LLM for Search Engines: Part 2

Chenyan Xiong

11-667

Disclaimer:

Pretraining for retrieval is a very premature field.
Anything we know now may be wrong.

Outline

Overview of Modern Information Retrieval Systems

- An example search component updated by LLMs
- Glances of other components using LLMs

Dense Retrieval, a different way of search with LLMs

- End-to-end learned retrieval
- Notable extensions

Pretrain retrieval representations

Mismatches Between LM Pretraining and Retrieval Needs

Various observations that pretrained LLMs are not as strong in retrieval than other language tasks

- Zero zero-shot performance from vanilla LMs, e.g., BERT, ELECTRA
- Required more complicated fine-tuning, e.g., smoothed self-negatives
- Prompting LLMs not really working

Mismatches Between LM Pretraining and Retrieval Needs

Various observations that pretrained LLMs are not as strong in retrieval than other language tasks

- Zero zero-shot performance from vanilla LMs, e.g., BERT, ELECTRA
- Required more complicated fine-tuning, e.g., smoothed self-negatives
- Prompting LLMs not really working

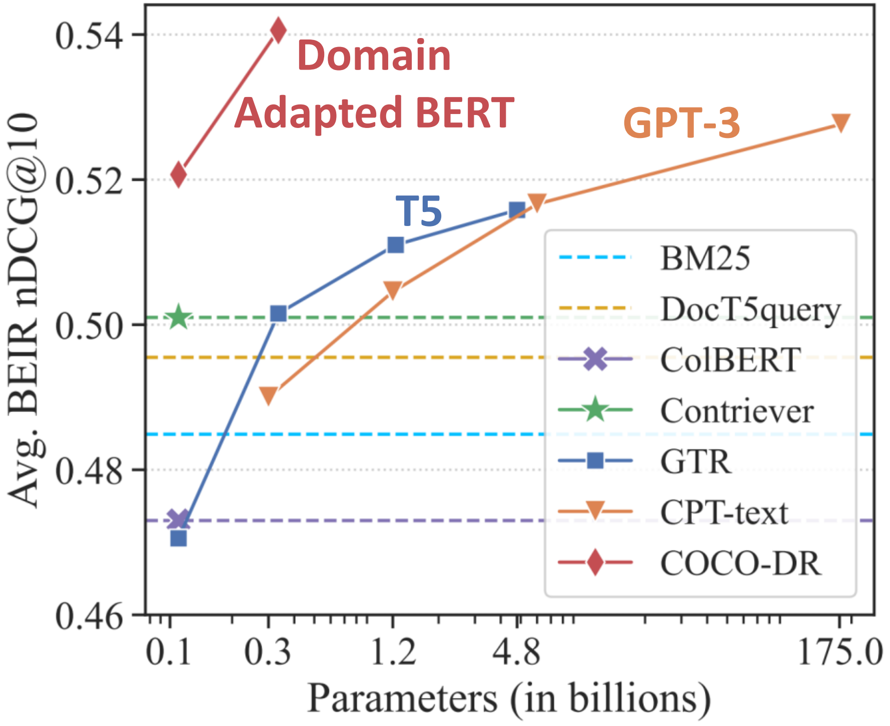


Figure 10: Scaling of LLMs on Zero-Shot Dense Retrieval [8]

Much worse scaling law from LLMs in retrieval

- GPT-3 much worse than T5 at similar scale
- More diminished return when scaling up
- Generalizing better with domain adapted pretraining

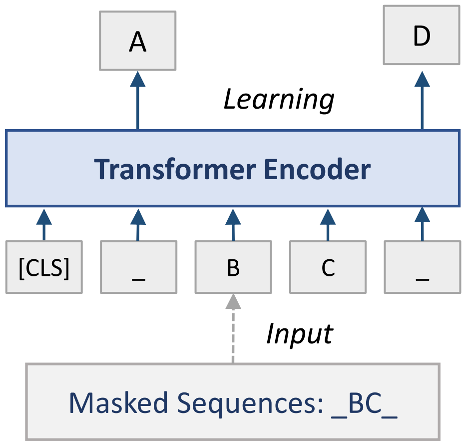
[8] Yu et al. "COCO-DR: Combating Distribution Shifts in Zero-Shot Dense Retrieval with Contrastive and Distributionally Robust Learning". EMNLP 2022.

Mismatch #1: Local versus Global

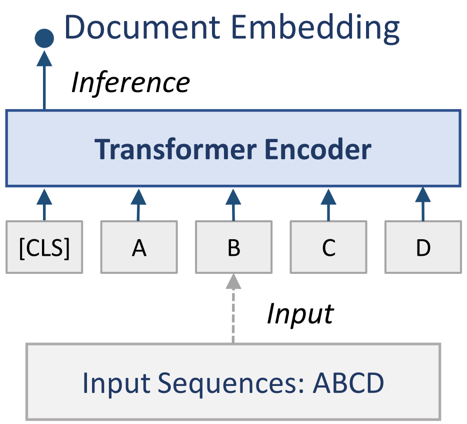
Language modeling is more about local contexts

Retrieval requires capturing information of the full document

Token Level Training



Document Level Needs

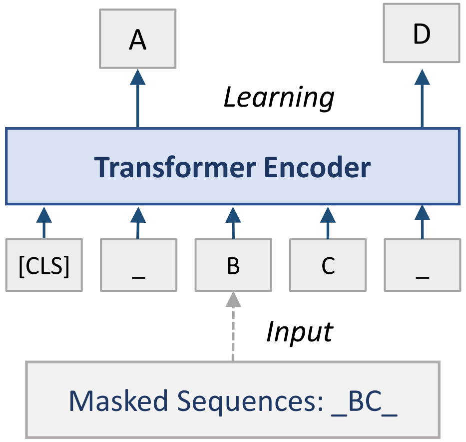


Mismatch #1: Local versus Global

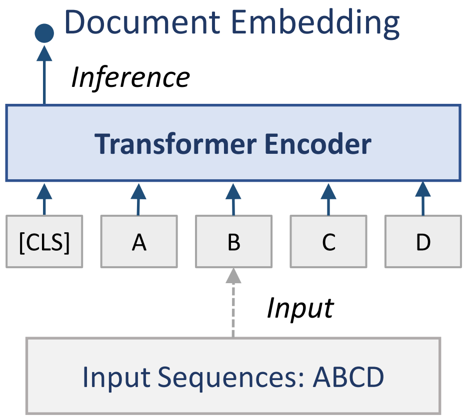
Language modeling is more about local contexts

