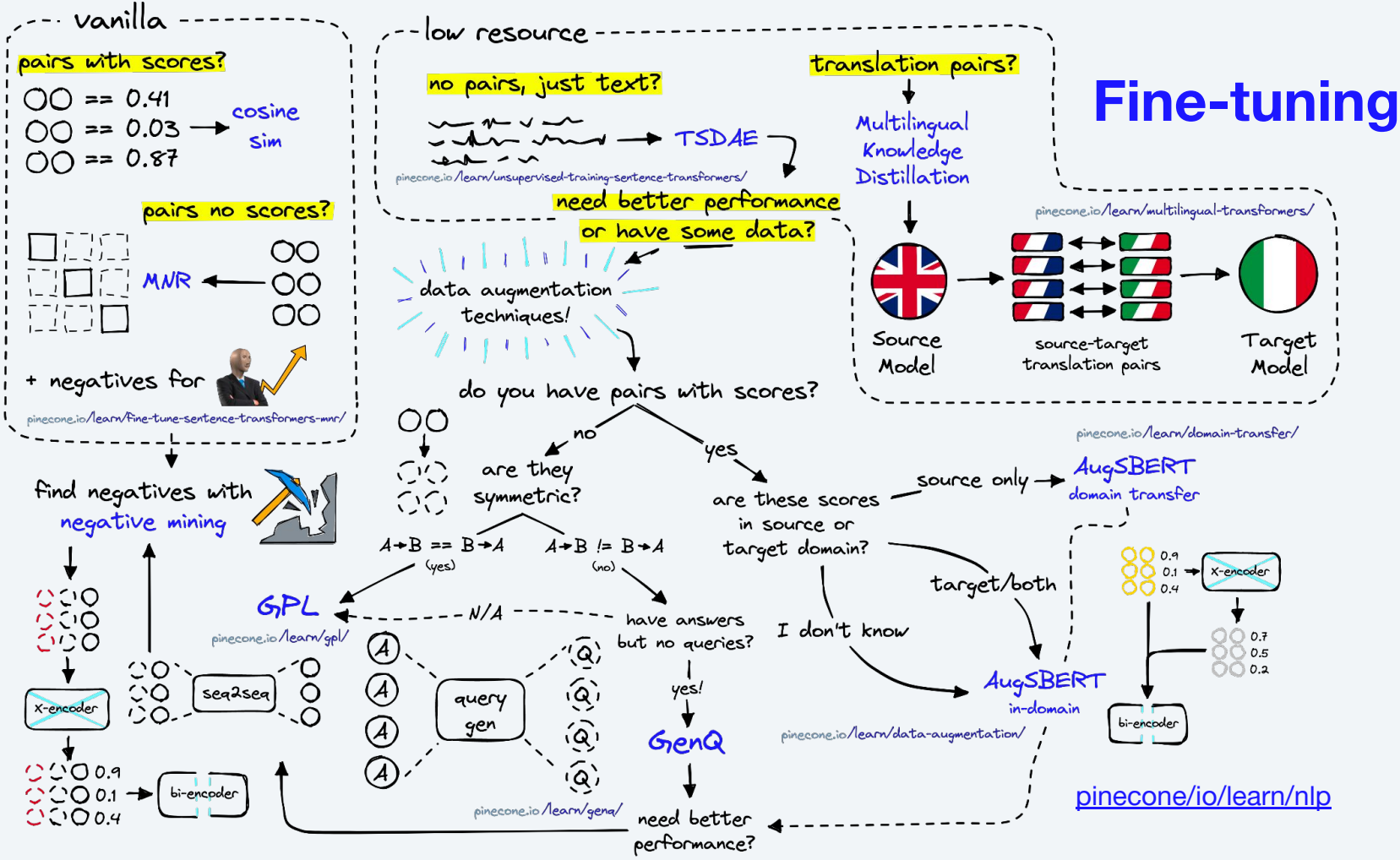
I've trained a model for sentence pair classification, but it seems unable to manage misspelt words, even if they are similar. Still, there is no such problem in sentence transformer its cosine similarity; it can recognise it with a higher score. Any ideas?

07/21/2022

Hello, have you tried running the training_nli_v2 training example script from the sentence-transformers library?

We can all benefit from some fine-tuning

Fine-tuning


vanilla
pairs with scores?
== 0.41
== 0.03
== 0.87
cosine sim
pairs no scores?
MNR
+ negatives for
pinecone.io/learn/fine-tune-sentence-transformers-mnr/
find negatives with negative mining
low resource
no pairs, just text?
TSDAE
pinecone.io/learn/unsupervised-training-sentence-transformers/
translation pairs?
Multilingual Knowledge Distillation
pinecone.io/learn/multilingual-transformers/
Source Model
source-target translation pairs
Target Model
need better performance or have some data?
data augmentation techniques!
do you have pairs with scores?
no
yes
are they symmetric?
A→B == B→A (yes)
A→B != B→A (no)
are these scores in source or target domain?
source only
target/both
I don't know
pinecone.io/learn/domain-transfer/
AugSBERT domain transfer
GPL
pinecone.io/learn/gpl/
N/A
have answers but no queries?
yes!
query gen
GenQ
pinecone.io/learn/gena/
AugSBERT in-domain
pinecone.io/learn/data-augmentation/
X-encoder
seq2seq
bi-encoder
0.9
0.1
0.4
0.7
0.5
0.2
need better performance?
pinecone/io/learn/nlp

Comparing items

Fine-tuning embedding models revolves around **contrast**

A is similar to B

B is *not* similar to C

etc

Labeled text-pairs

Fine-tuning with labels

Labeled text-pairs

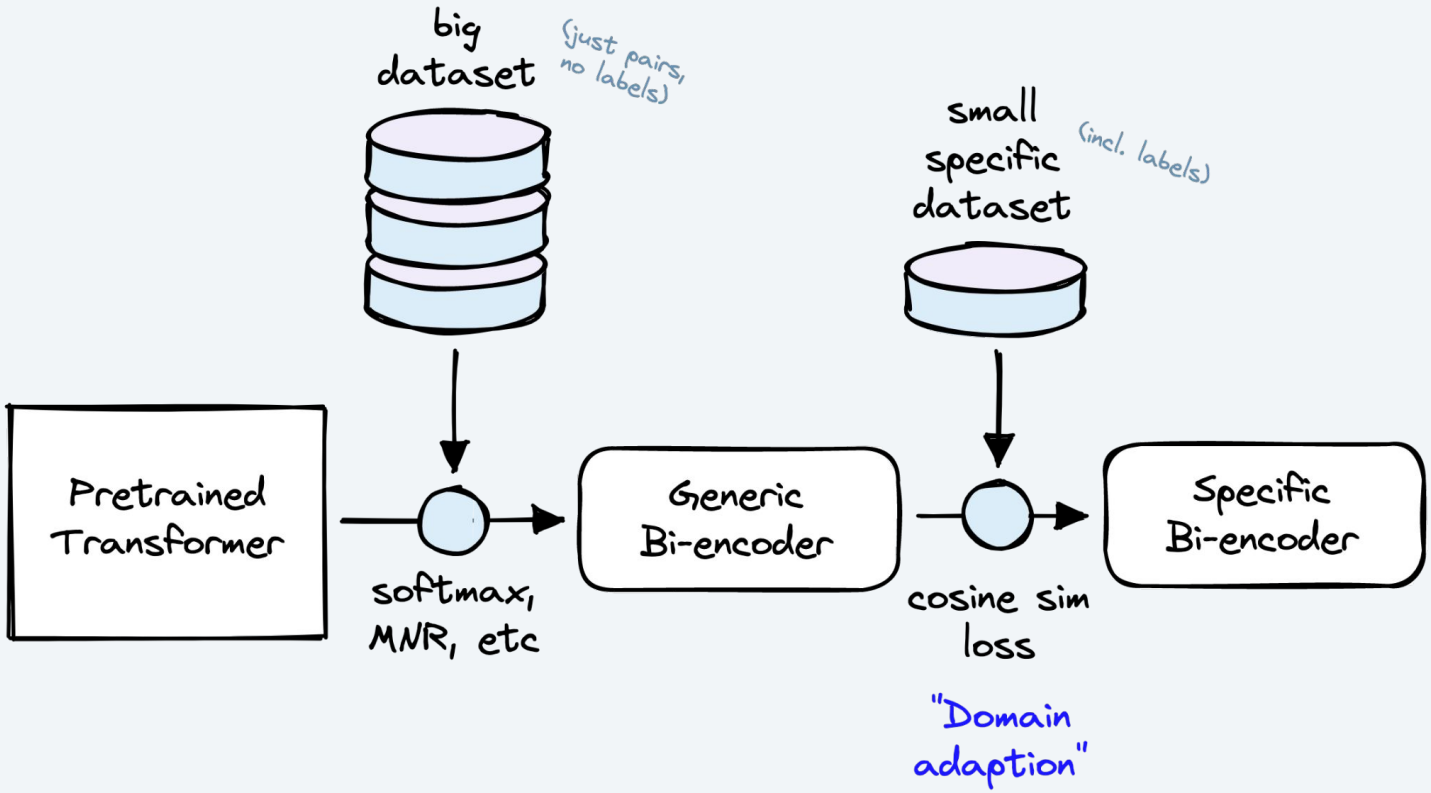
\+ Good performance

\- Hard to find data

sentence1	sentence2	label
"a plane is taking off"	"an air plane is taking off"	5
"a man is playing a large flute"	"a man is playing a flute"	3.8
"a man is spreading shreded..."	"a man is spreading shredding..."	3.8
"three men are playing chess"	"two men are playing chess"	2.6

GLUE **S**entence **T**extual **S**imilarity **b**enchmark (STSb)

