Biomedical Neural Information Retrieval

Ronak Pradeep

Outline

- Part 1: Background
  (text ranking, IR, transformers)
- Part 2: Expando-Mono-Duo-T5
  (passage, document)
- Part 3: Biomedical Search – TREC-COVID
- Part 4: Vera: Reducing Harmful Misinformation in Neural IR
- Part 5: Clinical Trial Matching

Background

Text Ranking

Ad Hoc Retrieval
Transformers

Focus: Ad Hoc Retrieval

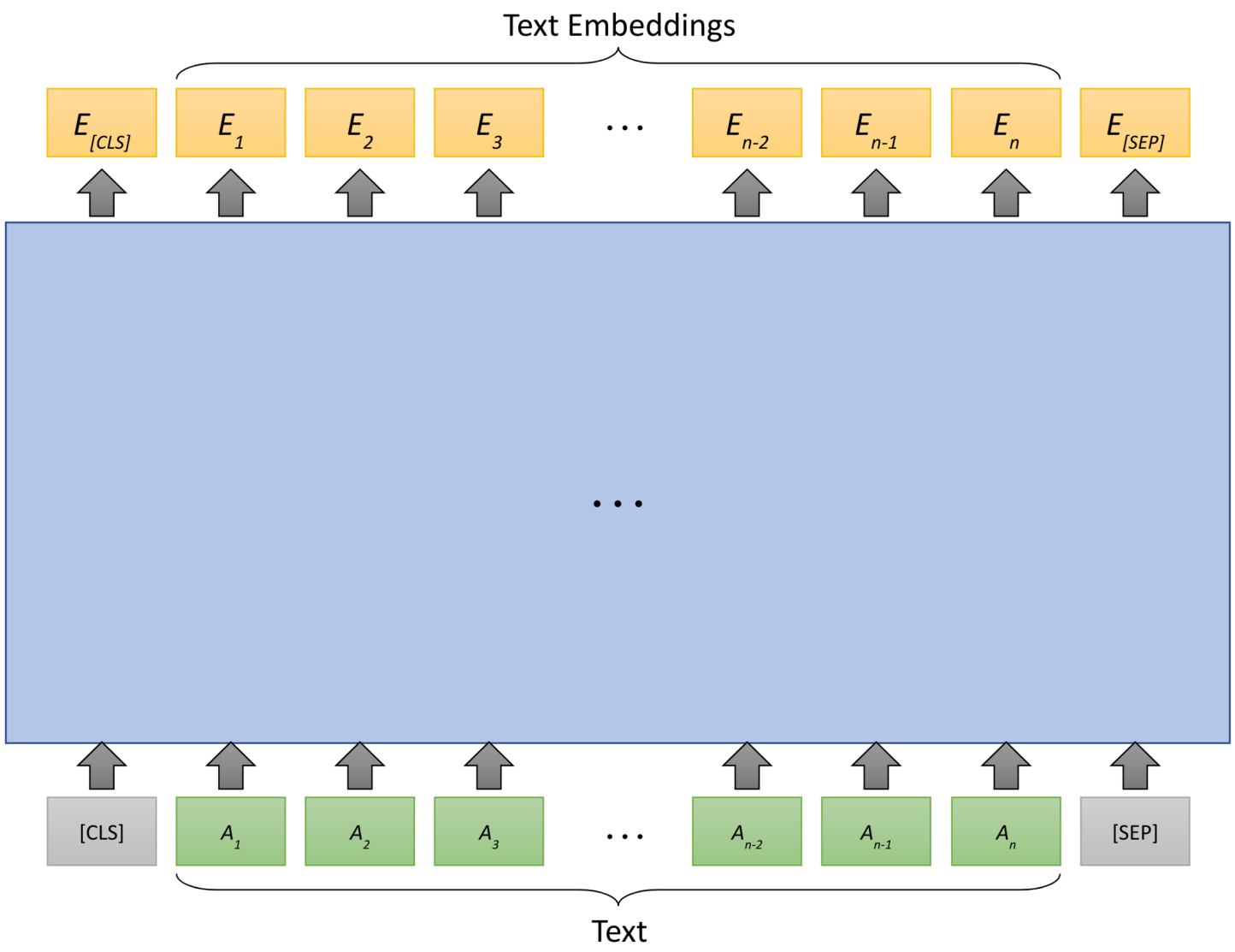
Given: query *q*

collection of texts

Return: a ranked list of *k* texts $d_1 \ldots d_k$

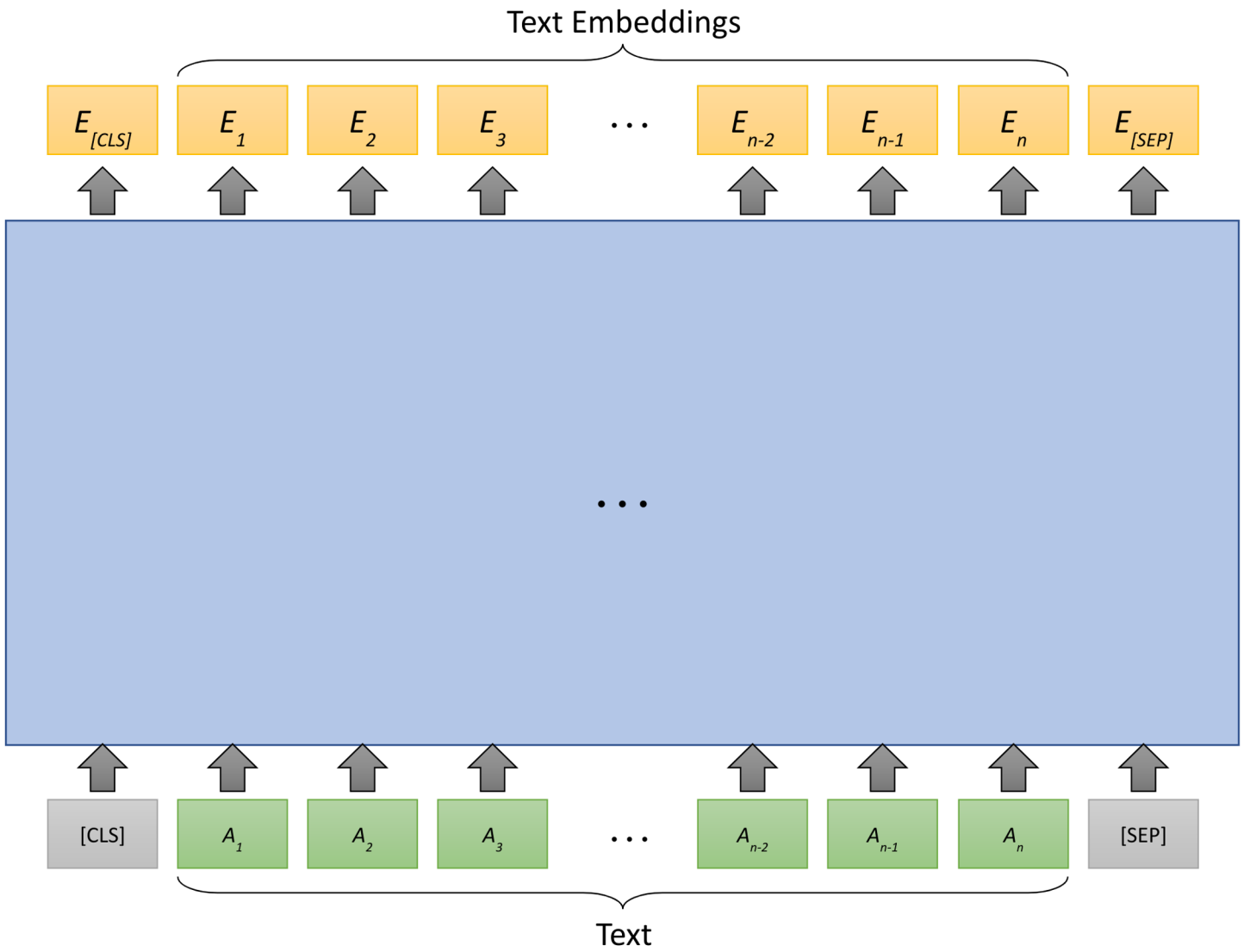
Maximizing: a metric of interest

Transformers



Pretrained Transformers

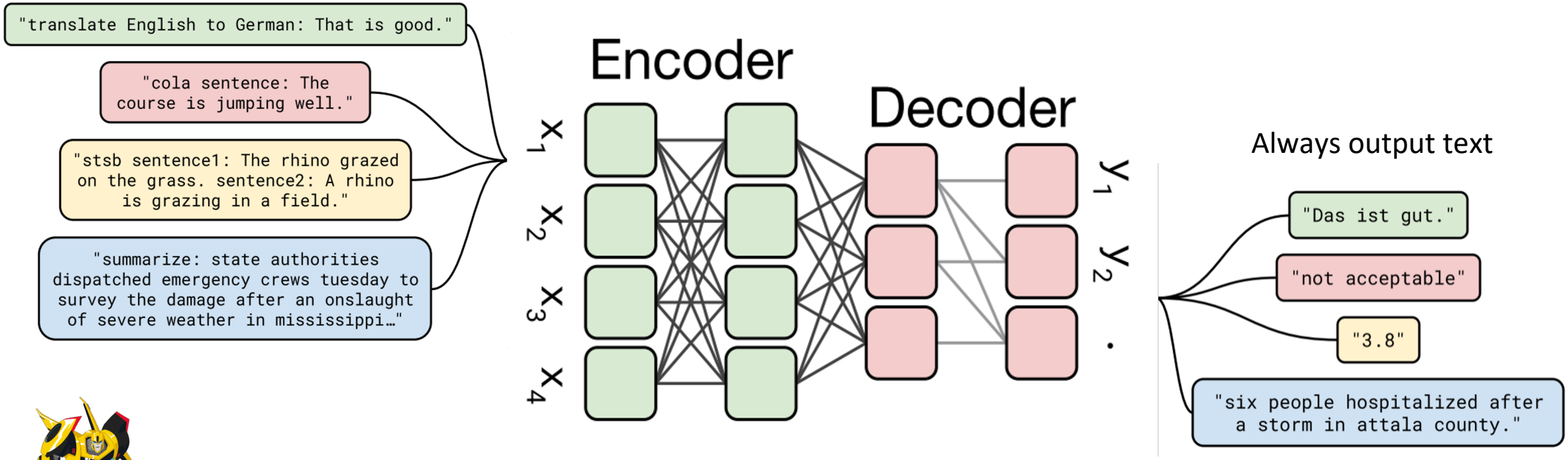
Initialized via pretraining



T5 (2020 ➡️ ?)