Retrieval requires capturing information of the full document

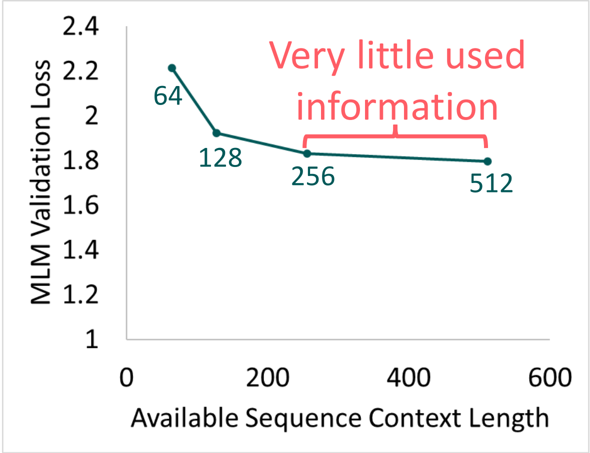
Token Level Training



Document Level Needs



BERT Base MLM Loss



Mismatch #1: Local versus Global

Language modeling is more about local contexts

- Long context methods mainly work on specific long-range tasks (not retrieval)
- "The longer context model retains strong performance on various general-purpose tasks" (LLaMA 2 [8])

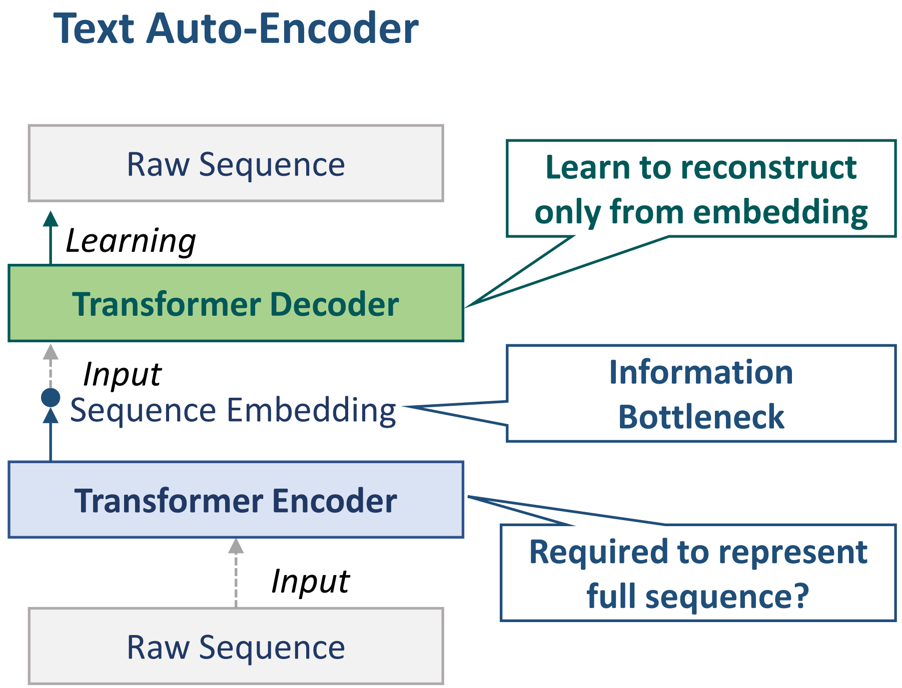
Context Length	Hella-Swag (0-shot)	NQ (64-shot)	TQA (64-shot)	GSM8K (8-shot)	Human-Eval (0-shot)
2k	75.1	25.5	53.7	4.9	7.9
4k	74.8	25.5	52.2	6.5	7.3

Table 2: LLaMA 2 performance on general-purpose tasks with different pretraining context length [8]

[8] Touvron et al. "LLaMA 2: Open Foundation and Fine-Tuned Chat Models". ArXiv 2023.

Mismatch #1 Solution: Auto-Encoder Training

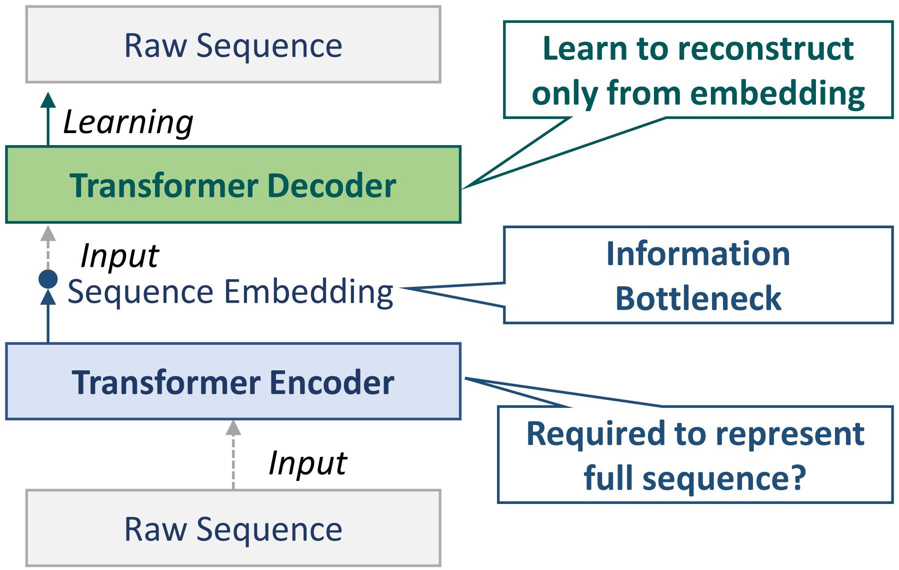
Reintroduce self-reconstruction loss on sequence embeddings to capture full sequence information [9]



Mismatch #1 Solution: Auto-Encoder Training

Reintroduce self-reconstruction loss on sequence embeddings to capture full sequence information [9]

Text Auto-Encoder



Reconstruction Loss:

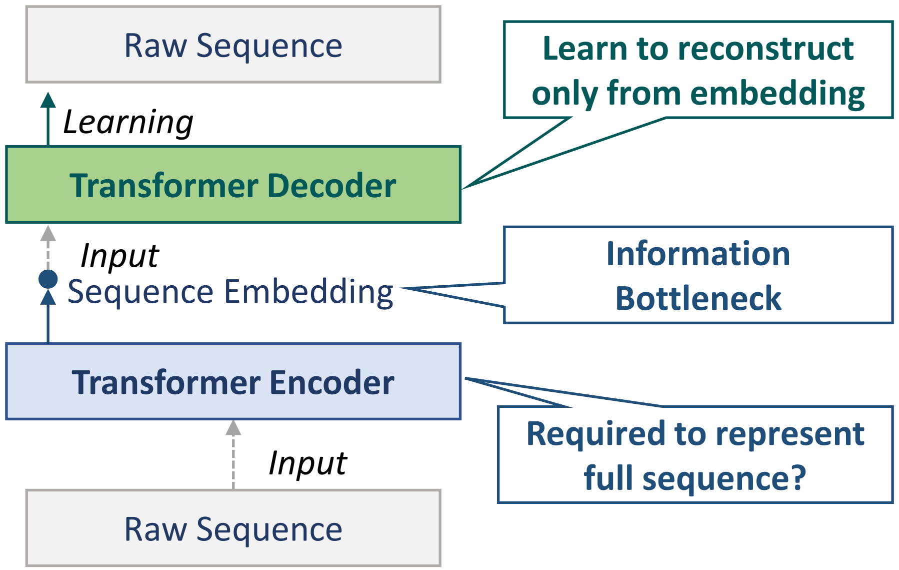
$$E_D[L_{dec}(X, \overset{\text{Decoder}}{\theta_{dec}})] = E_D\left[\sum_{t:1\sim n} -\log P\left(X_t \middle| X_{<t}, \overset{\text{Sequence Embedding}}{\overrightarrow{[CLS]}_{enc}}; \theta_{dec}\right)\right]$$