It's hard to find labeled data



Unlabeled text-pairs

Unlabeled text pairs

Most datasets don’t contain specific (or any) labels

premise	hypothesis	label
"choir sings to church audience"	"church with cracks in ceiling"	1 (neutral)
"choir sings to church audience"	"church is filled with song"	0 (entailment)
"choir sings to church audience"	"choir sings at baseball game"	2 (contradiction)
"woman with a green scarf"	"the woman is young"	1 (neutral)

Stanford **N**atural **L**anguage **I**nference (SNLI)

Multiple Negatives Ranking

All we need are *(anchor, positive)* pairs

premise	hypothesis	label
"choir sings to church audience"	"church is filled with song"	0 (entailment)
"woman with a big grin"	"the woman is very happy"	0 (entailment)
"old man poses in front of advert"	"man poses in front of ad"	0 (entailment)

Stanford **N**atural **L**anguage **I**nference (SNLI)

M. Henderson, et. al., Efficient Natural Language Response Suggestion for Smart Reply (2017), Google
N. Reimers, Losses - MultipleNegativesRanking, Sentence Transformers Docs

We mix-and-match to get negatives

A_0P_0 = ["choir sings to church audience", "church is filled with song"]

A_1P_1 = ["woman with a big grin", "the woman is very happy"]

A_2P_2 = ["old man poses in front of advert", "man poses in front of ad"]

Mix different pairs to create negatives

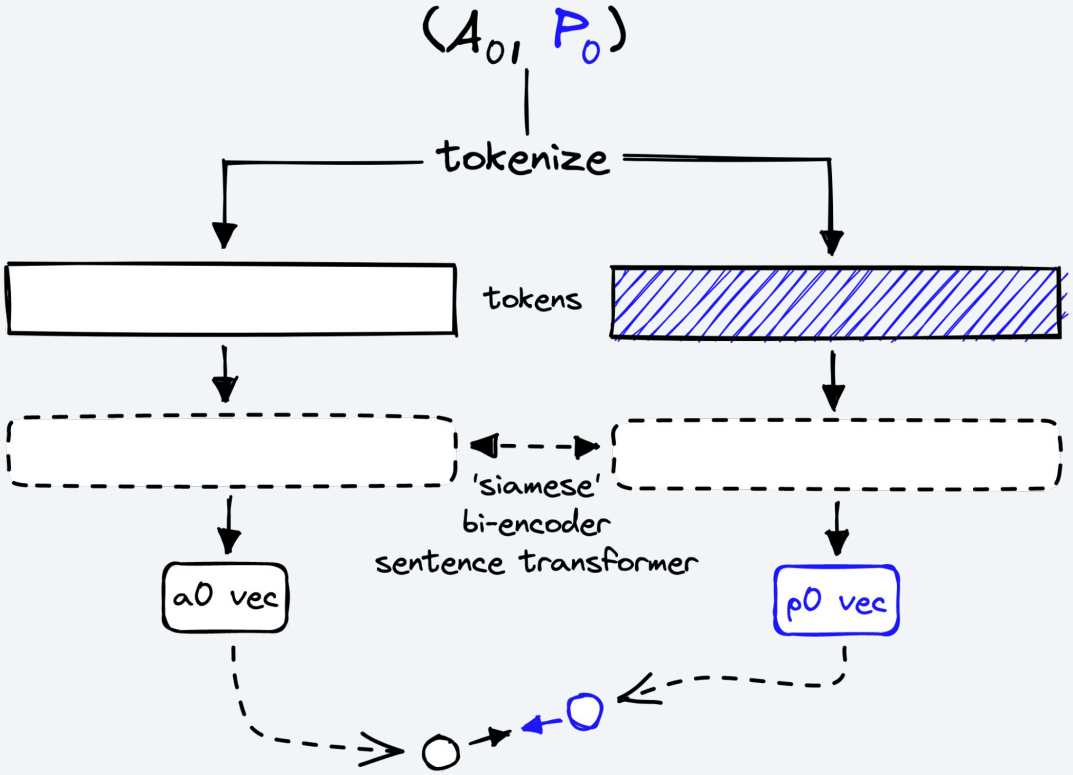
A_0P_0	A_1P_0	A_2P_0
A_0P_1	A_1P_1	A_2P_1
A_0P_2	A_1P_2	A_2P_2

Multiple Negatives Ranking

We mix-and-match to get negatives

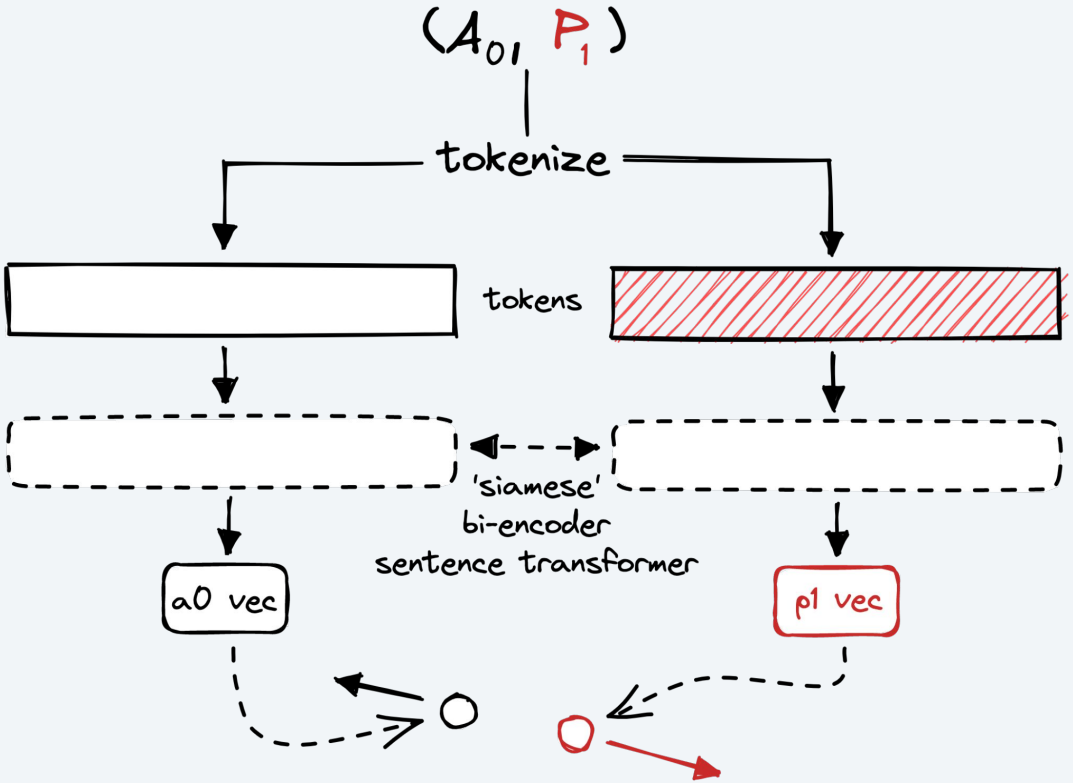
Source: pinecone.io/learn/genq/
pinecone.io/learn/fine-tune-sentence-transformers-mnr/

