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

`black bear attacks`

$collection$

metric: 0.66

1.
2.
3.
Transformers

Text Embeddings

E_{[CLS]}  E_1  E_2  E_3  ...  E_{n-2}  E_{n-1}  E_n  E_{[SEP]}

[CLS]  A_1  A_2  A_3  ...  A_{n-2}  A_{n-1}  A_n  [SEP]
Pretrained Transformers

Initialized via pretraining
T5 (2020 ➡️ ?)
Welcome T5

Pretraining Dataset
C4 (Colossal Clean Crawled Corpus)

Multitask Pretraining Mixture

"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi..."

IR Background

Unsupervised ranking methods
Unsupervised Ranking Methods

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

term score
input
Unsupervised Ranking Methods

Inverse Document Frequency

Term Frequency

more discriminative

the and but ...

HIV AIDS SARS ...

12
Unsupervised Ranking Methods – TF-IDF

\[ S(q, d) = \sum_{t \in q \cap d} \text{TF}(t) \times \text{IDF}(t) \]

Sparse representations
Unsupervised Ranking Methods – BM25

BM25(q, d) = \sum_{t \in q \cap d} \log \frac{N - df(t) + 0.5}{df(t) + 0.5} \cdot \frac{tf(t, d) \cdot (k_1 + 1)}{tf(t, d) + k_1 \cdot (1 - b + b \cdot \frac{L_d}{L})
Vocabulary Mismatch

- 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
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

H₁ document expansion

H₀ keyword search

top-\(k₀\) passages

H₁ pointwise reranker

top-\(k₁\) passages

H₂ pairwise reranker

multi-stage ranking architecture

keyword-based 1ˢᵗ stage retrieval (\(H₀\))
two reranking stages (\(H₁\) and \(H₂\))

pre-indexing document enrichment

July is the hottest month in Washington DC with an average temperature of 27°C (80°F) and the coldest is January at 4°C (39°F) with the most daily sunshine hours at 9 in July. The wettest month is May with an average of 100mm of rain.
July is the hottest month in Washington DC with an average temperature of 27°C (80°F) and the coldest is January at 4°C (39°F) with the most daily sunshine hours at 9 in July. The wettest month is May with an average of 100mm of rain.

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

...
Expando-Mono-Duo-T5

H_{i1}  
document expansion

\rightarrow

H_0  
keyword search

\rightarrow

H_1  
pointwise reranker

\rightarrow

H_2  
pairwise reranker

Anserini / Pyserini

(keyword) query

(top-k_0 texts)
Is this document relevant?

H₁: document expansion
H₀: keyword search
H₁: pointwise reranker
H₂: pairwise reranker

Query: [query] Document: [p] Relevant:

Encoder
Decoder

top-k₀ texts

true/false

sᵢ

top-k₁ texts

## Results – MS MARCO Passage Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Expand-Mono-Duo Variants</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_{-1}$</td>
<td>$H_0$</td>
<td>$H_1$</td>
</tr>
<tr>
<td>(1) -</td>
<td>-</td>
<td>BM25</td>
<td>-</td>
</tr>
<tr>
<td>(2) -</td>
<td>-</td>
<td>BM25</td>
<td>monoT5-base</td>
</tr>
<tr>
<td>(3) -</td>
<td>-</td>
<td>BM25</td>
<td>monoT5-large</td>
</tr>
<tr>
<td>(4) -</td>
<td>-</td>
<td>BM25</td>
<td>monoT5-3B</td>
</tr>
<tr>
<td>(5) doc2query-T5</td>
<td>-</td>
<td>BM25</td>
<td>-</td>
</tr>
<tr>
<td>(6) doc2query-T5</td>
<td>-</td>
<td>BM25</td>
<td>monoT5-3B</td>
</tr>
<tr>
<td>(7) doc2query-T5</td>
<td>-</td>
<td>BM25</td>
<td>monoT5-3B</td>
</tr>
<tr>
<td>(8) -</td>
<td>-</td>
<td>BM25</td>
<td>monoT5-base</td>
</tr>
<tr>
<td>(9) doc2query-T5</td>
<td>-</td>
<td>BM25</td>
<td>monoT5-3B</td>
</tr>
<tr>
<td>(10) BM25 + BERT-large [40]</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(11) TFR-BERT Ensemble [15]</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SOTA!
Towards Longer Texts
Passages ➡ Documents
Documents: too long for these poor Transformers!