$$= \sum_{t:1\sim n} E_D[D_{KL}(\underbrace{P_D(X_t|X_{<t}, \overrightarrow{[CLS]}_{enc})}_{\text{Data Distribution}} \parallel \underbrace{P_{\theta_{dec}}(X_t|X_{<t}, \overrightarrow{[CLS]}_{enc})}_{\text{Decoder's Distribution}})] + \underbrace{H_D(X_t|X_{<t}, \overrightarrow{[CLS]}_{enc})}_{\text{Language Entropy}}$$

Mismatch #1 Solution: Auto-Encoder Training

Reintroduce self-reconstruction loss on sequence embeddings to capture full sequence information [9]

Text Auto-Encoder



Reconstruction Loss:

$$E_D[L_{dec}(X, \overset{\text{Decoder}}{\theta_{dec}})] = E_D\left[\sum_{t:1\sim n} -\log P\left(X_t \middle| X_{<t}, \overset{\text{Sequence Embedding}}{\overrightarrow{[CLS]}_{enc}}; \theta_{dec}\right)\right]$$

$$= \sum_{t:1\sim n} E_D[D_{KL}(\underbrace{P_D(X_t|X_{<t}, \overrightarrow{[CLS]}_{enc})}_{\text{Data Distribution}} \parallel \underbrace{P_{\theta_{dec}}(X_t|X_{<t}, \overrightarrow{[CLS]}_{enc})}_{\text{Decoder's Distribution}})] + \underbrace{H_D(X_t|X_{<t}, \overrightarrow{[CLS]}_{enc})}_{\text{Language Entropy}}$$

What if:

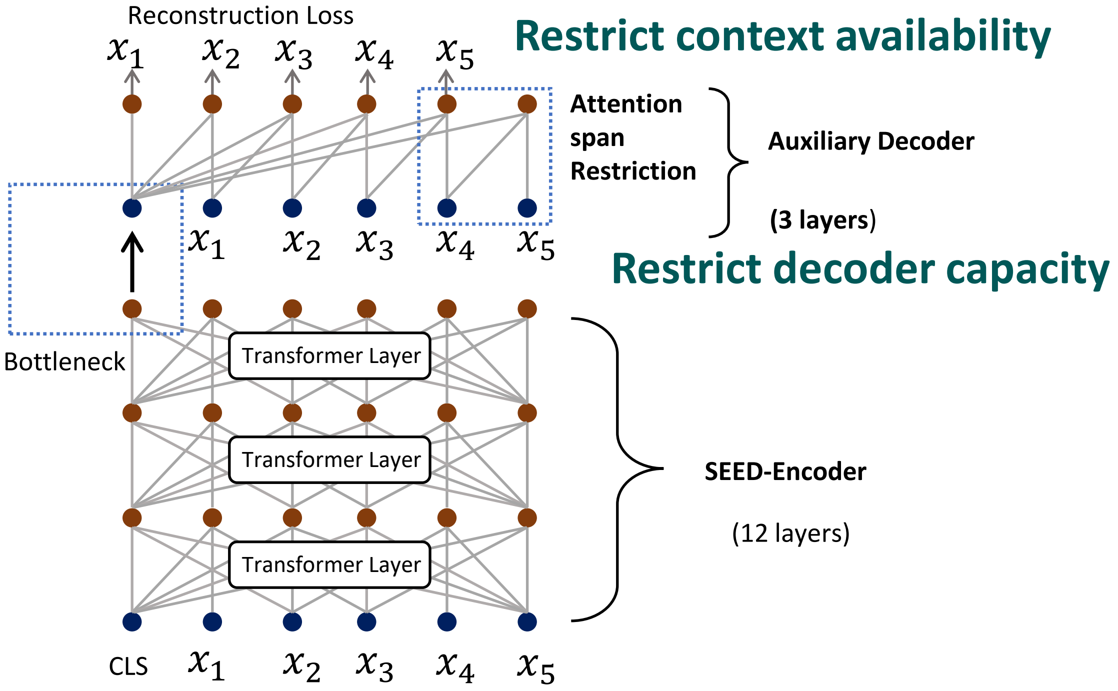
- Decoder is good at modeling language (GPT-*)?
- Language has strong patterns thus low entropy?

[9] Lu et al. "Less is More: Pre-train a Strong Text Encoder for Dense Retrieval Using a Weak Decoder". EMNLP 2021.

Mismatch #1 Solution: Auto-Encoder Training

Reintroduce self-reconstruction loss on sequence embeddings to capture full sequence information [9]

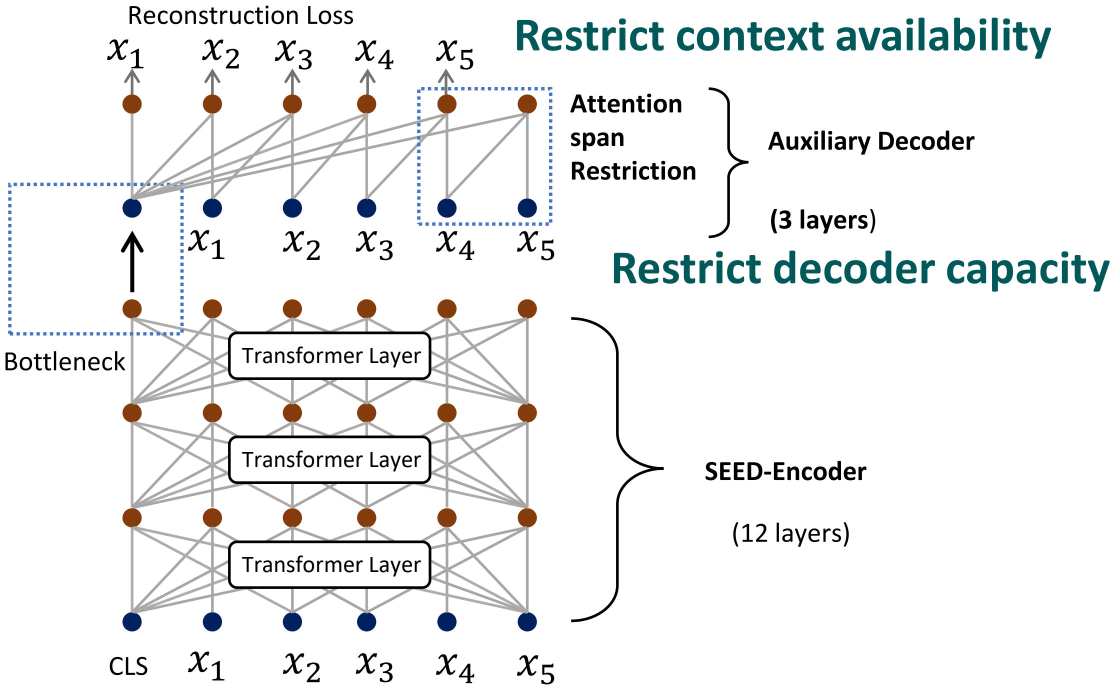
Information bottleneck on Document Encoding