Then optimize



Multiple Negatives Ranking

Then optimize

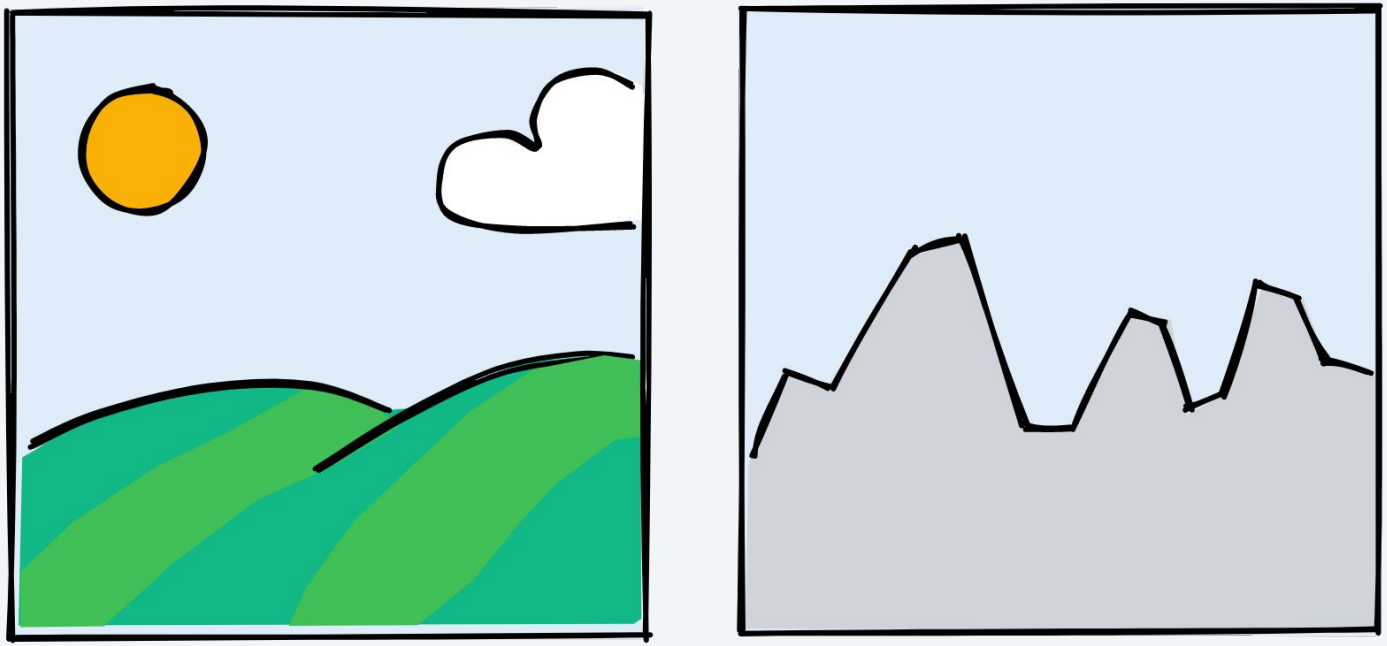


This works, but could be better

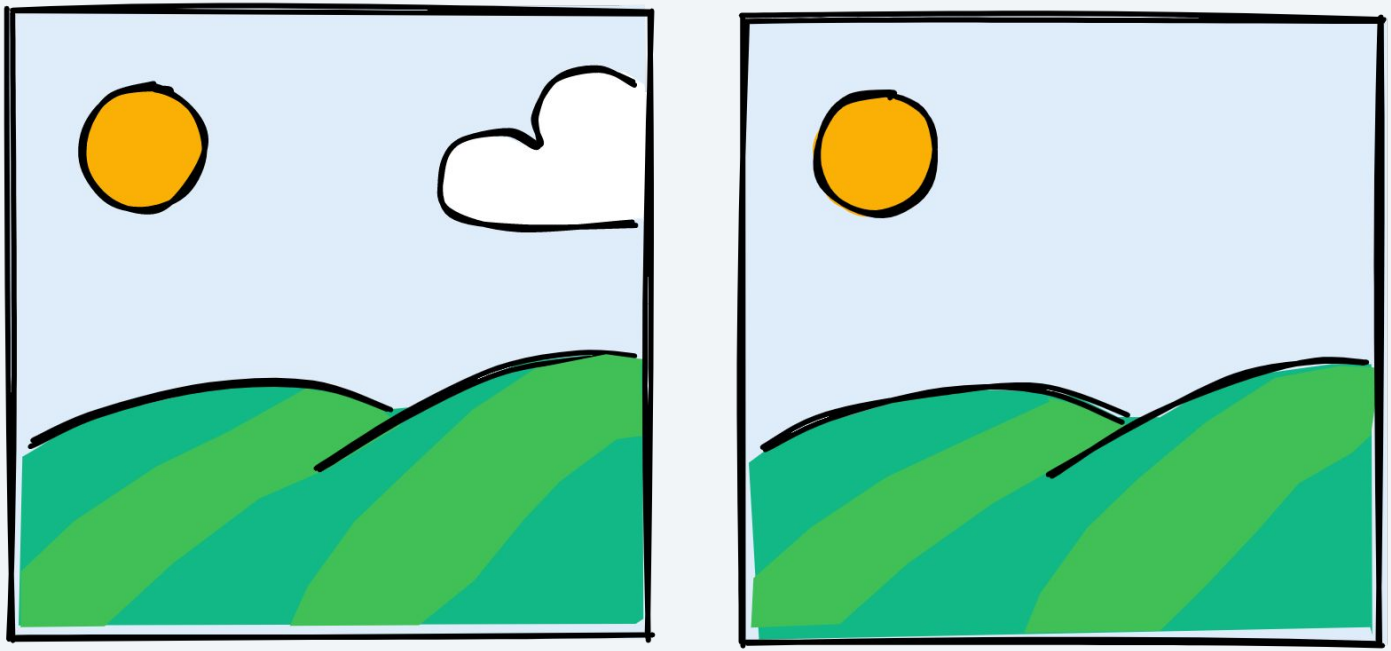
Randomly chosen pairs are often easy to distinguish

Multiple Negatives Ranking

It’s like asking the model to spot the difference

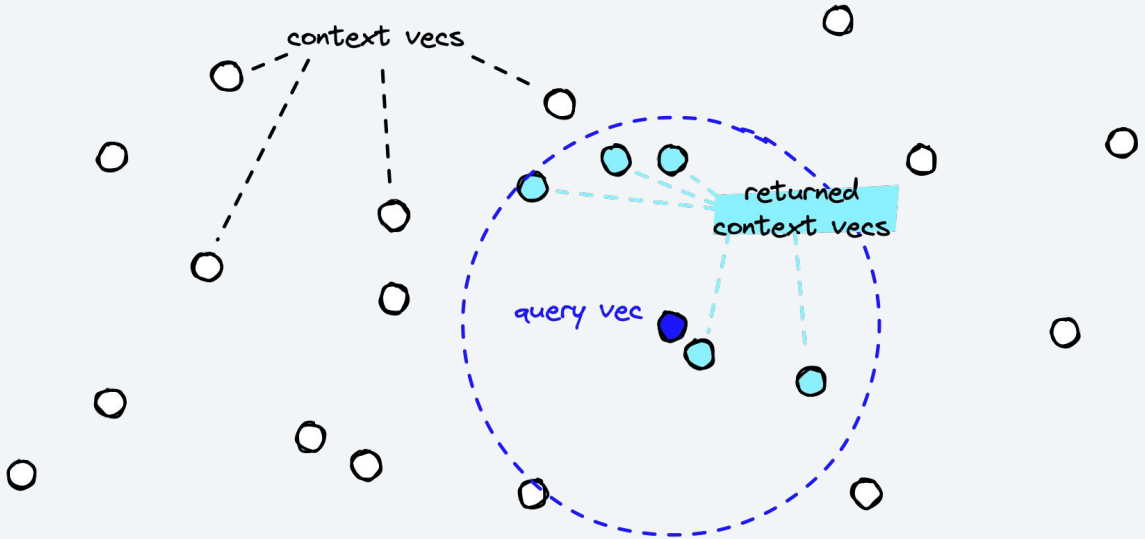


Ideally, we want to challenge the model



Hard-negatives

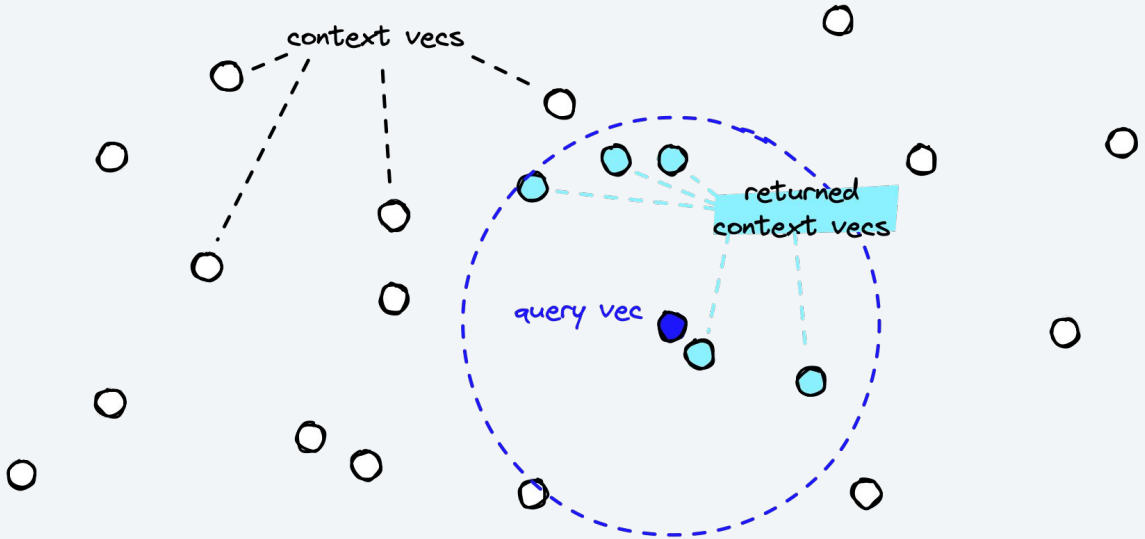
Hard-negatives are items that the model struggles to distinguish as negative



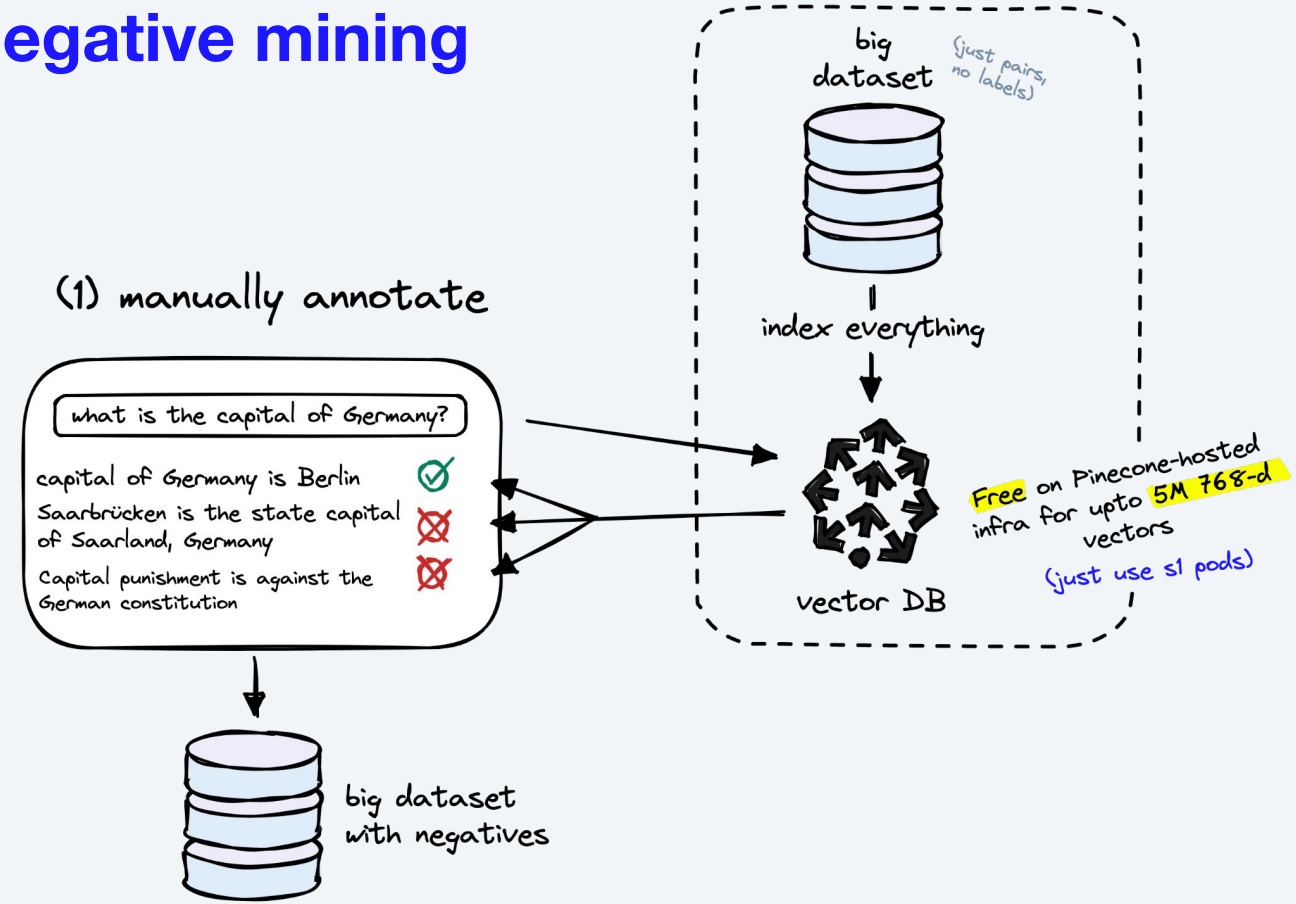
(Hard-negatives are not MNR-specific)