Welcome T5

<u>Pretraining Dataset</u>
C4 (Colossal Clean Crawled Corpus)
Multitask Pretraining Mixture



Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." JMLR 2020.

IR Background

Unsupervised ranking methods

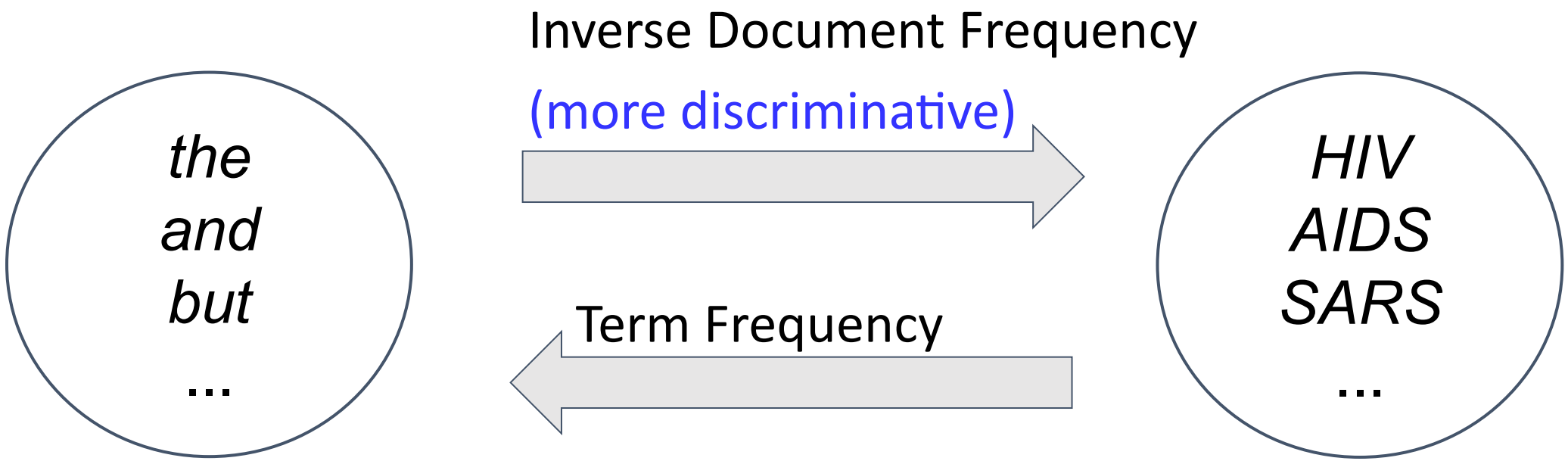
Unsupervised Ranking Methods

term score

$$S(q, d) = \sum_{t \in q \cap d} f(t)$$

input

Unsupervised Ranking Methods



Unsupervised Ranking Methods – TF-IDF

$$S(q, d) = \sum_{t \in q \cap d} \mathrm{TF}(t) \times \mathrm{IDF}(t)$$

Sparse representations

Unsupervised Ranking Methods – BM25

IDF component TF component

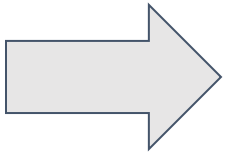
$$\mathrm{BM25}(q,d) = \sum_{t \in q \cap d} \log \frac{N - \mathrm{df}(t) + 0.5}{\mathrm{df}(t) + 0.5} \cdot \frac{\mathrm{tf}(t,d) \cdot (k_1 + 1)}{\mathrm{tf}(t,d) + k_1 \cdot \left(1 - b + b \cdot \frac{l_d}{L}\right)}$$

Vocabulary Mismatch

document ranking

web search

Expand



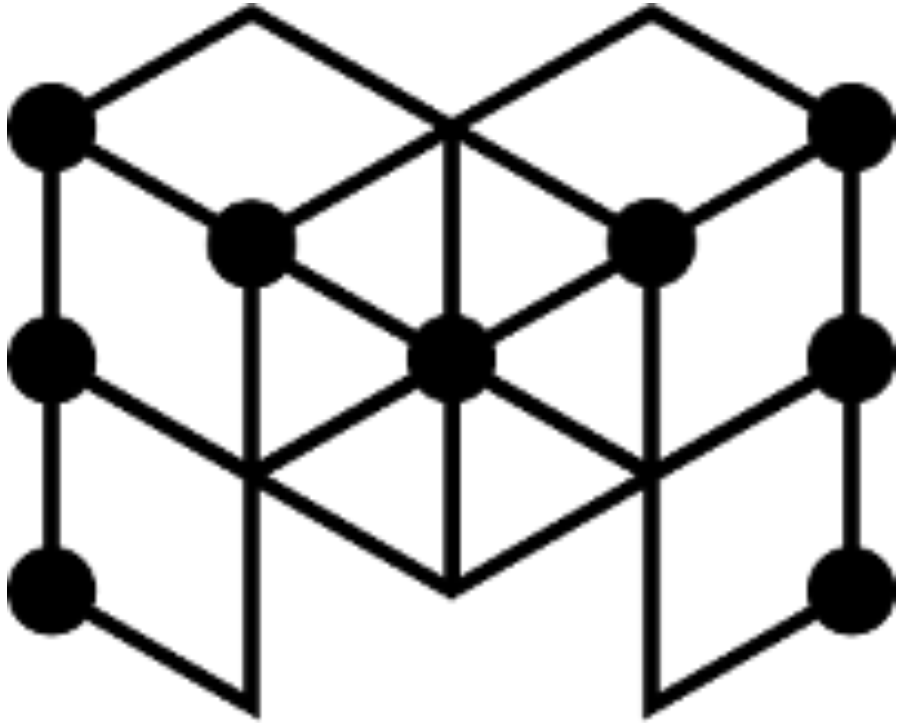
query, search,

ranking, retrieval, …

Enrich query or document representations → move beyond exact matching

- Unsupervised: pseudo-relevance feedback – RM3
- Later: document expansion with a neural method

Collections



MS MARCO

v1 Passage

v1 Passage Dataset

Query: when is the hottest month in washington dc?

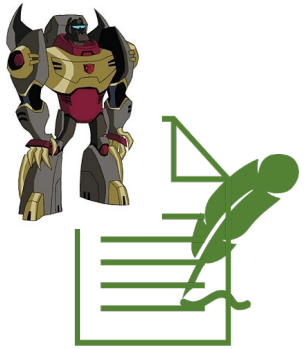
Relevant Passage: July is the hottest month in Washington DC with an average temperature of 27C (80F) and the coldest is January at 4C (38F) …

Non-Relevant Passage: April is a cruel time. Even though the sun may shine. And world looks in the shade as it slowly comes away.

- Large Training Set? 530K Pairs.
- Corpus? 9M Passages from BING search engine results.
- Passages? Mean length of 56 tokens, not too large for Transformers!
- Sparse Labels ☹ ~1 relevant "labelled" passage per query.

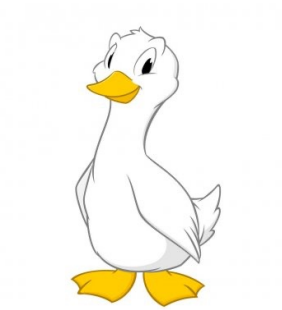
Expando-Mono-Duo-T5

Expando-Mono-Duo-T5



H_{-1}
document
expansion

→



H_0
keyword
search

top-k_0
passages
→

H_1
pointwise
reranker

top-k_1
passages
→

H_2
pairwise
reranker

multi-stage ranking architecture

keyword-based 1st stage retrieval (H_0)
two reranking stages (H_1 and H_2)

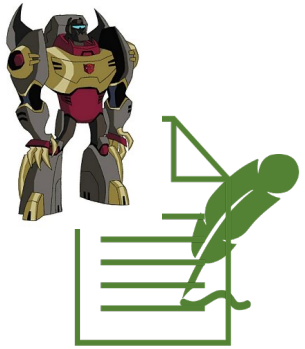
pre-indexing document enrichment

Pradeep et al. The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models. arXiv:2101.05667 (2021).

doc2query

Expando-Mono-Duo-T5

What queries are relevant for this document?



H_{-1}
document
expansion

H_0
keyword
search

H_1
pointwise
reranker

H_2
pairwise
reranker

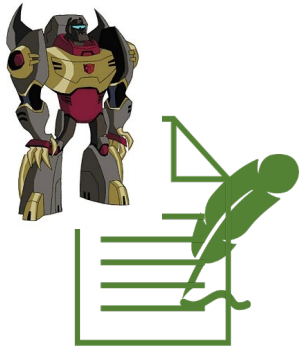
Original Corpus

July is the hottest month in Washington DC with an average temperature of 27C (80F) and the coldest is January at 4C (38F) with the most daily sunshine hours at 9 in July. The wettest month is May with an average of 100mm of rain.

doc2query

Expando-Mono-Duo-T5

What queries are relevant for this document?



H_{-1}
document
expansion

H_0
keyword
search

H_1
pointwise
reranker

H_2
pairwise
reranker

what is the weather in washington dc?
when is the hottest month in washington dc?
…

Original Corpus

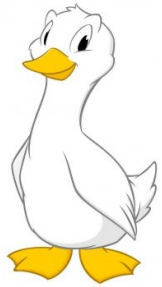
Expanded Corpus

Passed onto Anserini

July is the hottest month in Washington DC with an average temperature of 27C (80F) and the coldest is January at 4C (38F) with the most daily sunshine hours at 9 in July. The wettest month is May with an average of 100mm of rain.