Mismatch #1 Solution: Auto-Encoder Training

Reintroduce self-reconstruction loss on sequence embeddings to capture full sequence information [9]

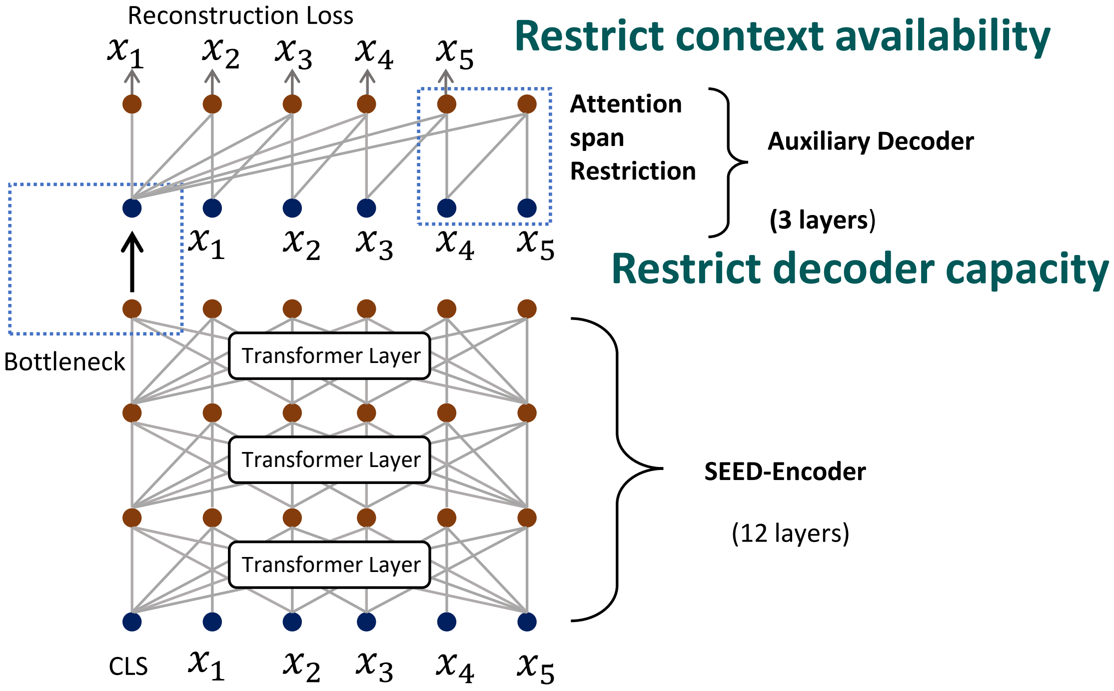
Information bottleneck on Document Encoding



Mismatch #1 Solution: Auto-Encoder Training

Reintroduce self-reconstruction loss on sequence embeddings to capture full sequence information [9]

Information bottleneck on Document Encoding



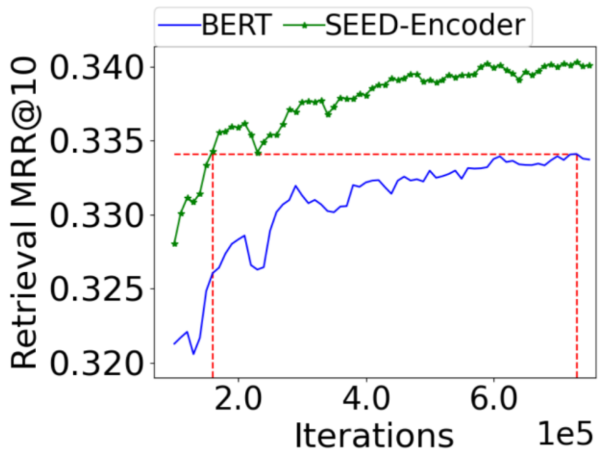
SEED-Encoder Pretraining

- Pretrain with Encoder with standard MLM
- Pretrain the Decoder with Auto-Regressive LM
- Two pretrained jointly:
    - Decoder pushes for better sequence encoding
    - Encoder is used in representation-centric tasks

Mismatch #1 Solution: Auto-Encoder Training

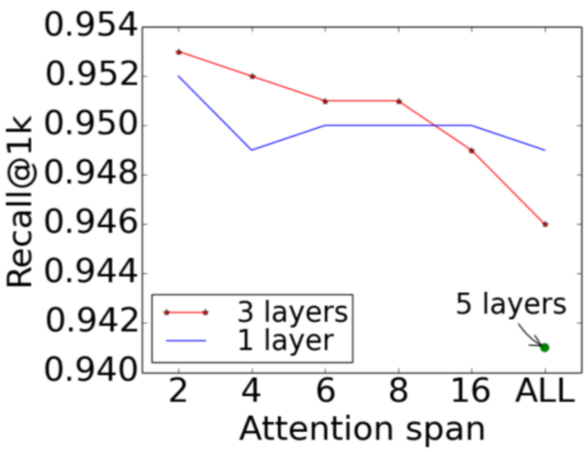
Better pretraining starting point for dense retrievers

MARCO Retrieval
(w.r.t. ANCE Fine-Tuning Steps)