Finding hard-negatives

1. In the dataset, if we're lucky
2. Hard-negative mining



Hard-negative mining

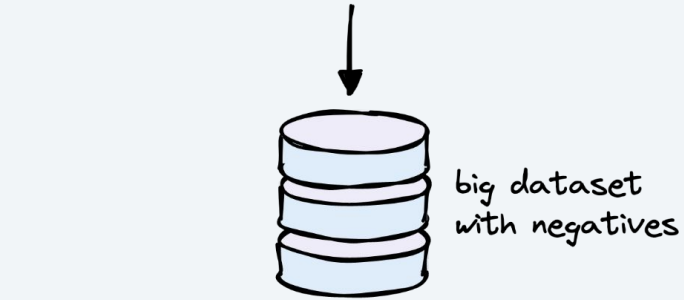


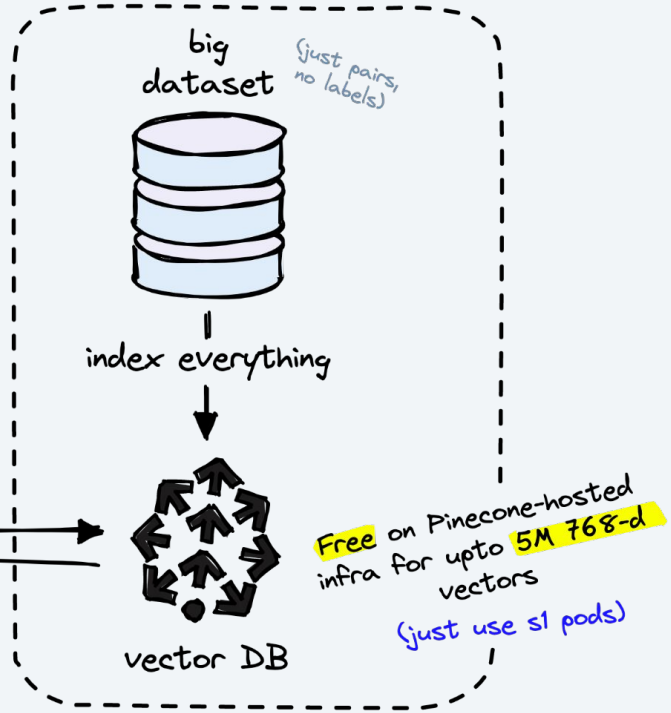
Hard-negative mining

(2) automatically annotate

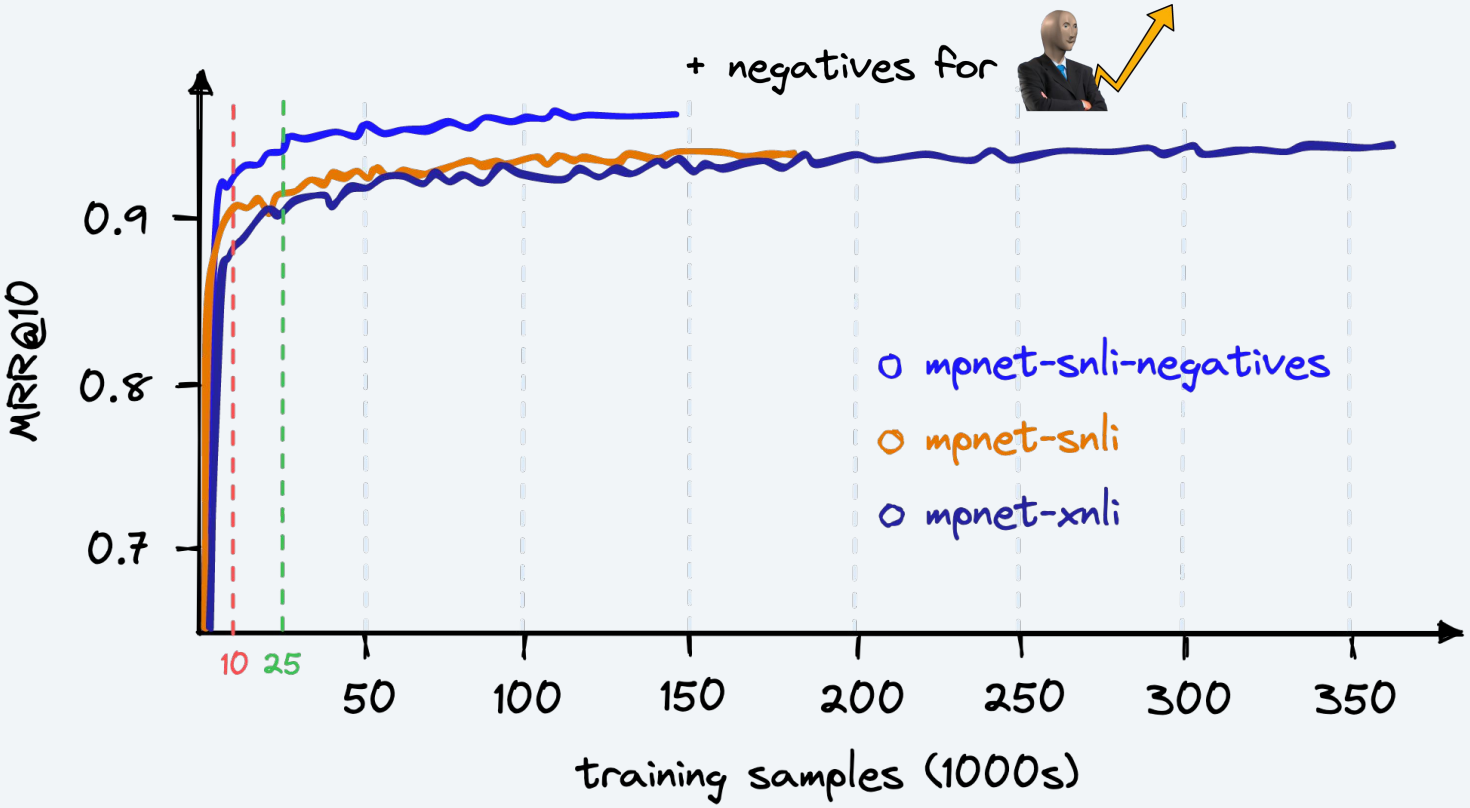
```python
negatives = []

for i, row in enumerate(dataset):
    anchor = row['anchor']
    xq = model.encode([anchor])
    # filter out the results that match to anchor
    meta_filter = {'idx': {'$neq': i}}
    # find negatives
    xc = index.query(xq, top_k=5, filter=meta_filter)
    negatives.append({anchor: xc['matches']})
```





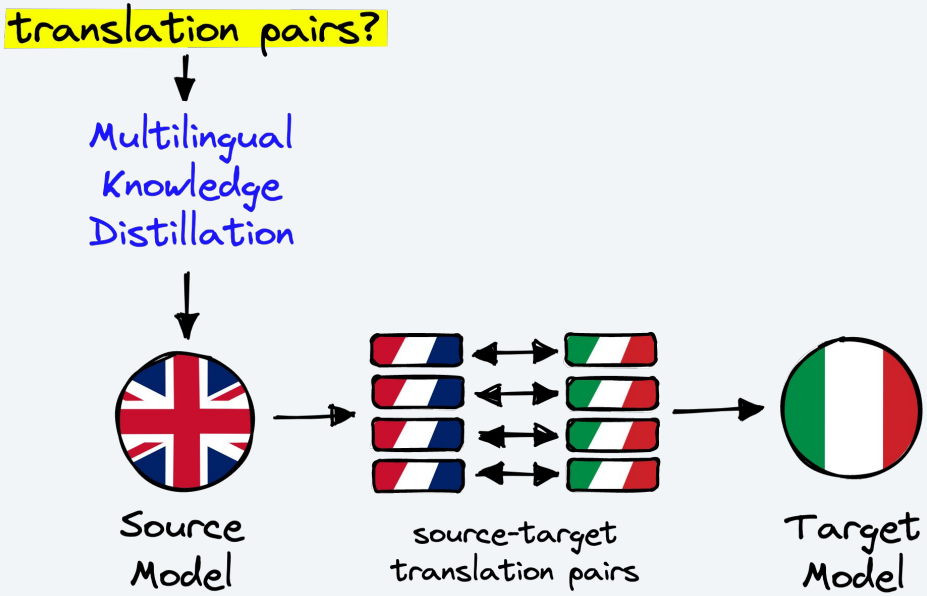
How many pairs are needed?