Expando-Mono-Duo-T5

H_{-1}
document
expansion



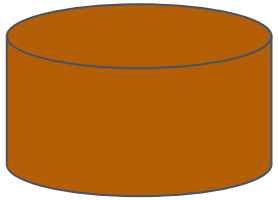
H_0
keyword
search

H_1
pointwise
reranker

H_2
pairwise
reranker

Anserini / Pyserini

(keyword) query



top-k_0 texts

Expando-Mono-Duo-T5

Is *this* document relevant?

H_{-1}
document
expansion

→

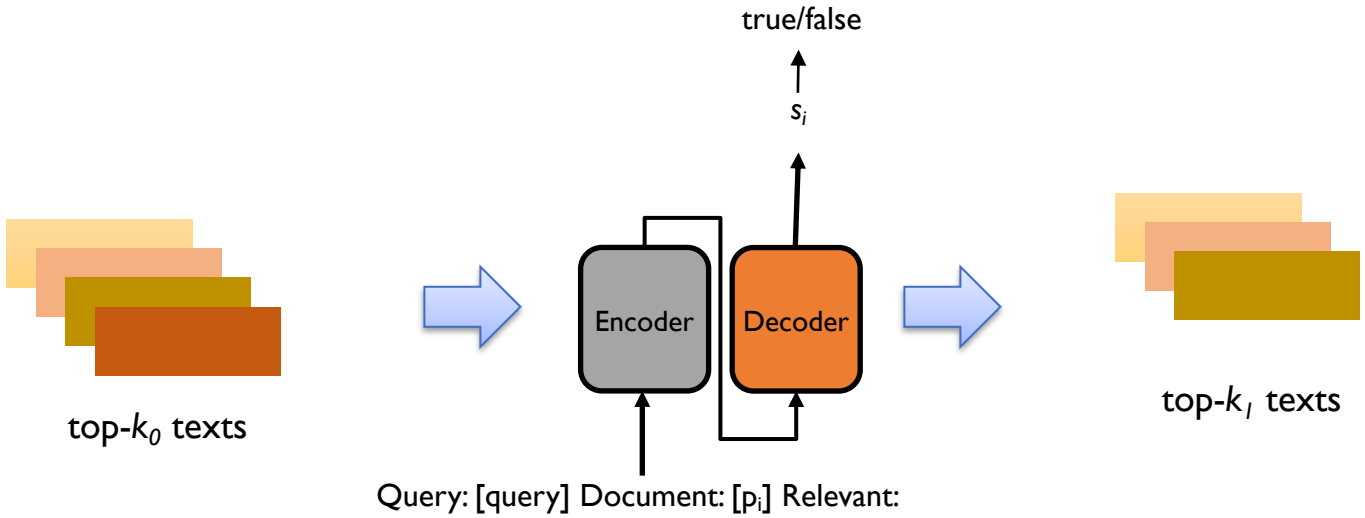
H_0
keyword
search

→

H_1
pointwise
reranker

→

H_2
pairwise
reranker



Nogueira et al. Document Ranking with a Pretrained Sequence-to-Sequence Model. Findings of EMNLP 2020.

Expando-Mono-Duo-T5

Is *this* document more relevant than *that* document?

H_{-1}
document
expansion

→

H_0
keyword
search

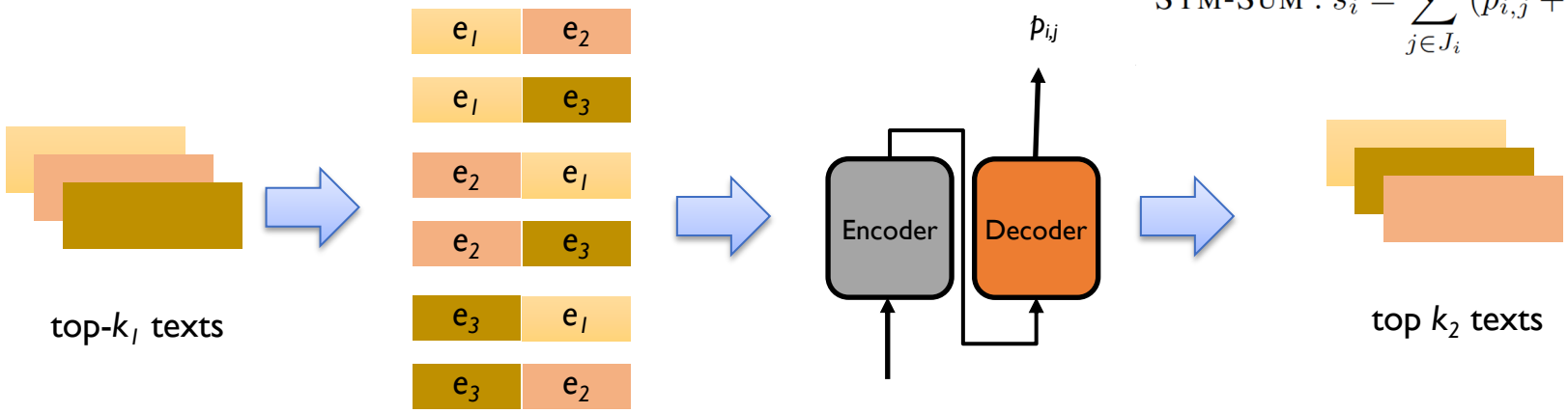
→

H_1
pointwise
reranker

→

H_2
pairwise
reranker

$$\text{SYM-SUM}: s_i = \sum_{j \in J_i} (p_{i,j} + (1 - p_{j,i}))$$



Pradeep et al. The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models. arXiv:2101.05667 (2021).

Results – MS MARCO Passage Dataset

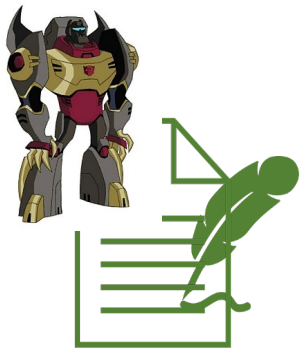
Model					Dev	Test
	Expand-Mono-Duo Variants					
	H_{-1}	H_0	H_1	H_2		
(1)	-	BM25	-	-	0.184	0.184
(2)	-	BM25	monoT5-base	-	0.381	-
(3)	-	BM25	monoT5-large	-	0.393	-
(4)	-	BM25	monoT5-3B	-	0.398	-
(5)	doc2query-T5	BM25	-	-	0.277	0.272
(6)	doc2query-T5	BM25	monoT5-3B	-	0.409	-
(7)	doc2query-T5	BM25	monoT5-3B	duoT5-3B	0.420	0.408
(8)	-	BM25	monoT5-base	duoT5-base	0.392	-
(9)	doc2query-T5	BM25	monoT5-3B	duoT5-base	0.409	-
(10)	BM25 + BERT-large [40]				0.372	0.365
(11)	TFR-BERT Ensemble [15]				0.405	0.395

SOTA!

Towards Longer Texts

Passages ➡️ Documents

Documents: too long for these poor Transformers!



H_{-1}
document
expansion

→

H_0
keyword
search

→

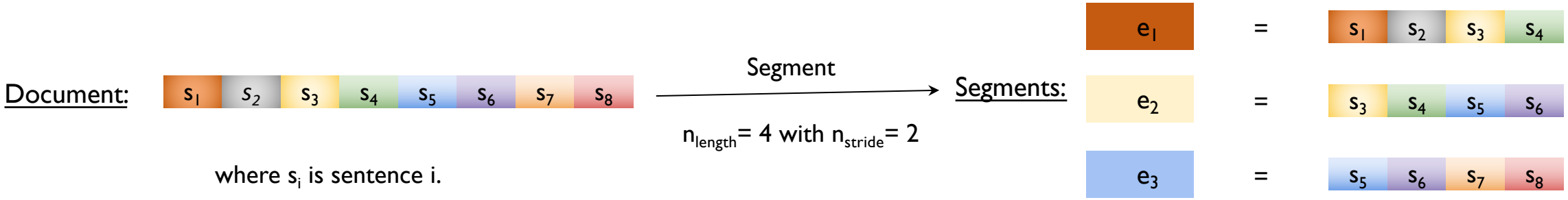
H_1
pointwise
reranker

→

H_2
pairwise
reranker

1. T5 can handle > 512 input tokens but pretrain/finetune step still use a maximum of 512 input tokens.
2. Running with arbitrarily long input sequences can be very computationally inefficient.

Our Solution: Sliding Window Segmentation! Use a window of size n_{length} sentences and a stride of n_{stride}



Expando-Mono-Duo-T5

Is *this* document relevant?

H_{-1}
document
expansion

→

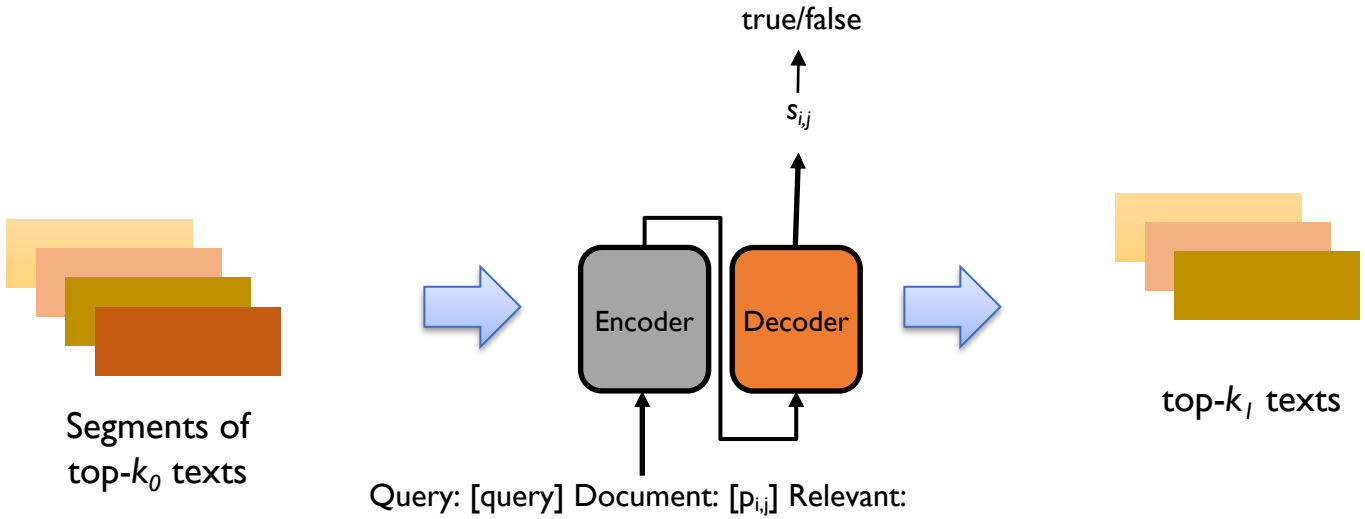
H_0
keyword
search

→

H_1
pointwise
reranker

→

H_2
pairwise
reranker



$s_i' = max\ s_{i,j}$

e_i : *highest scoring segment in the text*

Nogueira et al. Document Ranking with a Pretrained Sequence-to-Sequence Model. Findings of EMNLP 2020.