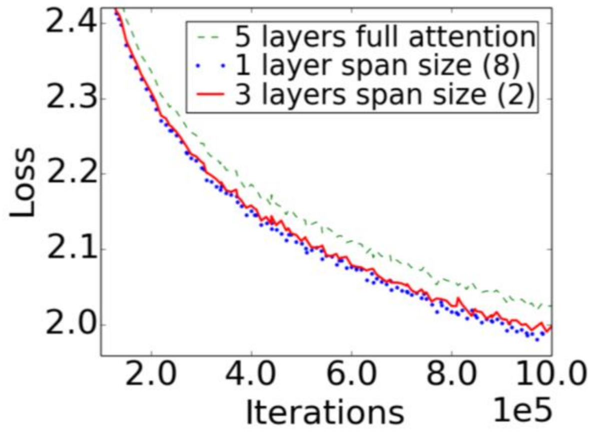
A Better Starting Point for Dense Retriever

MARCO Retrieval
(With ANCE)



Weaker Decoder Pushes for Better Encoder

Encoder Validation Loss
(During Pretraining)



Weaker Decoder Helps Encoder MLM Training

Mismatch #2: Anisotropy/Non-Uniformity

Zero-shot performance of pretrained embeddings on semantic text similarity (STS) tasks

- STS Task: producing a similarity score for a given pair of sentences
- Metric: by Pearson correlation with human rating (e.g., 5 being exact same meaning/paraphrase)

Model	STS12	STS13	STS14	STS15	STS16	STSb
Avg. GloVe embeddings	55.14	70.66	59.73	68.25	63.66	58.02
Avg. BERT embeddings	38.78	57.98	57.98	63.15	61.06	46.35
BERT CLS-vector	20.16	30.01	20.09	36.88	38.08	16.50

Table 3: BERT embedding similarity performances on STS tasks [10]

Much worse performance than GloVe Embeddings.

- [CLS] is near random.
- Mean-pooling over tokens is better but still much worse than word embeddings

[10] Reimers et al. “Sentence-BERT: Sentence Embeddings Using Siamese BERT-networks”. EMNLP 2019

Mismatch #2: Anisotropy/Non-Uniformity

The sequence embedding space of many pretrained LLMs are highly non-uniform

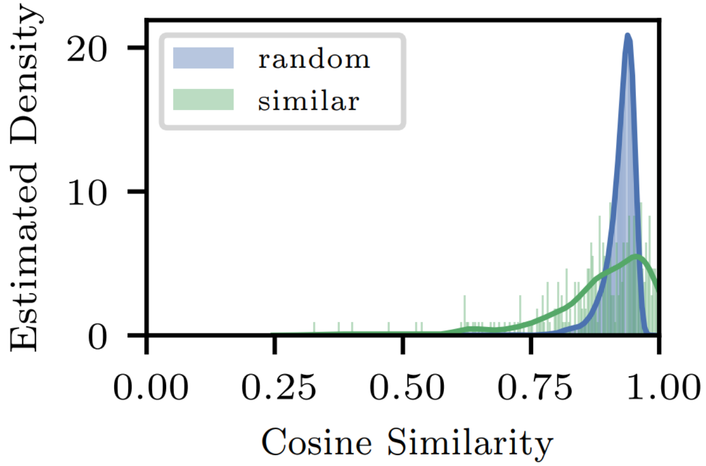


Figure 11: Similarity of RoBERTa $\overrightarrow{[\mathrm{CLS}]}$ on semantically similar and random pairs from STS-S [11]

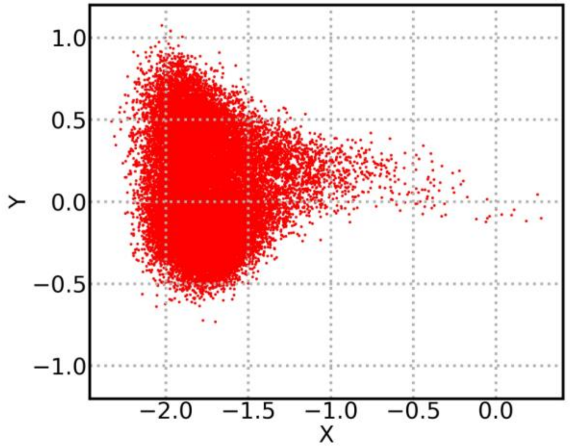


Figure 12: SVD 2-D mapping of word embeddings from Transformer trained on EN→DE [12]

Mismatch #2: Anisotropy/Non-Uniformity

The sequence embedding space of many pretrained LLMs are highly non-uniform

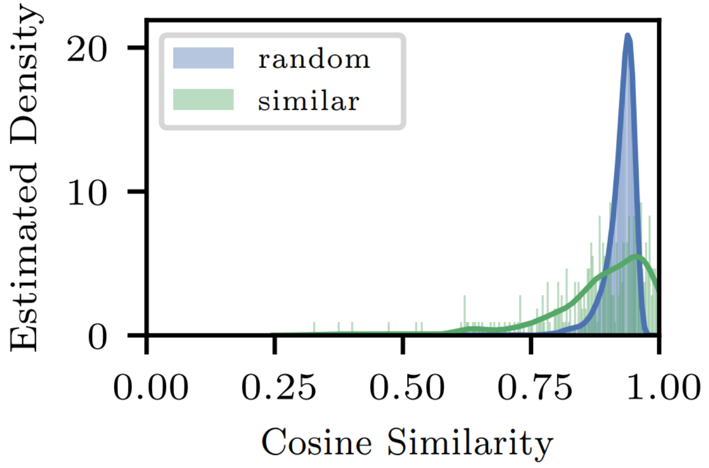


Figure 11: Similarity of RoBERTa $\overrightarrow{[\mathrm{CLS}]}$ on semantically similar and random pairs from STS-S [11]

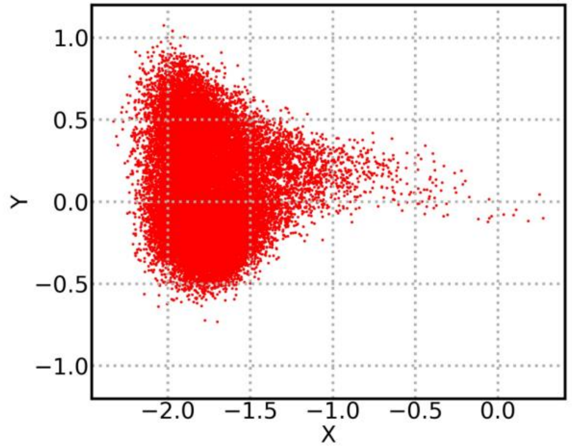


Figure 12: SVD 2-D mapping of word embeddings from Transformer trained on EN→DE [12]

Most rare tokens are pushed to a narrow cone in the space, and [CLS] is a rare token in learning

- Every training signal pushes all negatives away from the positive
- Rare tokens (without much or any positive pulls) are pushed away from all positives, into a narrow cone