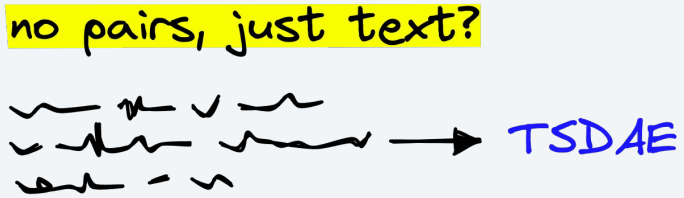
Low resource scenarios

Low-resource

Low-resource == very little data



N. Reimers, I. Gurevych, Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation (2020), EMNLP



K. Wang, et al., TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning (2021), EMNLP

Translation pairs

Translation pairs

“Parallel data” - text-pairs in a **source** language and a **target** language

\+ It’s easy to find parallel data

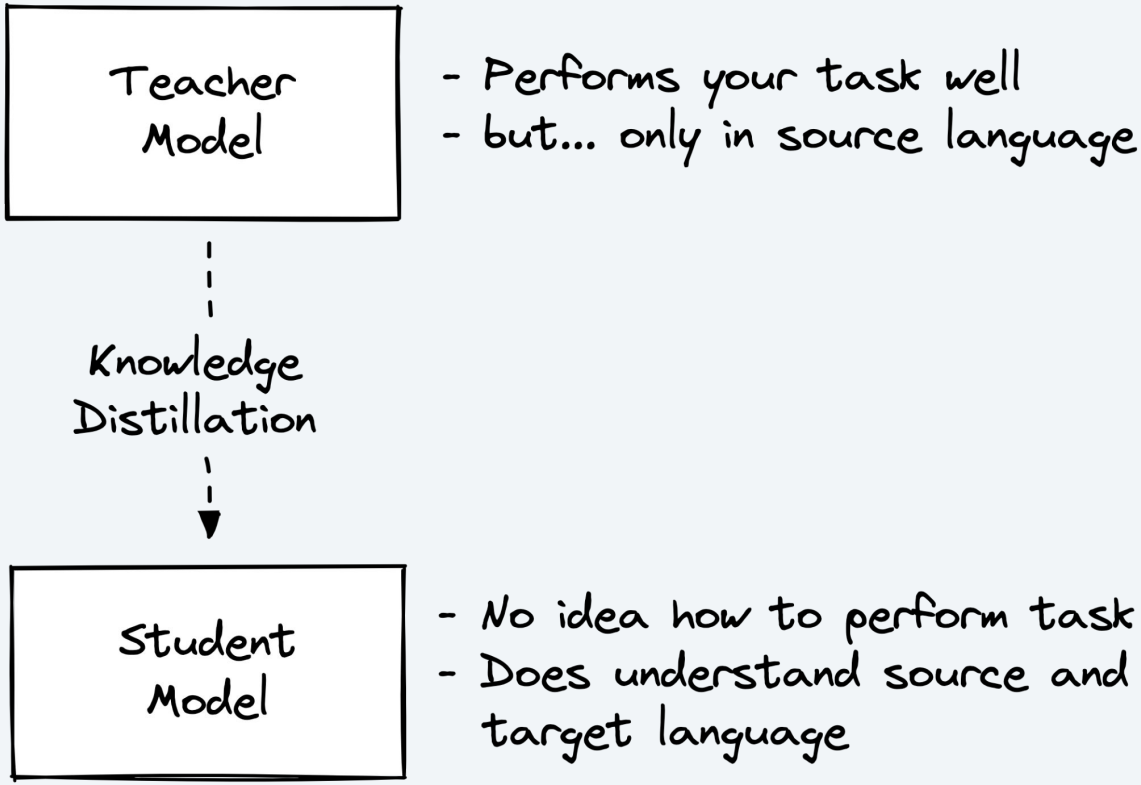
\- Requires existing use-case model in source language (**teacher**)

\- Requires pretrained model for source and target language (**student**)

With translation pairs we can use:

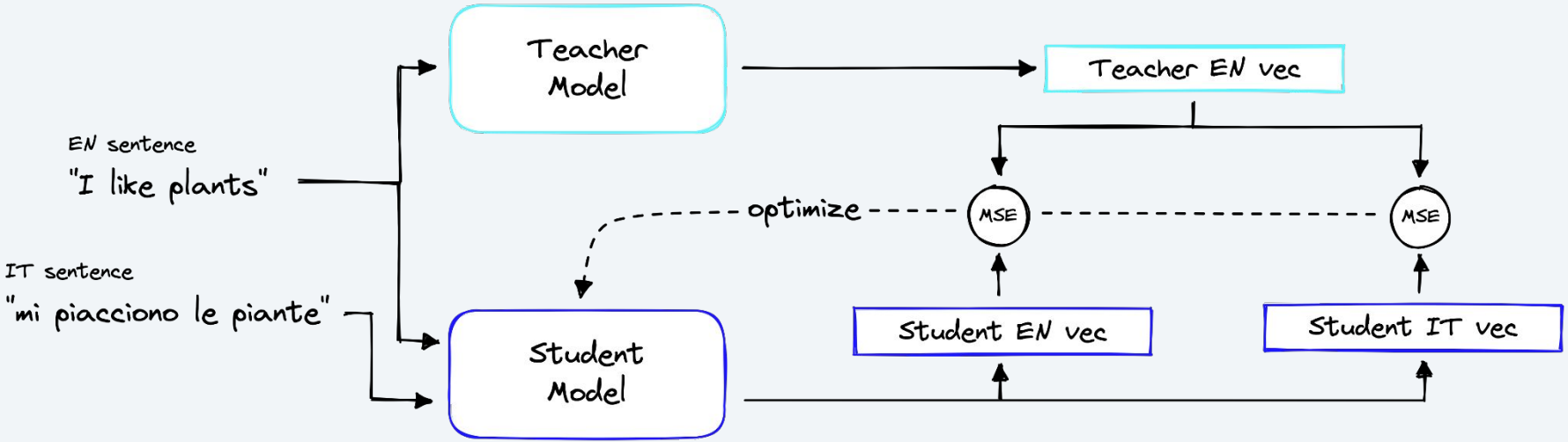
Multilingual knowledge distillation

Teacher and Student



Training Overview

Teacher model performance on **source** set == perf ceiling



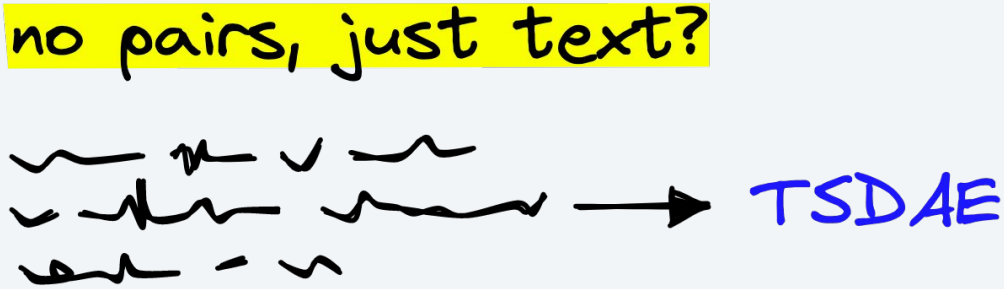
Student should approach perf ceiling
set by teacher on **target** set

… for several epochs

Only have text?

No pairs, just text?

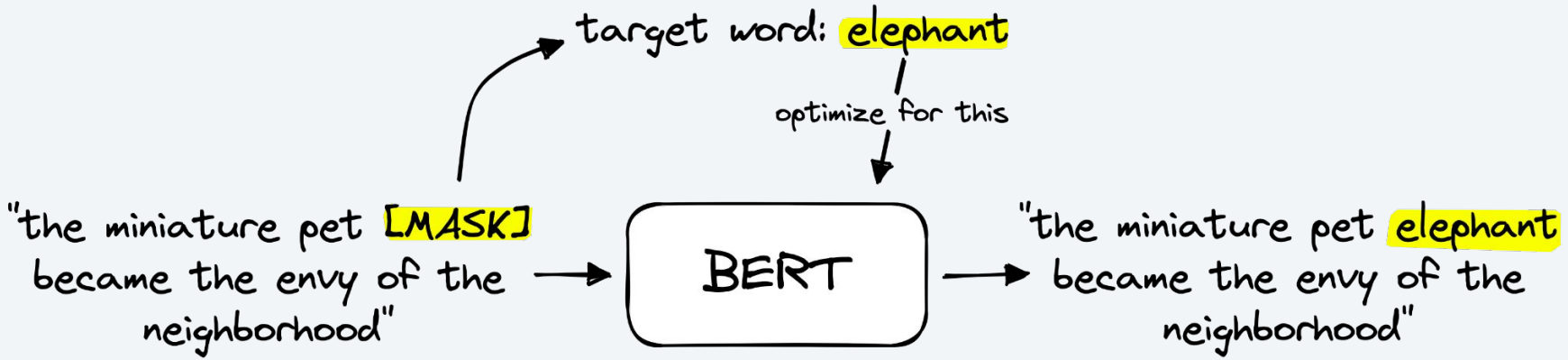
What if we don't have *any* text pairs?