Expando-Mono-Duo-T5

Is *this* document more relevant than *that* document?

H_{-1}
document
expansion

→

H_0
keyword
search

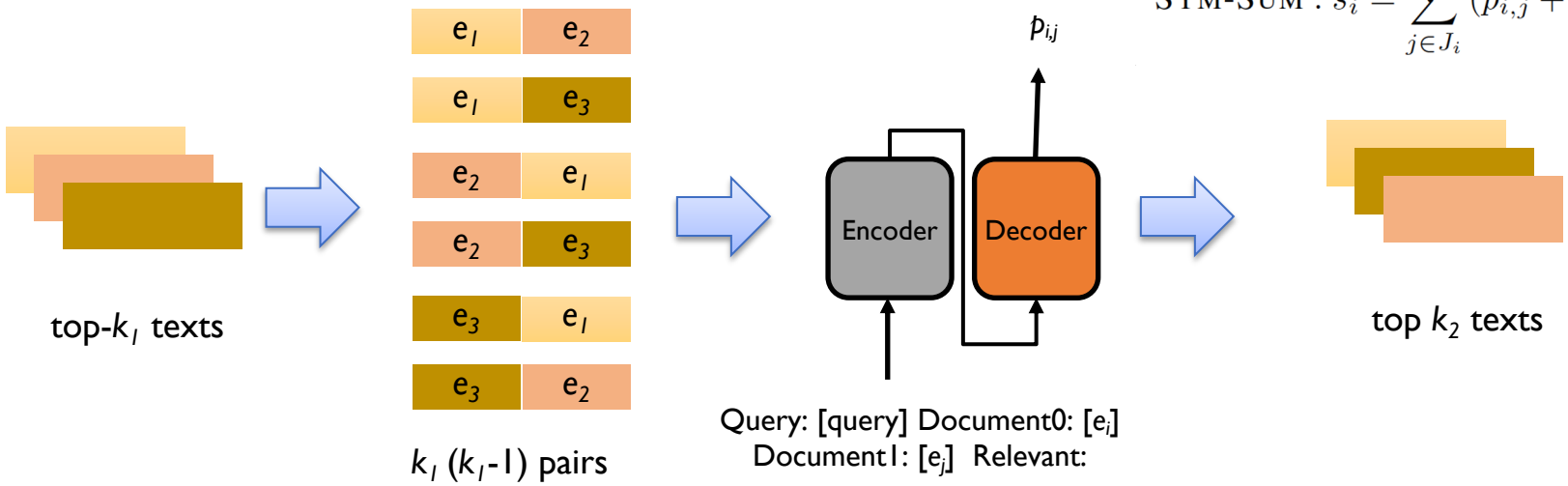
→

H_1
pointwise
reranker

→

H_2
pairwise
reranker

$$\text{SYM-SUM}: s_i = \sum_{j \in J_i} (p_{i,j} + (1 - p_{j,i}))$$



Pradeep et al. The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models. arXiv:2101.05667 (2021).

Biomedical Search

TREC-COVID

- Goal: Evaluate systems that help stakeholders access reliable evidence.
- Evolving nature of scientific literature and information needs.
- Series of 5 rounds, each using CORD-19 corpus at a snapshot in time.
- 50 topics by round 5.

Question
How does the coronavirus respond to changes in the weather?
Has social distancing had an impact on slowing the spread of COVID-19?
How long can the coronavirus live outside the body?
What is known about those infected with Covid-19 but are asymptomatic?
What evidence is there for the value of hydroxychloroquine in treating Covid-19?

SLEDGE

To improve zero-shot biomedical search:

- Filter MS MARCO training queries that contain biomedical terms (MED-MARCO).
- Finetune model first on MS MARCO and then on MED-MARCO.

MacAvaney et al. SLEDGE-Z: A Zero-Shot Baseline for COVID-19 Literature Search. *EMNLP 2020*

SLEDGE

Model	Training	Including Unjudged				Judged Only		
		nDCG@10	P@5	P@5 (F)	J@10	nDCG@10	P@5	P@5 (F)
BM25	-	* 0.368	* 0.469	* 0.331	75%	* 0.436	* 0.520	* 0.383
+ BERT	MSM	* 0.547	* 0.617	* 0.480	83%	* 0.617	* 0.703	* 0.549
+ BERT	MedM	0.625	* 0.697	* 0.571	92%	0.657	* 0.737	* 0.606
+ SciBERT	MSM	0.667	0.754	0.611	88%	**0.724**	* 0.789	0.646
+ SciBERT (SLEDGE-Z)	MedM	**0.681**	**0.800**	**0.663**	90%	0.719	**0.846**	**0.697**
+ ConvKNRM	MSM	0.536	0.617	0.491	86%	0.580	0.645	0.508
+ ConvKNRM	MedM	0.565	0.668	0.525	86%	0.621	0.714	0.565
+ CEDR-KNRM	MSM	0.514	0.617	0.468	86%	0.524	0.628	0.474
+ CEDR-KNRM	MedM	0.619	0.714	0.560	88%	0.649	0.742	0.582
+ Seq2seq T5	MSM	0.656	0.737	0.634	90%	0.685	0.765	0.651
+ Seq2seq T5	MedM	0.626	0.714	0.594	86%	0.678	0.754	0.628
Fusion1	-	0.519	0.640	0.457	94%	0.534	0.640	0.457
Fusion2	-	0.601	0.737	0.565	96%	0.605	0.737	0.565

Table 1: Ablation results and comparison of our approach and other zero-shot baselines on TREC-COVID Rounds 1 and 2. The top results are shown in bold. SciBERT with MedM (SLEDGE-Z) significantly outperforms values in the top (ablation) section marked with * ($p < 0.05$, paired t-test, Bonferroni correction).

MacAvaney et al. SLEDGE-Z: A Zero-Shot Baseline for COVID-19 Literature Search. *EMNLP 2020*

	Team	Run	Type	nDCG@20	P@20	AP
Round 4: 45 topics						
(4a)	unique_ptr	UPrrf38rrf3-r4†	feedback	0.7843	0.8211	0.4681
(4b)	covidex	covidex.r4.duot5.lr (= expando + monoT5 + duoT5 + LR)	feedback	0.7745	0.7967	0.3846
(4c)	covidex	covidex.r4.d2q.duot5 (= expando + monoT5 + duoT5)	automatic	0.7219	0.7267	0.3122
(4d)	covidex	covidex.r4.duot5 (= monoT5 + duoT5)	automatic	0.6877	0.6922	0.3283
(4e)	uogTr	uogTrDPH_QE_SCB1	automatic	0.6820	0.7144	0.3457
(4f)	anserini	r4.fusion2	automatic	0.6089	0.6589	0.3088
(4g)	anserini	r4.fusion1	automatic	0.5244	0.5611	0.2666
Round 5: 50 topics						
(5a)	unique_ptr	UPrrf93-wt-r5†	feedback	0.8496	0.8760	0.4718
(5b)	covidex	covidex.r5.2s.lr (= monoT5 + duoT5 + LR)	feedback	0.8311	0.8460	0.3922
(5c)	covidex	covidex.r5.d2q.2s.lr (= expando + monoT5 + duoT5 + LR)	feedback	0.8304	0.8380	0.3875
(5d)	covidex	covidex.r5.d2q.2s (= expando + monoT5 + duoT5)	automatic	0.7539	0.7700	0.3227
(5e)	covidex	covidex.r5.2s (= monoT5 + duoT5)	automatic	0.7457	0.7610	0.3212
(5f)	uogTr	uogTrDPH_QE_SB_CB	automatic	0.7427	0.7910	0.3305
(5g)	covidex	covidex.r5.d2q.1s (= expando + monoT5)	automatic	0.7121	0.7320	0.3150
(5h)	anserini	r5.fusion2	automatic	0.6007	0.6440	0.2734
(5i)	anserini	r5.fusion1	automatic	0.5313	0.5840	0.2314

Neural Covidex

covidex.ai/?query=What%20is%20the%20prognostic%20value%20of%20IL-6%20levels%20in%20COVID-19?&vertical=cord19

Neural Covidex

Feedback

Search CORD-19 for What is the prognostic value of IL-6 levels in COVID-19? Search

Filter your search

Publication Time

2016 2020

Source

- Elsevier
- MedRxiv
- Medline
- PMC
- WHO

Author

Find a filter...

- A. CANCELLIERE
- A. Charney
- A. Firpo-Betancourt
- A. GIUDICE
- A. Gianatti

More filters...

Journal

Find a filter...