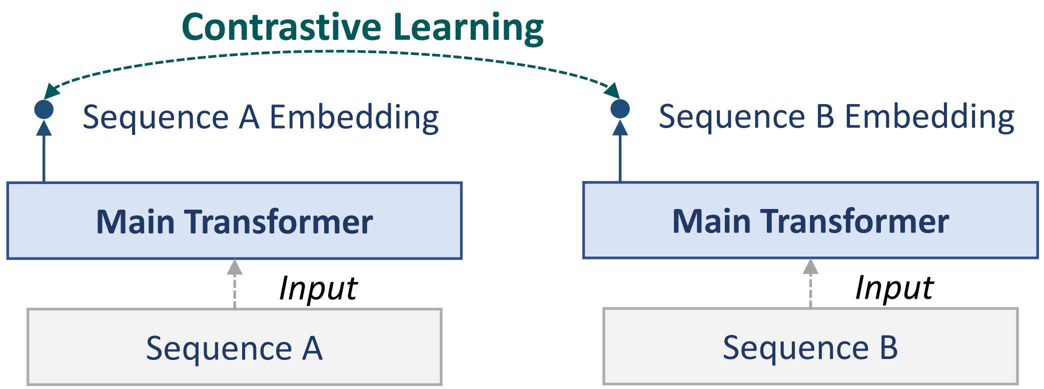
[11] Meng et al. "COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining". NeurIPS 2021.

[12] Gao et al. "Representation Degeneration Problem in Training Neural Language Generation Methods". ICLR 2019.

Mismatch #2 Solution: Sequence Contrastive Learning

Pretraining sequence representations with Sequence Contrastive Learning (SCL) [11]

Adding pretraining task: $L_{\text{SCL}} = \text{E}\left(\frac{\exp(\cos(\boldsymbol{s},\boldsymbol{s}^{+}))}{\exp(\cos(\boldsymbol{s},\boldsymbol{s}^{+}))+\sum_{s^-}\exp(\cos(\boldsymbol{s},\boldsymbol{s}^{-}))}\right)$



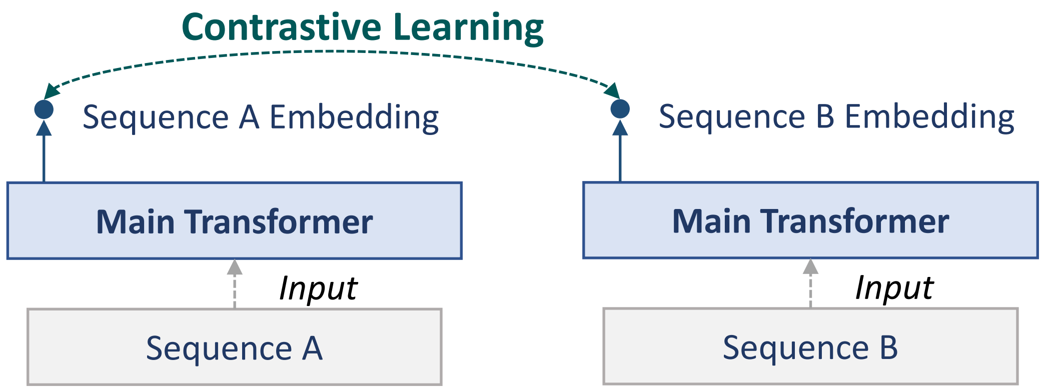
Mismatch #2 Solution: Sequence Contrastive Learning

Pretraining sequence representations with Sequence Contrastive Learning (SCL) [11]

Adding pretraining task: $L_{\mathrm{SCL}} = \mathrm{E}\left(\frac{\exp(\cos(\boxed{\boldsymbol{s},\boldsymbol{s}^{+}}))}{\exp(\cos(\boldsymbol{s},\boldsymbol{s}^{+}))+\sum_{s^{-}}\exp(\cos(\boxed{\boldsymbol{s},\boldsymbol{s}^{-}}))}\right)$

Embeddings of positive contrast sequence pairs

Embeddings of negative sequence pairs



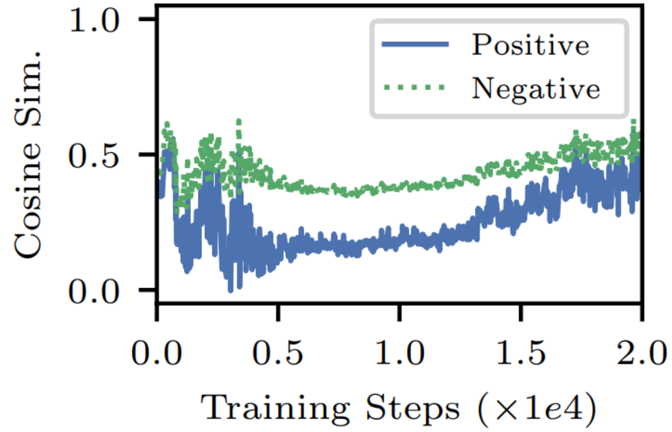
Construction of positive contrast sequence pairs:

- *Data augmentation*: cropping [11], random replacement, back translation, different dropout (SimCSE), etc.
- *Unsupervised pairs*: co-occurrence in doc (co-doc), etc.
- *Supervisions*: Web QA pairs, search query-clicked docs…

Mismatch #2 Solution: Sequence Contrastive Learning

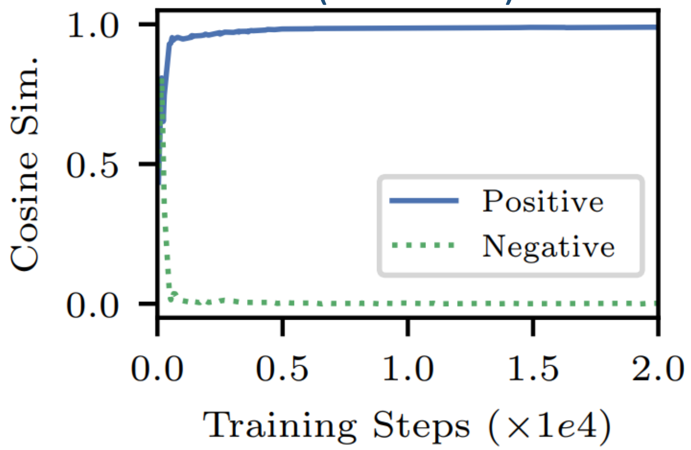
Recalibration of the embedding space, e.g., using cropped sequence pairs (90% overlap)

Training Pair Similarity
(Without SCL)



Failed without SCL
(Although 90% overlap!)