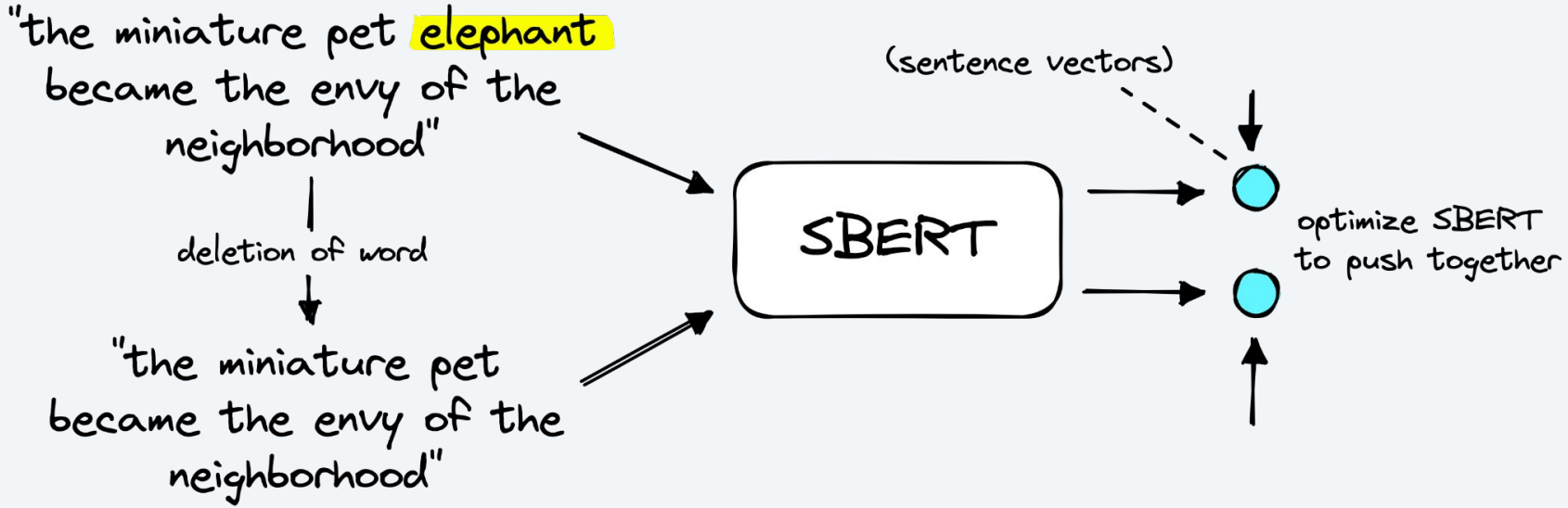
K. Wang, et al., TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning (2021), EMNLP

Similar to MLM

Masked-Language Modeling (MLM) is the pretraining approach of models like BERT



But at the sentence-level



TSDAE

+ Easy to fine-tune
+ All we need is unstructured text data
- Only works for vanilla similarity (i.e. no QA)
- Performance cannot compare to previous methods

Data Augmentation

Synthetic dataset augmentation?

If we have *some* data, or are willing to annotate a small 1-5K dataset

Data augmentation with transformers

Can we generate more pairs, and label them?

More pairs == random mix-and-match

Synthetic similarity scores?

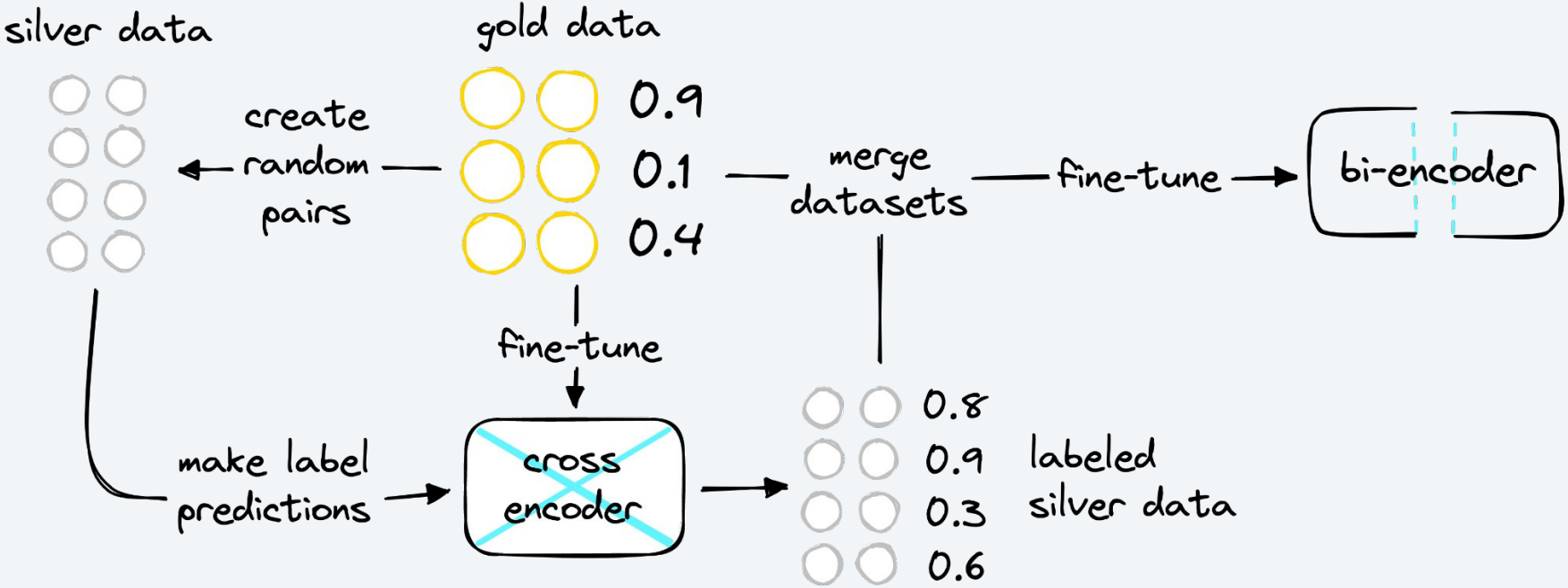
Cross-encoders

If we have a small 1-5K dataset of labeled pairs, we can fine-tune a cross-encoder

N. Thakur, et al., Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks (2021), NAACL

pinecone.io/learn/data-augmentation/

Process outline



N. Thakur, et al., Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks (2021), NAACL

pinecone.io/learn/data-augmentation/

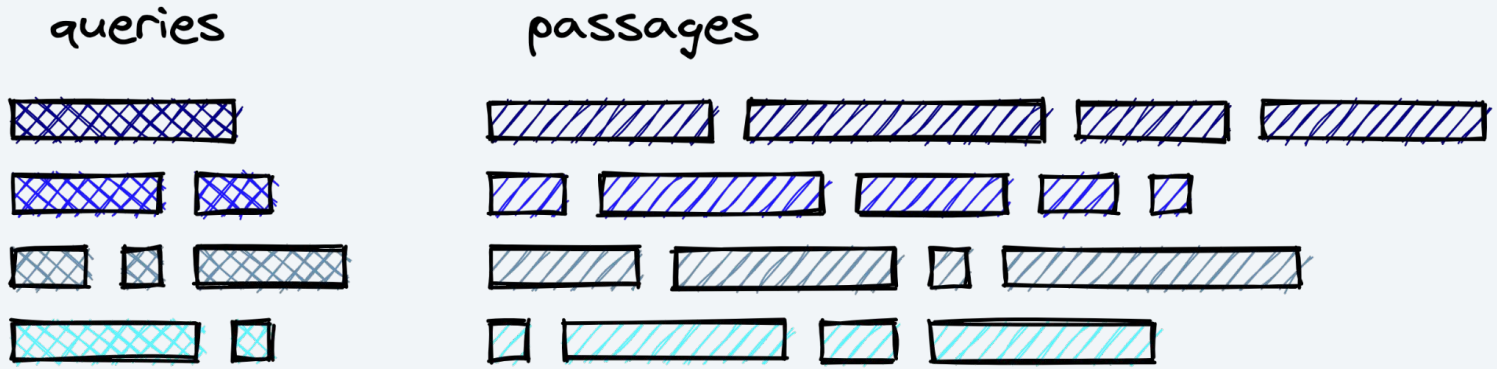
Considerations

+ Can start with small 1-5K dataset
+ Performance is reasonable
- Initial dataset must be labeled with similarity scores
- Must also fine-tune cross-encoder
- More complicated process