1. SARS-CoV-2 infection: the role of cytokines in COVID-19 disease

Víctor J. Costela-Ruiz, Rebeca Illescas-Montes, Jose M. Puerta-Puerta, Concepción Ruiz, Lucia Melguizo-Rodríguez. *Cytokine Growth Factor Rev* (2020-06-02)

… detected elevated **IL-6 levels** in one-third of patients with mild symptoms and three-quarters of those with severe symptoms, concluding that **IL-6**, alongside **IL**-10, may be of **prognostic value** in patients with **COVID-19** [54]. … Various authors have detected this interleukin in patients with **COVID-19** and related its **levels** to disease severity and progression, as in the case of other cytokines [17,43,57,64,82,83,86,88,89] and it has been reported to have possible **prognostic value** [54]. …

Show more Related articles

2. Rethinking interleukin-6 blockade for treatment of COVID-19

S. Scherger, A. Henao-Martínez, C. Franco-Paredes, L. Shapiro. *Med Hypotheses* (2020-06-27)

… This report shows elevated circulating **IL-6** concentrations in **COVID-19** patients and association between elevated **IL-6 levels** and disease severity. … While elevated **IL-6 levels** may have **prognostic value**, correlation between elevated **IL-6** and severity of illness does not prove causation. …

Show more Related articles

3. The Prevention and Management of COVID-19: Seeking a Practical and Timely Solution

Charulata Jindal, Sandeep Kumar, Sunil Sharma, Yuk Ming Choi, Jimmy T. Efird. *Int J Environ Res Public Health* (2020-06-04)

… D-dimer **values** correlate with disease severity and are a reliable **prognostic** marker for in-hospital mortality in patients with **COVID-19** [32]. … Raised baseline **IL-6 levels** were positively correlated to maximal body temperature during hospitalization and were also associated with greater progression of CT findings [35]. …

Show more Related articles

Vera

Reducing Harmful Misinformation in Consumer Health Search

McGill University

Americans are super-spreaders of COVID-19 misinformation

Misinformation about COVID-19 is spreading from the United States into ... Science at McGill University under the supervision of Dietlind Stolle.

BBC

Covid vaccine: Social media urged to remove 'disinfo dozen'

Covid vaccine: Social media urged to remove 'disinfo dozen' ... Twitter and Google have not responded to the BBC on the specific request to ... dangers of Covid-19 and spread misinformation about the safety of vaccines". ... with

Wired UK

Covid lies are tearing through India's family WhatsApp groups

In a see-saw movement, Covid disinformation and misinformation is ... The practice of Ayurveda is rooted in sacred Hindu texts revered by ...

5 days ago

Can multi-stage ranking systems do this?


NOOOOOO!!!


THE VERY BEST OF
VERA LYNN
We'll Meet Again
75014

TREC 2020 Health Misinformation Ad Hoc Retrieval Task

- Credible and correct information over incorrect information
- CommonCrawl News crawl - January 1st to April 30th 2020
- ~ 65 Million Articles
- Top 1K documents for a set of 46 COVID-19 related topics.

Before we rank: Topics

```
<topic>
<number>0</number>
<title>ibuprofen COVID-19</title>
<description>Can ibuprofen worsen COVID-19?</description>
<answer>no</answer>
<evidence>https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7287029</evidence>
<narrative>Ibuprofen is an anti-inflammatory drug used to reduce fever and
treat pain and inflammation. Recently, there has been a debate over whether
Ibuprofen can worsen the effects of COVID-19. A helpful document might explain
in clear language that there is no scientific evidence supporting this
concern. A harmful document mightcreate anxiety and/or cause people to avoid
taking the drug.</narrative>
</topic>
<topic>
```

A sample topic

- A [description] is more generally "Can X Y COVID-19?" where X is a treatment and Y is one of five action words.
- An [answer] is one of "yes" or "no".

Before we rank: Query Variants

```
<topic>
<number>0</number>
<title>ibuprofen COVID-19</title>
<description>Can ibuprofen worsen COVID-19?</description>
<answer>no</answer>
<evidence>https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7287029</evidence>
<narrative>Ibuprofen is an anti-inflammatory drug used to reduce fever and
treat pain and inflammation. Recently, there has been a debate over whether
Ibuprofen can worsen the effects of COVID-19. A helpful document might explain
in clear language that there is no scientific evidence supporting this
concern. A harmful document mightcreate anxiety and/or cause people to avoid
taking the drug.</narrative>
</topic>
<topic>
```

A sample topic

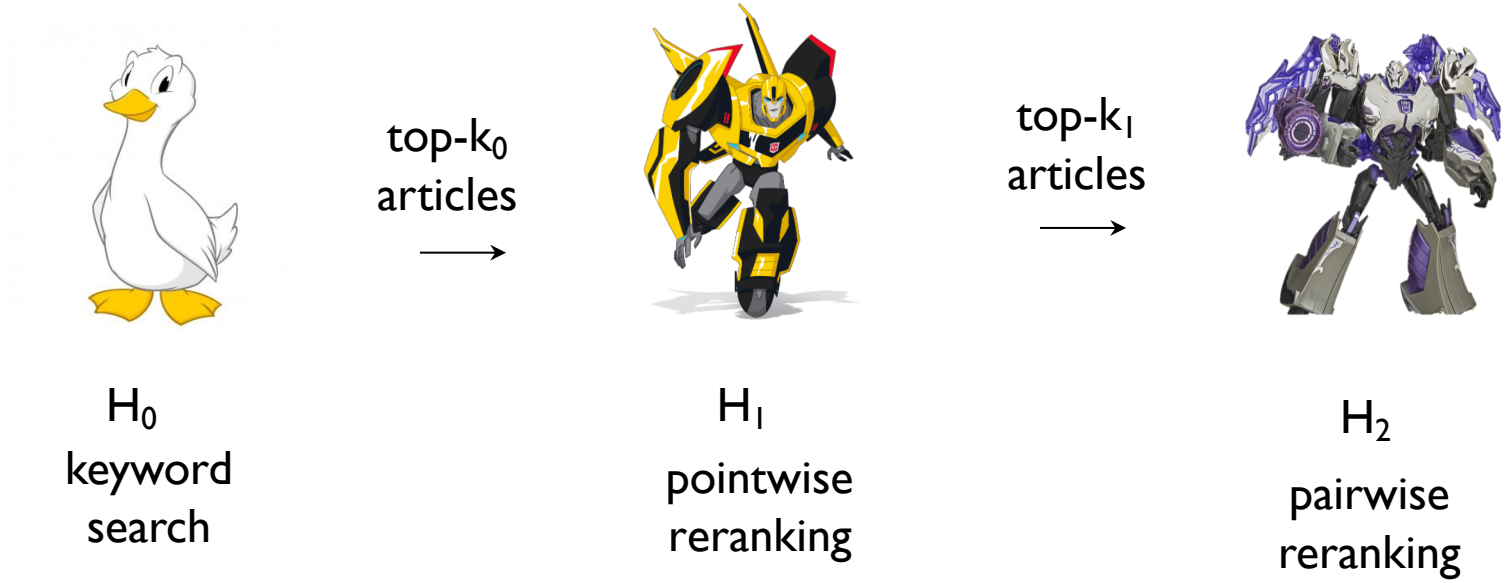
- $\text{Query}_{\text{base}}$: "[description]" - Example: Can ibuprofen worsen COVID-19?
- Query_{NL} : "X can Y COVID-19" if [answer] is "yes" else "X can not Y COVID-19"
  Example: Ibuprofen can not worsen COVID-19

Before we rank: Metrics

Three metrics are considered

- $COMP_{HELP}$
- $COMP_{HARM}$
- $COMP_{\Delta} = COMP_{HELP} - COMP_{HARM}$

First Attempt: Mono-Duo-T5



multi-stage document retrieval

keyword-based retrieval (H_0)
two reranking stages (H_1, H_2)

Pradeep et al. "The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models." arXiv:2101.05667 (2021)

Results 1 – Relevance Ranking

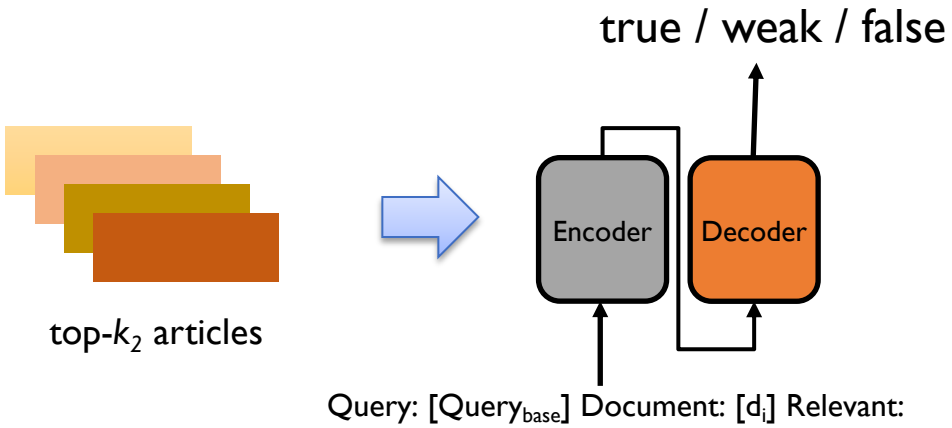
Model	$\text{COMP}_{\text{HELP}}$	$\text{COMP}_{\text{HARM}}$	COMP_{Δ}
(a) Median	0.334	0.075	0.259
(b) BM25	0.368	0.120	0.248
(c) + $\text{monoT5}_{\text{base}}$	0.440	0.113	0.327
(d) + $\text{duoT5}_{\text{base}}$	0.466	0.120	0.346
(e) + $\text{monoT5}_{\text{NL}}$	0.511	0.075	0.436
(f) + duoT5_{NL}	0.549	0.080	0.469