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 $n_{\text{length}}$ sentences and a stride of $n_{\text{stride}}$

**Document:** $s_1 s_2 s_3 s_4 s_5 s_6 s_7 s_8$

$H_1$ document expansion

$H_0$ keyword search

$H_1$ pointwise reranker

$H_2$ pairwise reranker

Segment

$n_{\text{length}} = 4$ with $n_{\text{stride}} = 2$

Segments:

$e_1 = s_1 s_2 s_3 s_4$

$e_2 = s_3 s_4 s_5 s_6$

$e_3 = s_5 s_6 s_7 s_8$

where $s_i$ is sentence $i$. 
Expando-Mono-Duo-T5

Is this document relevant?

H₁ document expansion

H₀ keyword search

H₁ pointwise reranker

H₂ pairwise reranker

Segments of top-k₀ texts

Query: [query] Document: [pᵢ,j] Relevant:

true/false

sᵢ = max sᵢ,j

eᵢ: highest scoring segment in the text

top-k₁ texts

Expando-Mono-Duo-T5

Is this document more relevant than that document?

$H_0$ keyword search

$H_1$ pointwise reranker

$H_2$ pairwise reranker

Biomedical Search
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.
<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>How does the coronavirus respond to changes in the weather?</td>
</tr>
<tr>
<td>Has social distancing had an impact on slowing the spread of COVID-19?</td>
</tr>
<tr>
<td>How long can the coronavirus live outside the body?</td>
</tr>
<tr>
<td>What is known about those infected with Covid-19 but are asymptomatic?</td>
</tr>
<tr>
<td>What evidence is there for the value of hydroxychloroquine in treating Covid-19?</td>
</tr>
</tbody>
</table>
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.
## SLEDGE

<table>
<thead>
<tr>
<th>Model</th>
<th>Training</th>
<th>Including Unjudged</th>
<th>Judged Only</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>nDCG@10</td>
<td>P@5</td>
</tr>
<tr>
<td>BM25</td>
<td>-</td>
<td>* 0.368</td>
<td>* 0.469</td>
</tr>
<tr>
<td>+ BERT</td>
<td>MSM</td>
<td>* 0.547</td>
<td>* 0.617</td>
</tr>
<tr>
<td>+ BERT</td>
<td>MedM</td>
<td>0.625</td>
<td>* 0.697</td>
</tr>
<tr>
<td>+ SciBERT</td>
<td>MSM</td>
<td>0.667</td>
<td>0.754</td>
</tr>
<tr>
<td>+ SciBERT (SLEDGE-Z)</td>
<td>MedM</td>
<td><strong>0.681</strong></td>
<td><strong>0.800</strong></td>
</tr>
</tbody>
</table>