Asymmetric Search

For asymmetric search

Search can be symmetric or *asymmetric*

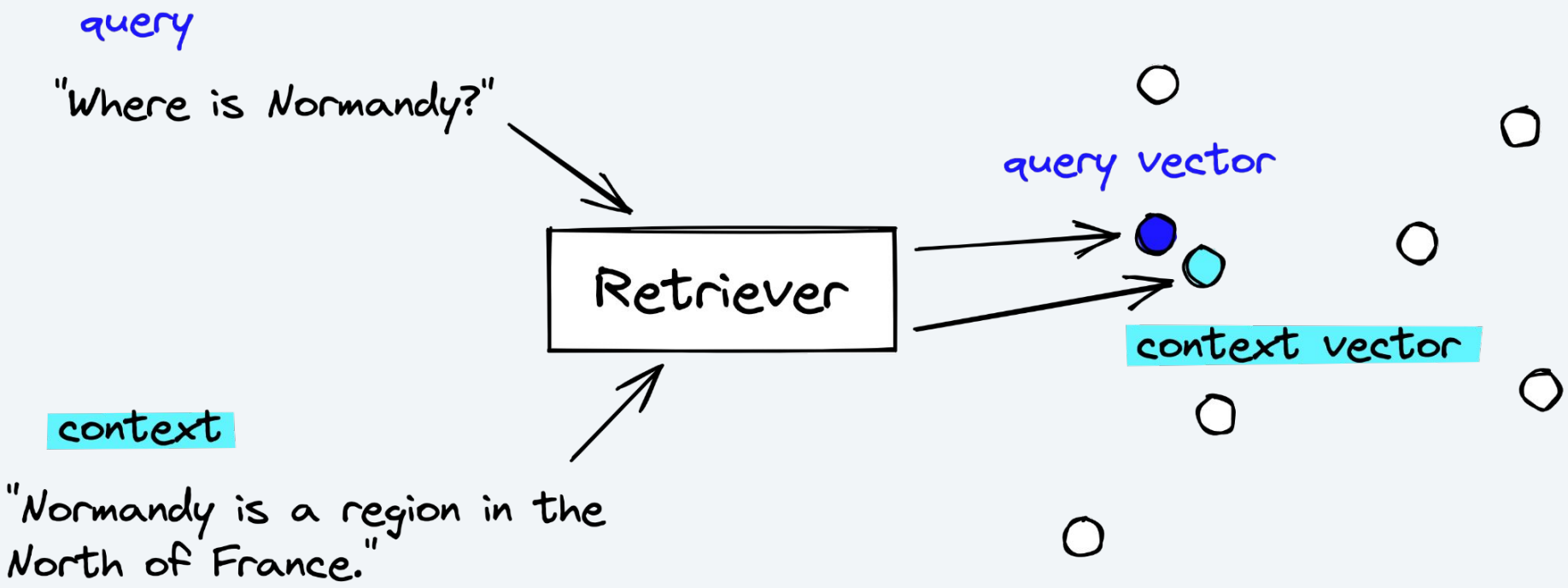


Asymmetric search, source: pinecone.io/learn/genq

Symmetric: query texts and passages texts are expected to be the same (size, form, etc)

Asymmetric: query texts are not equal size/form to passages texts, for example QA

Question-answering



Training methods

MNR is good when we have many query-context pairs

But with low-resource domains?

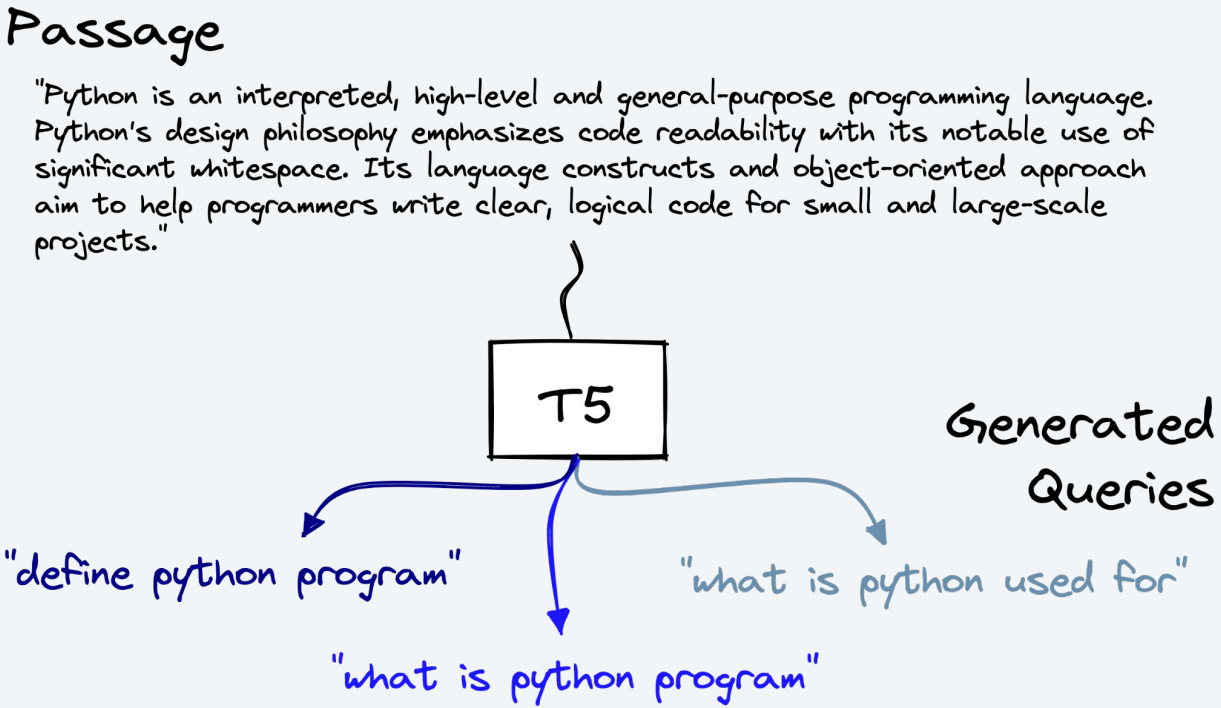
- TSDAE cannot (only for vanilla similarity)
- AugSBERT format with scores not ideal

Query gen models

There are seq2seq transformer models

+ Given a “context” or passage, they can generate queries
- Needs pretrained seq2seq model or fine-tune on *(query, context)* pairs

Query generation



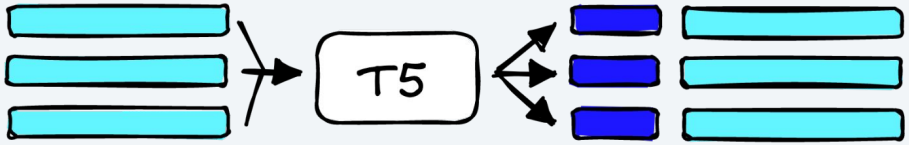
source: /learn/genq

BeIR/query-gen-msmarco-t5-large-v1

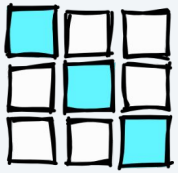
GenQ

1. Find contexts/passages from target domain

2. Generate queries for each passage



3. Train on (query, passage) pairs with MNR loss



J. Ma, et al., Zero-shot Neural Passage Retrieval via Domain-targeted Synthetic Question Generation (2021), ACL

GenQ

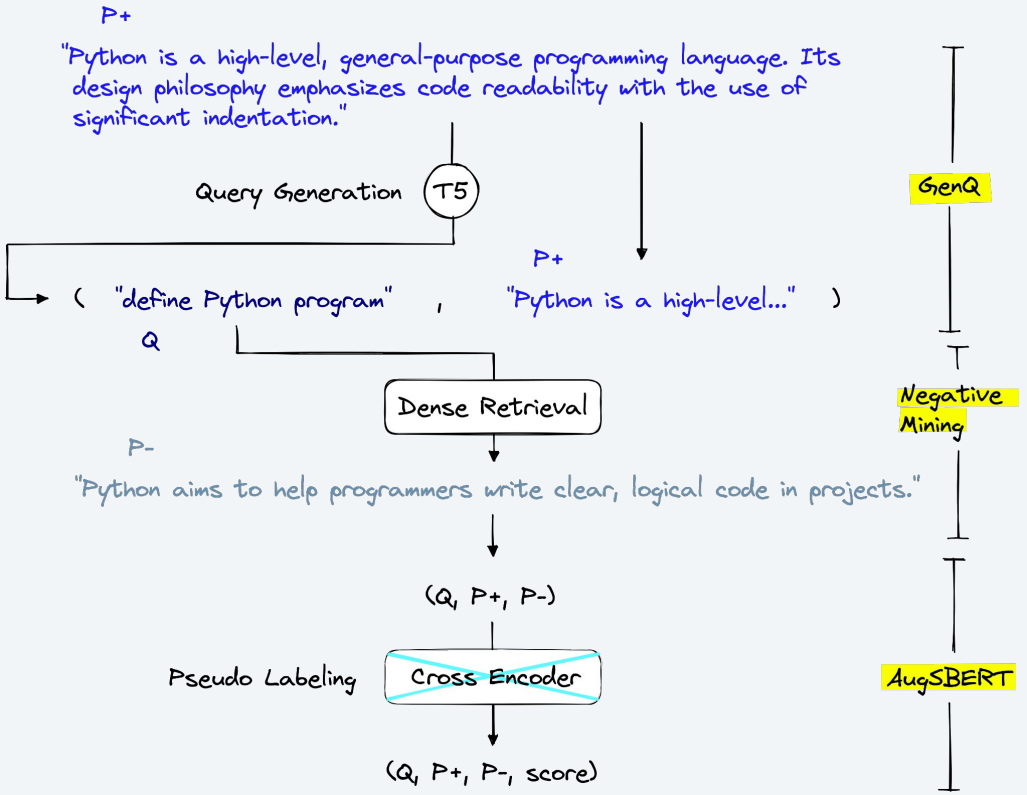
+ Only needs plain text data
+ Good performance *if* quality data
- Synthetic query generation are prone to poor quality
- Needs pretrained query gen model
- Whole process relies on high quality queries

Can we avoid bad performance from poor queries?