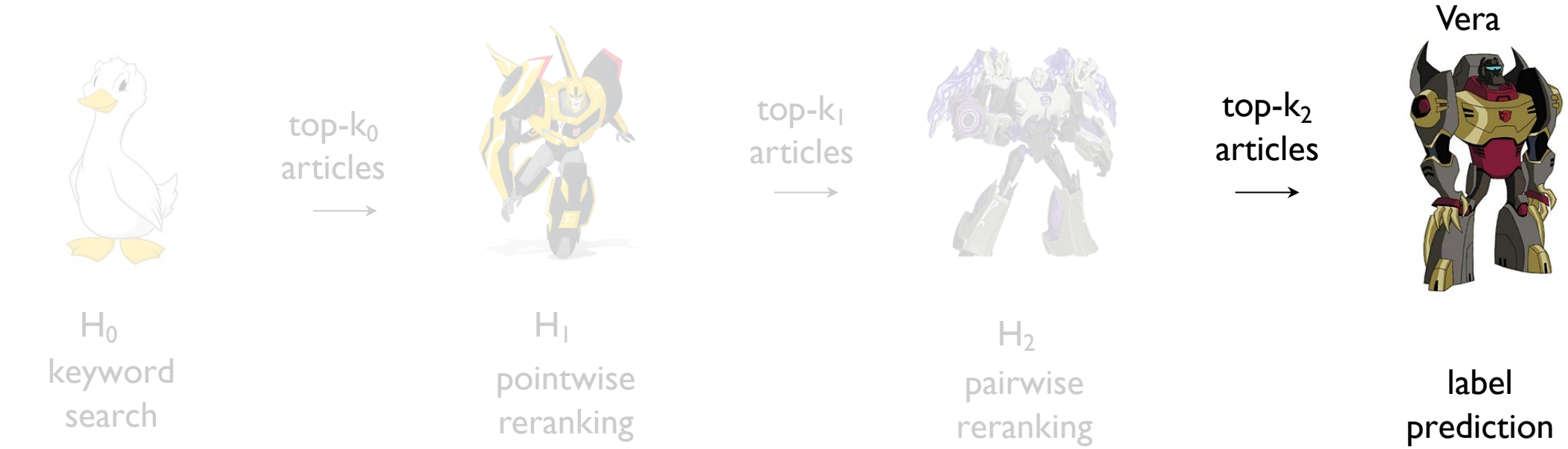
Vera(city) Prediction

FINE-TUNING

Effectiveness Judgements from TREC 2019: Decision (Medical Misinformation) Track.

Effective → true
Ineffective → false
No Info / Not Relevant → weak



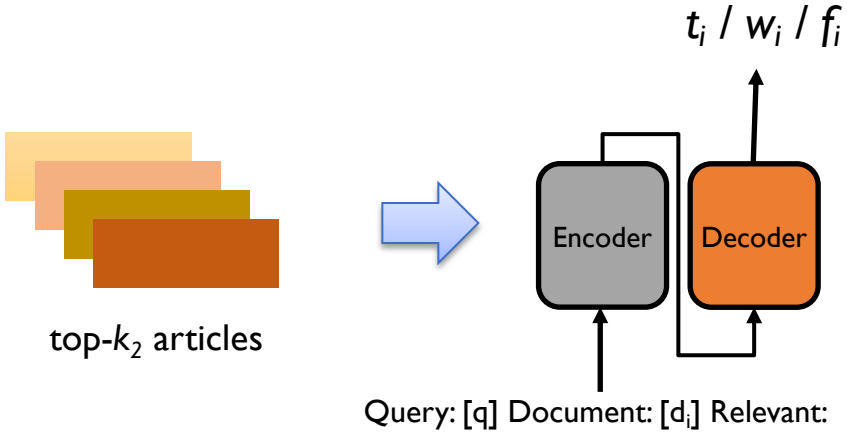


INFERENCE

$$s_i^{\text{final}} = \underbrace{\lambda \cdot s_i^z}_{\textit{Relevancy}} + \underbrace{(1-\lambda) \cdot \begin{cases} t_i - f_i, \text{answer field is "yes"} \\ f_i - t_i, \text{answer field is "no"} \end{cases}}_{\textit{Veracity}}$$

where z is one of mono or duo in this case.

We denote this as $\text{Vera}\ (\lambda, z)$



Results 2 – Do no harm!

Model	$\text{COMP}_{\text{HELP}}$	$\text{COMP}_{\text{HARM}}$	COMP_{Δ}
(a) Median	0.334	0.075	0.259
(b) BM25	0.368	0.120	0.248
(c) + $\text{monoT5}_{\text{base}}$	0.440	0.113	0.327
(d) + $\text{duoT5}_{\text{base}}$	0.466	0.120	0.346
(e) + $\text{monoT5}_{\text{NL}}$	0.511	0.075	0.436
(f) + duoT5_{NL}	0.549	0.080	0.469
(g) Vera ($\lambda = 0.0, z = \text{mono}$)	0.449	0.015	0.434

Results 3 – (Help!) And do no harm!

Model	$\text{COMP}_{\text{HELP}}$	$\text{COMP}_{\text{HARM}}$	COMP_{Δ}
(a) Median	0.334	0.075	0.259
(b) BM25	0.368	0.120	0.248
(c) + $\text{monoT5}_{\text{base}}$	0.440	0.113	0.327
(d) + $\text{duoT5}_{\text{base}}$	0.466	0.120	0.346
(e) + $\text{monoT5}_{\text{NL}}$	0.511	0.075	0.436
(f) + duoT5_{NL}	0.549	0.080	0.469
(g) Vera ($\lambda = 0.0, z = \text{mono}$)	0.449	0.015	0.434
(h) Vera ($\lambda = 0.5, z = \text{mono}$)	0.490	0.016	0.474
(i) Vera ($\lambda = 0.95, z = \text{mono}$)	0.507	0.019	0.488
(j) Vera ($\lambda = 0.95, z = \text{duo}$)	0.520	0.018	0.502

Submitted Runs

Post-hoc Runs

Results 4 – Together We Stand

Model	$\text{COMP}_{\text{HELP}}$	$\text{COMP}_{\text{HARM}}$	COMP_{Δ}
(a) Median	0.334	0.075	0.259
(b) cn-kq-td (Webis)	0.334	0.052	0.282
(c) adhoc_run3 (KU)	0.401	0.121	0.280
(d) Vera ($\lambda = 0.5, z = \text{mono}$)	0.490	0.016	0.474
(e) Vera ($\lambda = 0.95, z = \text{mono}$)	0.507	0.019	0.488
(f) Vera ($\lambda = 0.95, z = \text{duo}$)	0.520	0.018	0.502

Clinical Trials Matching

Collections

Clinical Trials (Koopman et al. 2016)

TREC Clinical Trials 2021

Clinical Trials Collection (Koopman et al. 2016)

Query: A 58-year old African American woman presents to the ER with episodic pressing/burning anterior chest pain that began two days earlier for the first time in her life. The pain started while she was walking, radiates to the back, and is accompanied by nausea …

Relevant Passage: ID: *NCT00005485* Title: Environmental and Genetic Factors That Influence Cardiovascular Disease in African Americans

Non-Relevant Passage: ID: *NCT00000408* Title: Low Back Pain Patient Education Evaluation

- Graded-Relevance: Scale – 0 - 2
- Dataset size? 0: 2764, 1: 685, 2: 421, only ~4000 labelled examples ☹.

TREC Clinical Trials 2021

- Graded-Relevance: Scale – 0 - 2
- Corpus: 380k clinical trials.
- Fields: Long – Description & Eligibility; Short – Title, Condition & Summary.
- Topics – 75.

Trial: NCT03863457

Inclusion Criteria:

- Participants will be $\geq$ 18 years of age
- Known or suspected primary or metastatic breast cancer.
- At least one lesion $\geq$ 1.5 cm that is seen on standard imaging (e.g. CT, MRI, mammogram, ultrasound, FDG-PET/CT). Only one type of imaging is required to show a lesion.
- Participants must be informed of the investigational nature of this study and be willing to provide written informed consent and participate in this study in accordance with institutional and federal guidelines prior to study-specific procedures.

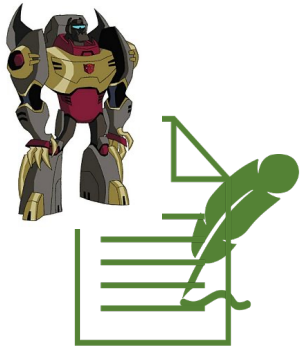
Exclusion Criteria:

- Females who are pregnant or breast feeding at the time of screening; a urine pregnancy test will be performed in women of child-bearing potential at screening.
- Inability to tolerate imaging procedures in the opinion of an investigator or treating physician.
- Any current medical condition, illness, or disorder, as assessed by medical record review and/or self-reported, that is considered by a physician investigator to be a condition that could compromise participant safety or successful participation in the study.

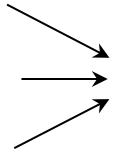
Topic 3	A 32 yo woman who presents following a severe 'exploding' headache. She and her husband report that yesterday she was in the kitchen and stood up and hit her head on the corner of a cabinet. The next morning she developed a sudden 'exploding' headache. She came to the hospital where head CT showed a significant amount of blood in her right ventricle. NSGY evaluated her for spontaneous intraventricular hemorrhage with a concern for an underlying vascular malformation. Cerebral angiogram was done which showed abnormal vasculature with a draining vein from L temporal lobe penetrating deep white matter consistent with AVM. The patient did continue to have a headaches but they were improving with pain medication. The patient refused PT evaluation but was ambulating independently without difficulty. She was discharged to home with her husband on [**2155-12-6**].

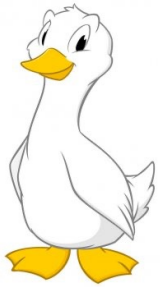
Neural Query Synthesis (NQS) using doc2query-T5

Expando^{kinda}-Mono^{2}-~~Duo~~-T5



H_{-1}
question
synthesis





H_0
keyword
search

top-k_0
passages
→

H_1
pointwise
reranker

top-k_1
passages
→

H_1'
pointwise
reranker
prime

multi-stage ranking architecture

keyword-based 1st stage retrieval (H_0)

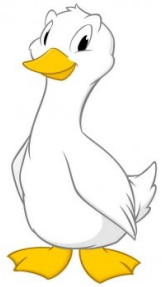
two reranking stages using pointwise reranker (H_1 & H_1')

question synthesis using doc2query-T5 (H_{-1})

Work Under Review

1st stage retrieval

H_{-1}
question
decomposition



H_0
keyword
search

top-k_0
passages

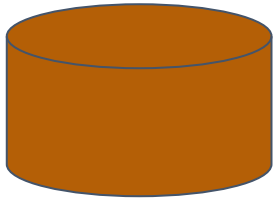
H_1
pointwise
reranker

top-k_1
passages

H_1'
pointwise
reranker
prime

Pyserini

Patient
Description

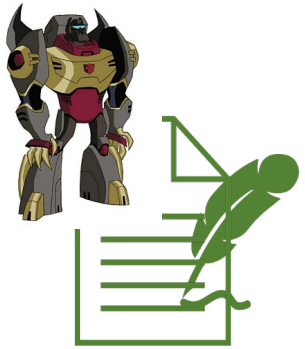


top-k_0 texts

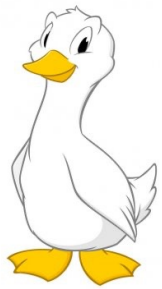
doc2query

Expando-Mono-Duo-T5

What queries are relevant for this document?