Training Pair Similarity
(With SCL)

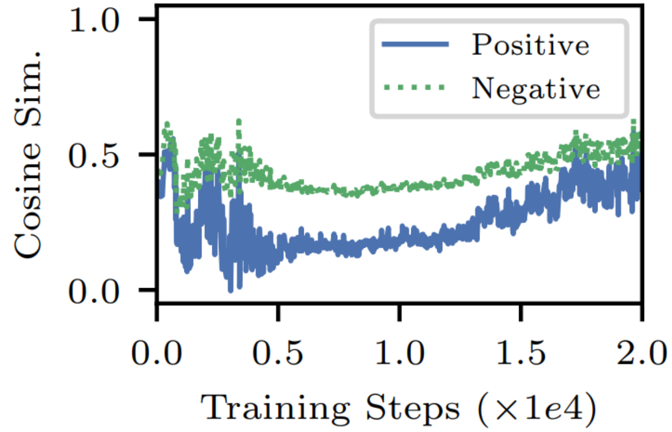


Easy-to-Learn Task
(90% overlap, after all)

Mismatch #2 Solution: Sequence Contrastive Learning

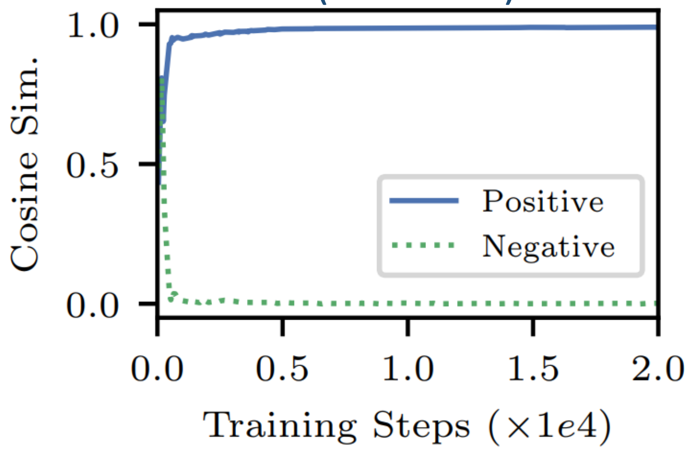
Recalibration of the embedding space, e.g., using cropped sequence pairs (90% overlap)

Training Pair Similarity
(Without SCL)



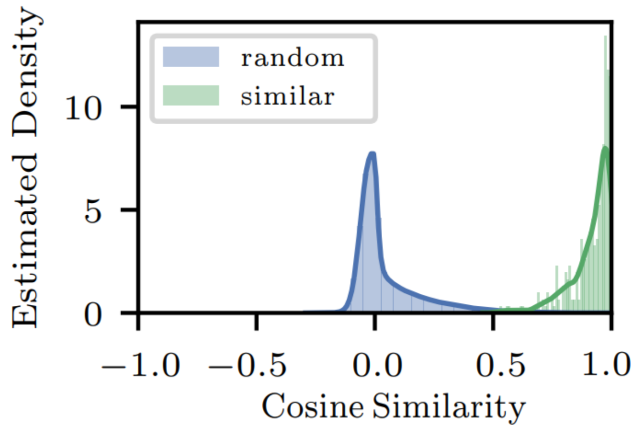
Failed without SCL
(Although 90% overlap!)

Training Pair Similarity
(With SCL)



Easy-to-Learn Task
(90% overlap, after all)

STS-B Similarity
(With SCL)



Effective Calibration
& Good Zero-Shot Ability

Decent zero-shot performance on many sequence similarity tasks and non-random performance on retrieval

[11] Meng et al. "COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining". NeurIPS 2021.

Deeper Look into Contrastive Learning

Two forces in contrastive learning: Alignment and Uniformity [13]

$$L_{\text{SCL}} = \mathrm{E}\left(\frac{\exp(\cos(\boldsymbol{s},\boldsymbol{s}^{+}))}{\exp(\cos(\boldsymbol{s},\boldsymbol{s}^{+}))+\sum_{s^{-}}\exp(\cos(\boldsymbol{s},\boldsymbol{s}^{-}))}\right)$$

$$\sim \underbrace{\cos(\boldsymbol{s},\boldsymbol{s}^{+})}_{\text{Align positive pairs together}} + \underbrace{\log\left(\exp(\cos(\boldsymbol{s},\boldsymbol{s}^{+})) + \sum_{s^{-}}\exp(\cos(\boldsymbol{s},\boldsymbol{s}^{-}))\right)}_{\text{Uniformly spread random pairs in the space}}$$

- Proof in Wang et al. [12] that, if exist, perfectly aligned/uniform encoders minimize the two terms
- Note: here negatives are sampled uniformly, not from a long tail distribution

[13] Wang et al. "Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere". ICML 2020.

Deeper Look into Contrastive Learning

Two forces in contrastive learning: Alignment and Uniformity [13]