Generative Pseudo Labeling (GPL)

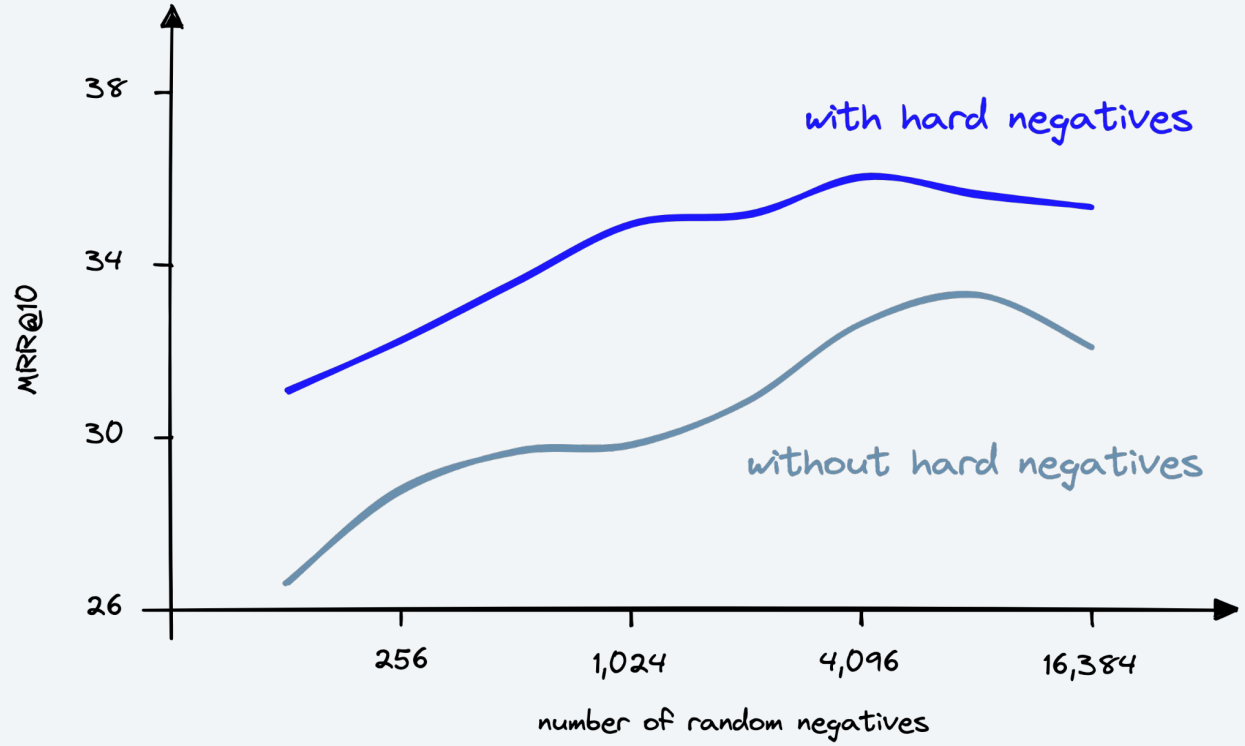
Negative mining means better model performance

Adding pseudo-labeling reduces reliance on good query generation because the margin score the indicates relevance



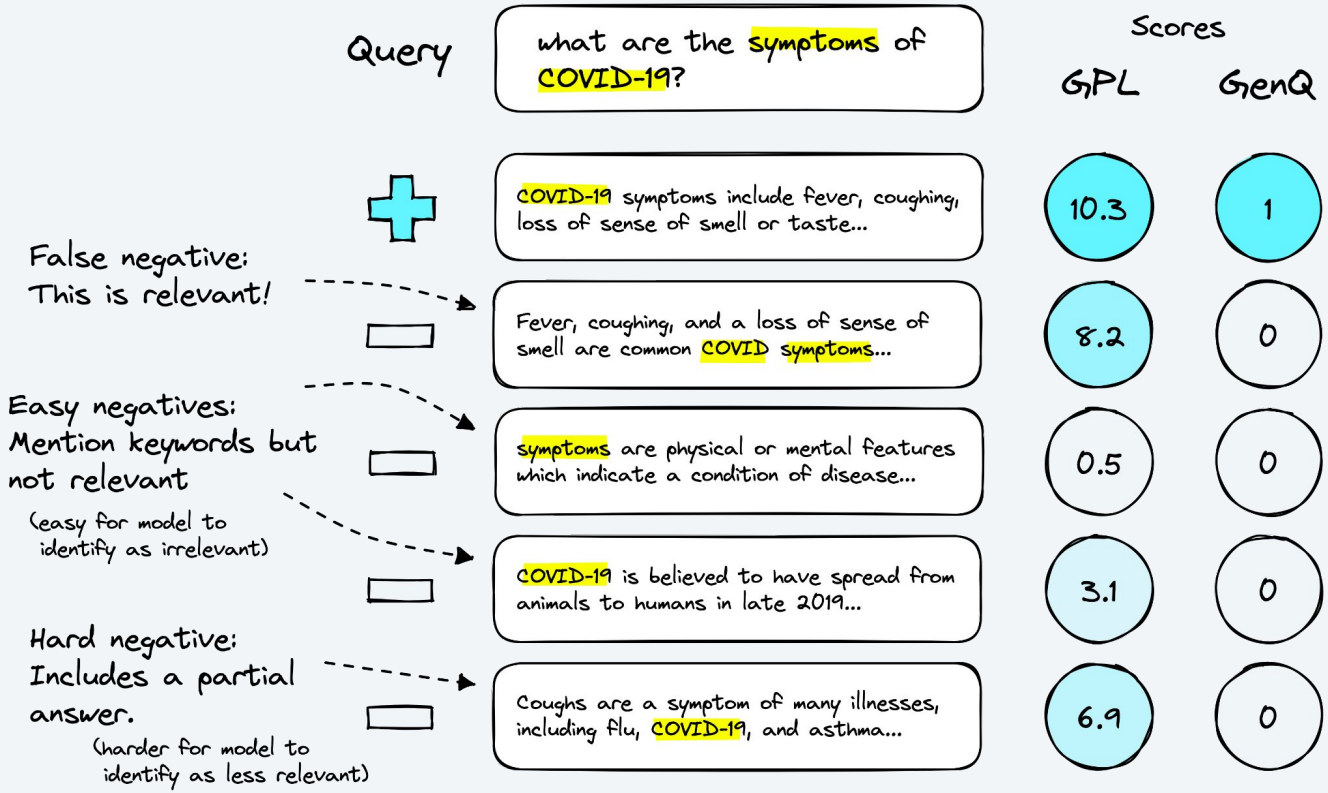
K. Wang, et al., GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval, NAACL 2022

GPL and negative mining



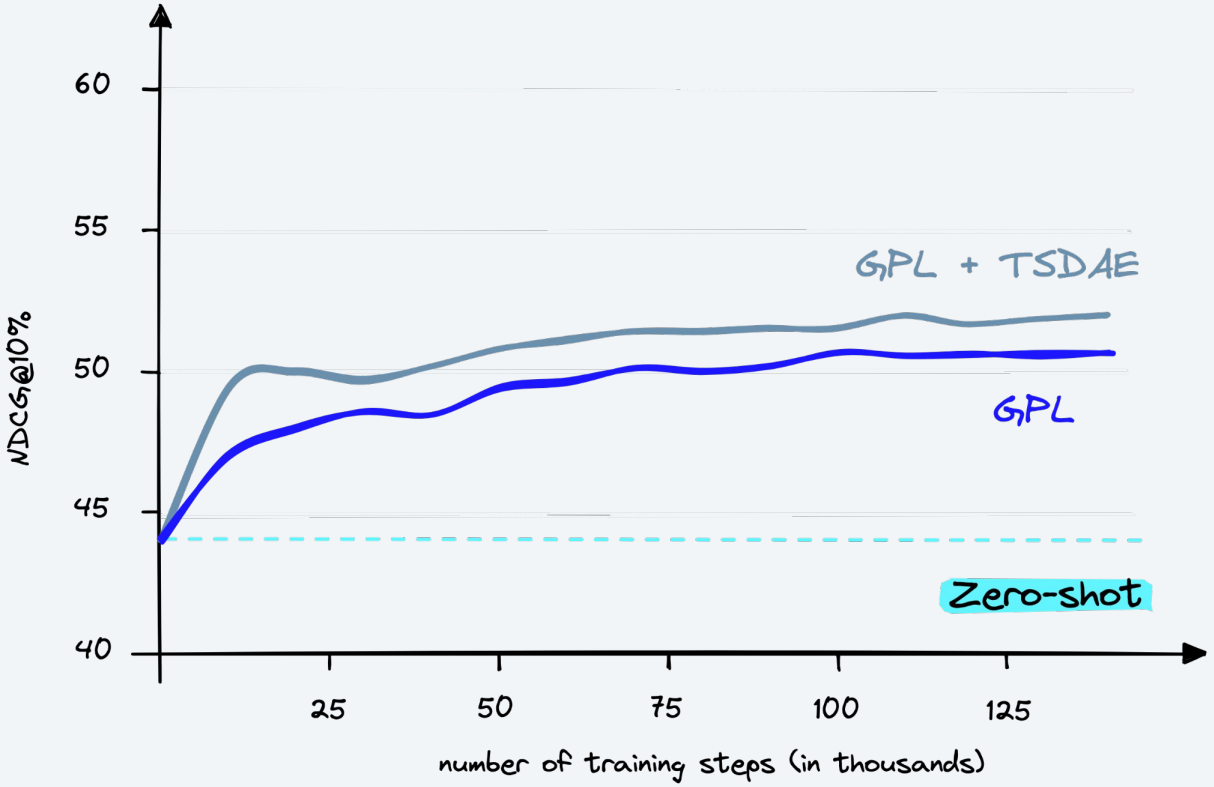
source: pinecone.io/learn/gpl

GPL and pseudo labeling



source: pinecone.io/learn/gpl

GPL and other techniques



source: pinecone.io/learn/gpl

How to fine-tune?

Pairs and labels? Cosine similarity loss

Just pairs? MNR

Not much of anything but translation pairs? Multilingual knowledge distillation

Small set of pairs and labels? AugSBERT

Asymmetric use-case, small dataset? GenQ

GenQ performing badly? GPL

Any Questions?

James

 YouTube.com/c/JamesBriggs

 discord.gg/c5QtDB9RAP

 twitter.com/jamescalam

Pinecone

 Vector database for millions or billions of records, free and paid services available

 Semantic Search ebook:

pinecone.io/learn/nlp

 Careers: pinecone.io/careers

 twitter.com/pinecone