H_{-1}
document
expansion



H_0
keyword
search

H_1
pointwise
reranker

H_2
pairwise
reranker

what is the weather in washington dc?
when is the hottest month in washington dc?
…

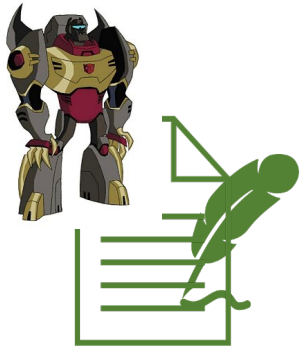
Original Corpus

Expanded Corpus

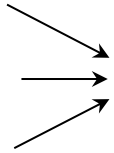
Passed onto Anserini

July is the hottest month in Washington DC with an average temperature of 27C (80F) and the coldest is January at 4C (38F) with the most daily sunshine hours at 9 in July. The wettest month is May with an average of 100mm of rain.

Query Synthesis using doc2query-T5



H_{-1}
question
synthesis



H_0
keyword
search

top-k_0
passages

H_1
pointwise
reranker

top-k_1
passages

H_1'
pointwise
reranker
prime

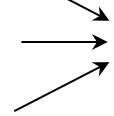
The patient is a 55-year-old man who was recently diagnosed with Parkinson's disease. He is complaining of slowness of movement and tremors. His disease is ranked as mild, Hoehn-Yahr Stage I. His past medical history is significant for hypertension and hypercholesterolemia. He lives with his wife. They have three children. He used to be active with gardening before his diagnosis. He complains of shaking and slow movement. He had difficulty entering through a door, as he was frozen and needed guidance to step in. His handwriting is getting smaller. He is offered Levodopa and Trihexyphenidyl. He is an alert and cooperative man who does not have any signs of dementia. He does not smoke or use any illicit drugs.

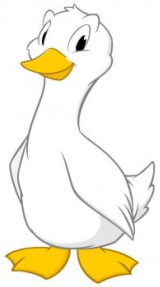
[q_1] what is hoehn yahr stage
[q_2] what is the order of increasing a patient's anticholinergic medication during a parkinson's disease episode
…

Passed onto Pyserini

1st stage retrieval (revisited)

H_{-1}
question synthesis





H_0
keyword search

top-k_0 passages

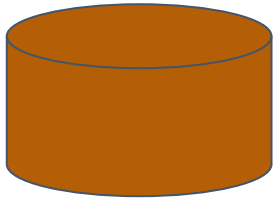
H_1
pointwise reranker

top-k_1 passages

H_1'
pointwise reranker prime

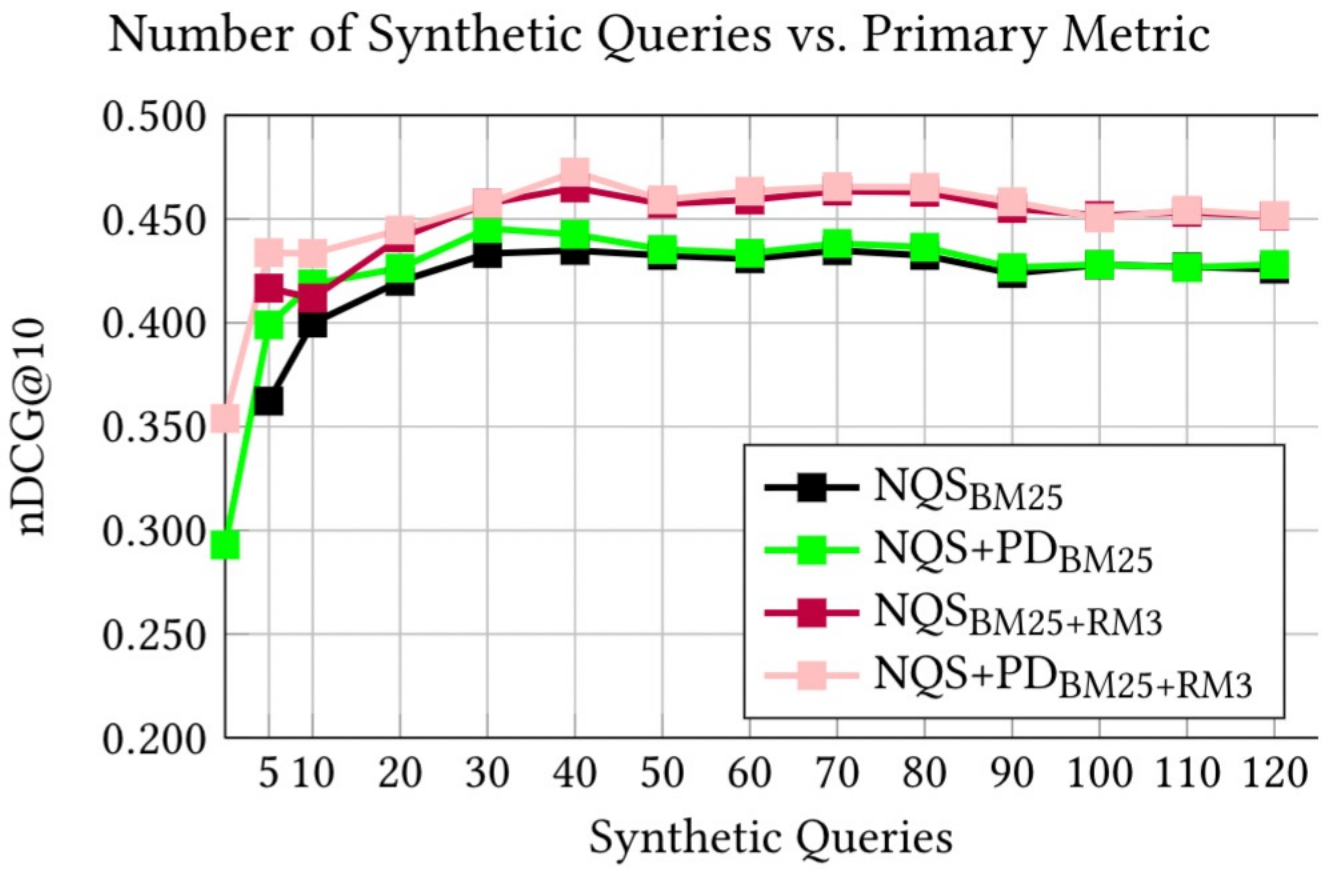
Topic Description (PD)
q_1
q_2
...
q_j

Pyserini



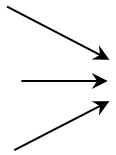
top-k_0 texts

Results 1– First Stage



H_1 - monoT5$_{MED}$

H_{-1}
question synthesis



H_0
keyword search

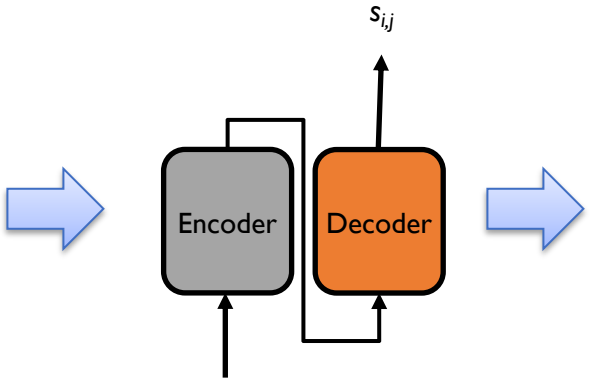
top-k_0 passages

H_1
pointwise reranker

top-k_1 passages

H_1'
pointwise reranker prime

Segments of top-k_0 texts



Query: [query] Document: [$p_{i,j}$] Relevant:

top-k_1 texts

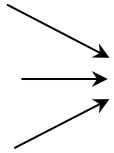
$$e_i' = \max_{\text{eligibility segments}} s_{i,j}$$
$$d_i' = \max_{\text{description segments}} s_{i,j}$$
$$a_i' = \max_{\text{all segments}} s_{i,j}$$

Eligibility or Description

[$p_{i,j}$] = title: [t_{title}] condition: [$t_{condition}$] eligibility: [$t_{eligibility,j}$] or [$p_{i,j}$] = title: [t_{title}] condition: [$t_{condition}$] description: [$t_{description,j}$]

monoT5$_{CT}$

H_{-1}
question
decomposition



H_0
keyword
search

top-k_0
passages
→

H_1
pointwise
reranker

top-k_1
passages
→

H_1'
pointwise
reranker
prime

FINE-TUNING

Judgements from Koopman et al. 2016.