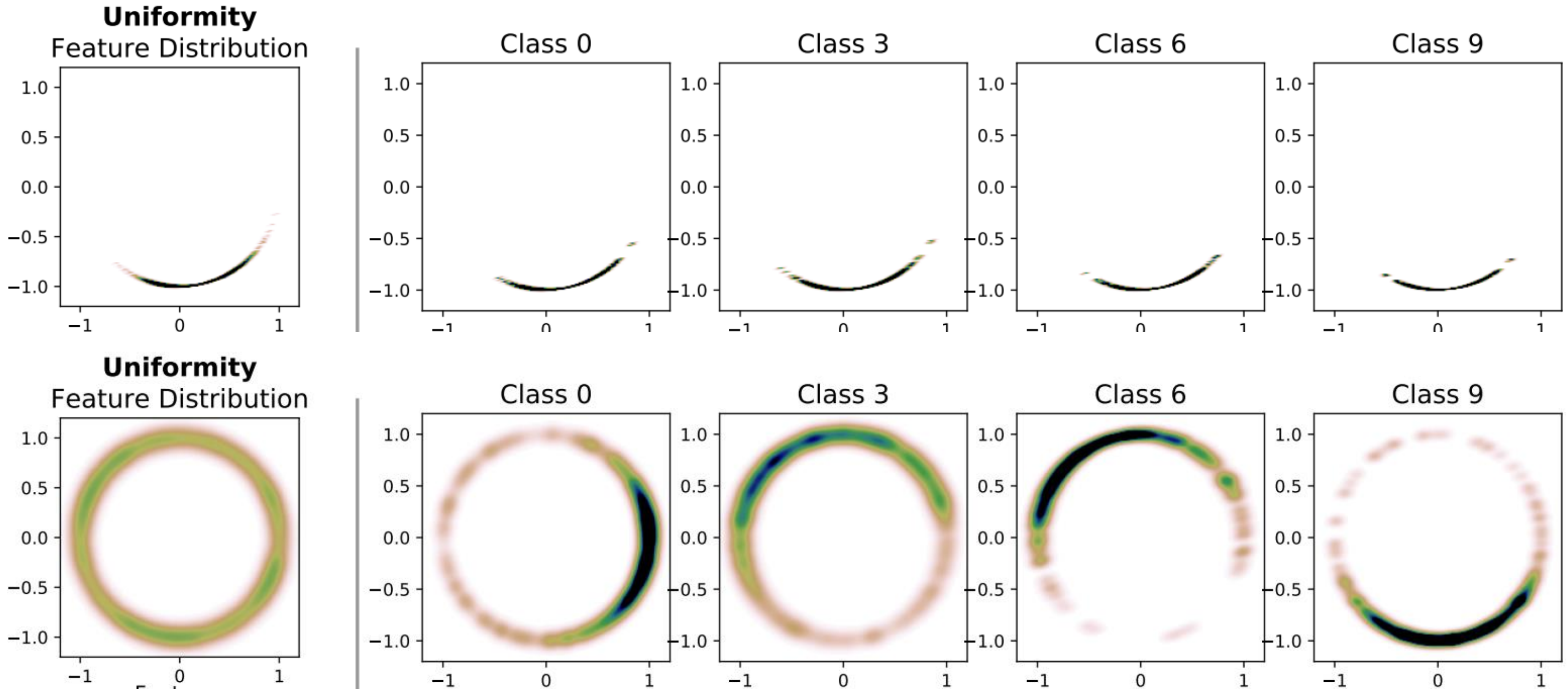


Figure 13: Uniformity of image features in CIFAR-10 from random network (top) and unsupervised contrastive learning (bottom) [12]

Mismatch #3: Alignments

What information does unsupervised contrastive pairs bring in to align the embedding space?

Method	Sequence A	Sequence B
SimCSE	The Steelers enjoy a large, widespread fanbase nicknamed Steeler Nation.	The Steelers enjoy a large, widespread fanbase nicknamed Steeler Nation.
Inverse Cloze Task (ICT)	The Steelers enjoy a large, widespread fanbase nicknamed Steeler Nation.	They currently play their home games at Acrisure Stadium on Pittsburgh's North Side in the North Shore neighborhood,
Cropping Augmentation	The Steelers enjoy a large, widespread fanbase nicknamed ___	_____ enjoy a large, widespread fanbase nicknamed Steeler Nation.
Co-document	The Steelers enjoy a large, widespread fanbase nicknamed Steeler Nation.	In the NFL's "modern era" (since the AFL–NFL merger in 1970) the Steelers have posted the best record in the league.

Very limited semantic signals in the alignment for search relevance

- Either strong term overlaps or loosely correlated

Mismatch #3 Solution: Weak Supervision from Web Graph

Leverage Anchor Texts and the document they point to pseudo query-relevant document pairs

Method	Sequence A	Sequence B
Anchor-Document	Vegetarian Society of Ireland	The Vegetarian Society of Ireland is a registered charity. Our aim is to increase awareness of vegetarianism in relation to health,
Actual Argument Retrieval Data	Becoming a vegetarian is an environmentally friendly thing to do.	Health general weight philosophy ethics You don't have to be vegetarian to be green. Many special environments have been created by

Mismatch #3 Solution: Weak Supervision from Web Graph

Leverage Anchor Texts and the document they point to pseudo query-relevant document pairs

Method	Sequence A	Sequence B
Anchor-Document	Vegetarian Society of Ireland	The Vegetarian Society of Ireland is a registered charity. Our aim is to increase awareness of vegetarianism in relation to health,
Actual Argument Retrieval Data	Becoming a vegetarian is an environmentally friendly thing to do.	Health general weight philosophy ethics You don't have to be vegetarian to be green. Many special environments have been created by

Web graph and anchor information is widely used in many web and search applications

- Determine the importance of a web page (Page Rank)
- Enrich the representation of a document , using 3rd party information (Document Expansion)
- Serve as pseudo queries for feature-based ranking models

[14] Nie et al. "Unsupervised Dense Retrieval Training with Web Anchors". SIGIR 2023.

Mismatch #3 Solution: Weak Supervision from Web Graph

Leverage Anchor Texts and the document they point to pseudo query-relevant document pairs

Method	Sequence A	Sequence B
Anchor-Document	Vegetarian Society of Ireland	The Vegetarian Society of Ireland is a registered charity. Our aim is to increase awareness of vegetarianism in relation to health,
Actual Argument Retrieval Data	Becoming a vegetarian is an environmentally friendly thing to do.	Health general weight philosophy ethics You don't have to be vegetarian to be green. Many special environments have been created by

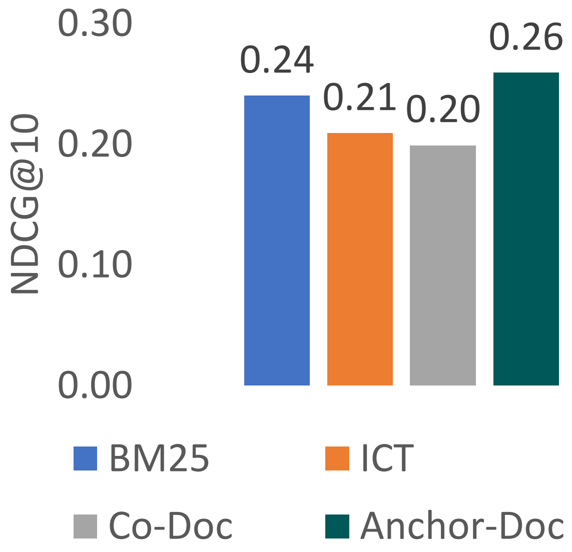


Figure 14: MARCO NDCG@10 of BM25 and dense retrievers trained by different unsupervised signals

Anchor-Doc the only unsupervised signal source outperforms BM25

- Data cleaning required to filter out functional anchors, e.g., "homepage"

A widely useful information in standard web search

- Page Rank, Document Expansion, etc.

Still a weakly supervised method, rather than a pretraining method

- Behavior closer to weak supervision/transfer learning, not pretraining

Mismatch Between LLM and Retrieval: Recap

Mismatch	Local vs. Global	Non-uniformity	Semantic Alignment
Solution?	Auto-Encoder	Contrastive Training	Weak Supervision

Mismatch Between LLM and Retrieval: Recap

Mismatch	Local vs. Global	Non-uniformity	Semantic Alignment
Solution?	Auto-Encoder	Contrastive Training	Weak Supervision

We are still not seeing the emergent power of LLMs in embedding-based retrieval

- The fact we need these solutions/mitigations shows there is something missing

Auto-regressive LM + scaling up solved a lot of problems, but not everything

- Web search is perhaps the biggest money-making AI application, yet not fully covered by GPT-X

“Bitter lesson”, more compute and large-scale trump specific designs, is deemed to happen

- But that may not achieved all by current language models

Quiz: Why data augmentation based contrastive learning work better in vision tasks like ImageNet classification but not as much in search?