Graded Relevance

$\geq 1 \rightarrow$ true

$< 1 \rightarrow$ false

Three input templates for training

$[p_{i,j}]$ = title: $[t_{title}]$ condition: $[t_{condition}]$ eligibility: $[t_{eligibility,j}]$

$[p_{i,j}]$ = title: $[t_{title}]$ condition: $[t_{condition}]$ description: $[t_{description,j}]$

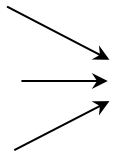
$[p_{i,j}]$ = title: $[t_{title}]$ condition: $[t_{condition}]$ eligibility: $[t_{eligibility,j}]$ description: $[t_{description,j}]$

Results 2 – Pointwise rerankers

	Run	nDCG@10	P@10	RR
(a)	Median	0.3040	0.1613	0.2942
(b)	damoebrtog	0.5953	0.4093	0.6083
(c)	CSIROmed_inc	0.5320	0.3173	-
(d)	BM25	0.2923	0.1680	0.3015
(e)	BM25+RM3	0.3539	0.2040	0.3659
(f)	NQS+PD$_{\text{BM25+RM3}}$	0.4726	0.2760	0.4304
(g)	+ monoT5$^{\text{A}}_{\text{MED}}$	0.2994	0.1973	0.3560
(h)	+ monoT5$^{\text{D}}_{\text{MED}}$	0.2311	0.1507	0.3223
(i)	+ monoT5$^{\text{E}}_{\text{MED}}$	0.4715	0.2987	0.4830
(j)	+ monoT5$^{\text{A}}_{\text{CT}}$	0.6763	0.5480	0.7253
(k)	+ monoT5$^{\text{D}}_{\text{CT}}$	0.4493	0.3267	0.6260
(l)	+ monoT5$^{\text{E}}_{\text{CT}}$	0.6792	0.5493	0.7161
(m)	+ monoT5'$_{\text{CT}}$	0.7118	0.5933	0.8162

H_1' - monoT5$_{CT}$'

H_{-1}
question
decomposition



H_0
keyword
search

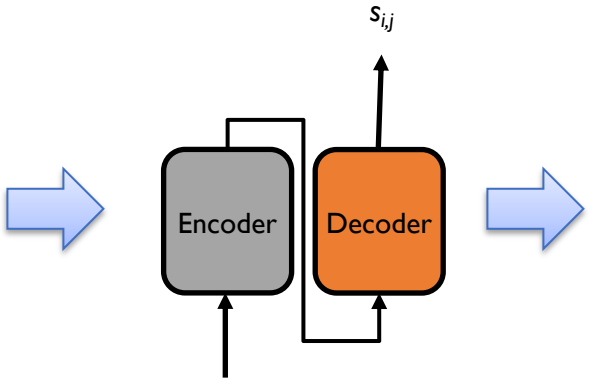
top-k_0
passages

H_1
pointwise
reranker

top-k_1
passages

H_1'
pointwise
reranker
prime

Segments of
top-k_1 texts



top-k_2 texts

[$p_{i,j}$] = title: [t_{title}] condition: [$t_{condition}$] eligibility: [$t_{eligibility,best}$] description: [$t_{description,best}$]

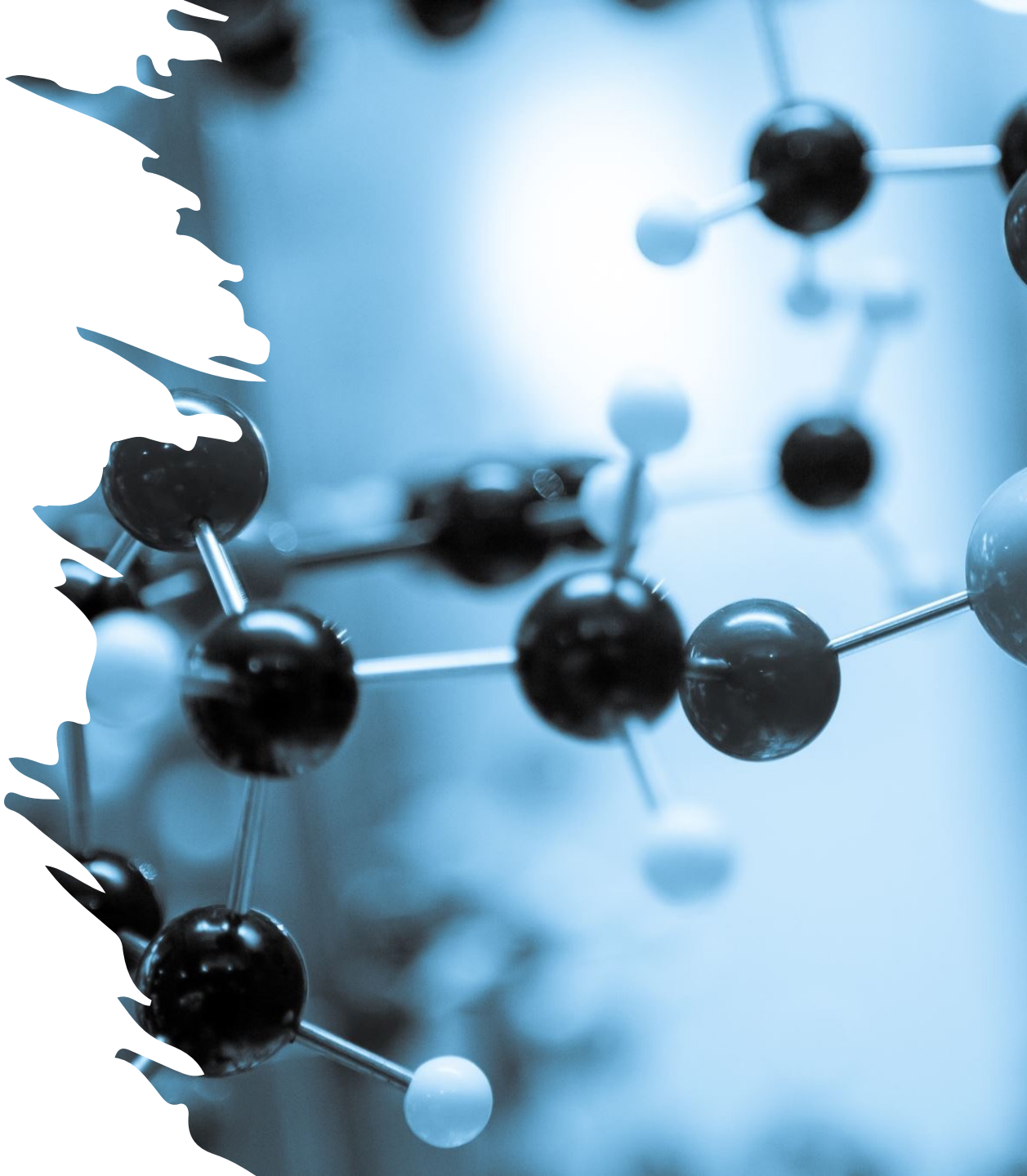
Results 3 – State of the Art!

	Run	nDCG@10	P@10	RR
(a)	Median	0.3040	0.1613	0.2942
(b)	`damoebrtog`	0.5953	0.4093	0.6083
(c)	`CSIROmed_inc`	0.5320	0.3173	-
(d)	BM25	0.2923	0.1680	0.3015
(e)	BM25+RM3	0.3539	0.2040	0.3659
(f)	$\text{NQS+PD}_{\text{BM25+RM3}}$	0.4726	0.2760	0.4304
(g)	$+ \text{monoT5}^{\text{A}}_{\text{MED}}$	0.2994	0.1973	0.3560
(h)	$+ \text{monoT5}^{\text{D}}_{\text{MED}}$	0.2311	0.1507	0.3223
(i)	$+ \text{monoT5}^{\text{E}}_{\text{MED}}$	0.4715	0.2987	0.4830
(j)	$+ \text{monoT5}^{\text{A}}_{\text{CT}}$	0.6763	0.5480	0.7253
(k)	$+ \text{monoT5}^{\text{D}}_{\text{CT}}$	0.4493	0.3267	0.6260
(l)	$+ \text{monoT5}^{\text{E}}_{\text{CT}}$	0.6792	0.5493	0.7161
(m)	$+ \text{monoT5'}_{\text{CT}}$	0.7118	0.5933	0.8162

Drawbacks


Not the friendliest Muppet in Sesame Street!

Lack of Biomedical Knowledge



	Team	Run	Type	nDCG@20	P@20	AP
Round 4: 45 topics						
(4a)	unique_ptr	UPrrf38rrf3-r4†	feedback	0.7843	0.8211	0.4681
(4b)	covidex	covidex.r4.duot5.lr (= expando + monoT5 + duoT5 + LR)	feedback	0.7745	0.7967	0.3846
(4c)	covidex	covidex.r4.d2q.duot5 (= expando + monoT5 + duoT5)	automatic	0.7219	0.7267	0.3122
(4d)	covidex	covidex.r4.duot5 (= monoT5 + duoT5)	automatic	0.6877	0.6922	0.3283
(4e)	uogTr	uogTrDPH_QE_SCB1	automatic	0.6820	0.7144	0.3457
(4f)	anserini	r4.fusion2	automatic	0.6089	0.6589	0.3088
(4g)	anserini	r4.fusion1	automatic	0.5244	0.5611	0.2666
Round 5: 50 topics						
(5a)	unique_ptr	UPrrf93-wt-r5†	feedback	0.8496	0.8760	0.4718
(5b)	covidex	covidex.r5.2s.lr (= monoT5 + duoT5 + LR)	feedback	0.8311	0.8460	0.3922
(5c)	covidex	covidex.r5.d2q.2s.lr (= expando + monoT5 + duoT5 + LR)	feedback	0.8304	0.8380	0.3875
(5d)	covidex	covidex.r5.d2q.2s (= expando + monoT5 + duoT5)	automatic	0.7539	0.7700	0.3227
(5e)	covidex	covidex.r5.2s (= monoT5 + duoT5)	automatic	0.7457	0.7610	0.3212
(5f)	uogTr	uogTrDPH_QE_SB_CB	automatic	0.7427	0.7910	0.3305
(5g)	covidex	covidex.r5.d2q.1s (= expando + monoT5)	automatic	0.7121	0.7320	0.3150
(5h)	anserini	r5.fusion2	automatic	0.6007	0.6440	0.2734
(5i)	anserini	r5.fusion1	automatic	0.5313	0.5840	0.2314

Conclusions

- System - zero-shot & few-shot capabilities.
- Wins everywhere – doc2query, pointwise reranking, pairwise reranking, finetuning and query synthesis.
- Medical subset MS MARCO => Better Biomedical Search.
- Linear Combinations => models less likely bring up harmful misinformation
- Better Templates => rankers learn and perform better on Clinical Trial Matching.