Table 1: Comparison results and comparison of our approach and other baselines on the TREC-2020 COVID Rounds 1 and 2. The top results are shown in bold. The differences with MedM for SLEDGE-Z, significantly improves the F score in the main evaluation section marked with * in the final column. Friedman correction.
<table>
<thead>
<tr>
<th>Round 4: 45 topics</th>
<th>Team</th>
<th>Run</th>
<th>Type</th>
<th>nDCG@20</th>
<th>P@20</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>(4a) unique.ptr</td>
<td>UPrrf38rrf3-r4†</td>
<td>feedback</td>
<td>0.7843</td>
<td>0.8211</td>
<td>0.4681</td>
<td></td>
</tr>
<tr>
<td>(4b) covidex</td>
<td>covidex.r4.duot5.1r</td>
<td>feedback</td>
<td>0.7745</td>
<td>0.7967</td>
<td>0.3846</td>
<td></td>
</tr>
<tr>
<td>(4c) covidex</td>
<td>covidex.r4.d2q.duot5 (= expando + monoT5 + duoT5 + LR)</td>
<td>automatic</td>
<td>0.7219</td>
<td>0.7267</td>
<td>0.3122</td>
<td></td>
</tr>
<tr>
<td>(4d) covidex</td>
<td>covidex.r4.duot5 (= monoT5 + duoT5)</td>
<td>automatic</td>
<td>0.6877</td>
<td>0.6922</td>
<td>0.3283</td>
<td></td>
</tr>
<tr>
<td>(4e) uogTr</td>
<td>uogTrDHF.QE_SCB1</td>
<td>automatic</td>
<td>0.6820</td>
<td>0.7144</td>
<td>0.3457</td>
<td></td>
</tr>
<tr>
<td>(4f) anserini</td>
<td>r4.fusion2</td>
<td>automatic</td>
<td>0.6089</td>
<td>0.6589</td>
<td>0.3088</td>
<td></td>
</tr>
<tr>
<td>(4g) anserini</td>
<td>r4.fusion1</td>
<td>automatic</td>
<td>0.5244</td>
<td>0.5611</td>
<td>0.2666</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Round 5: 50 topics</th>
<th>Team</th>
<th>Run</th>
<th>Type</th>
<th>nDCG@20</th>
<th>P@20</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5a) unique.ptr</td>
<td>UPrrf93-wt-r5†</td>
<td>feedback</td>
<td>0.8496</td>
<td>0.8760</td>
<td>0.4718</td>
<td></td>
</tr>
<tr>
<td>(5b) covidex</td>
<td>covidex.r5.2s.1r (= monoT5 + duoT5 + LR)</td>
<td>feedback</td>
<td>0.8311</td>
<td>0.8460</td>
<td>0.3922</td>
<td></td>
</tr>
<tr>
<td>(5c) covidex</td>
<td>covidex.r5.d2q.2s.1r (= expando + monoT5 + duoT5 + LR)</td>
<td>feedback</td>
<td>0.8304</td>
<td>0.8380</td>
<td>0.3875</td>
<td></td>
</tr>
<tr>
<td>(5d) covidex</td>
<td>covidex.r5.d2q.2s (= expando + monoT5 + duoT5)</td>
<td>automatic</td>
<td>0.7539</td>
<td>0.7700</td>
<td>0.3227</td>
<td></td>
</tr>
<tr>
<td>(5e) covidex</td>
<td>covidex.r5.2s (= monoT5 + duoT5)</td>
<td>automatic</td>
<td>0.7457</td>
<td>0.7610</td>
<td>0.3212</td>
<td></td>
</tr>
<tr>
<td>(5f) uogTr</td>
<td>uogTrDHF.QE_SB_CB</td>
<td>automatic</td>
<td>0.7427</td>
<td>0.7910</td>
<td>0.3305</td>
<td></td>
</tr>
<tr>
<td>(5g) covidex</td>
<td>covidex.r5.d2q.1s (= expando + monoT5)</td>
<td>automatic</td>
<td>0.7121</td>
<td>0.7320</td>
<td>0.3150</td>
<td></td>
</tr>
<tr>
<td>(5h) anserini</td>
<td>r5.fusion2</td>
<td>automatic</td>
<td>0.6007</td>
<td>0.6440</td>
<td>0.2734</td>
<td></td>
</tr>
<tr>
<td>(5i) anserini</td>
<td>r5.fusion1</td>
<td>automatic</td>
<td>0.5313</td>
<td>0.5840</td>
<td>0.2314</td>
<td></td>
</tr>
</tbody>
</table>
1. SARS-CoV-2 infection: the role of cytokines in COVID-19 disease

Víctor J. Costela-Ruiz, Rebeca Blescas-Montes, José M. Puerta-Puerta, Concepción Ruiz, Lucía 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 [17,43,57,84,82,83,96,93,99] and it has been reported to have possible prognostic value [54].

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


... 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.

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


... 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 [15].
Vera

Reducing Harmful Misinformation in Consumer Health Search
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

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

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 reverred by...
5 days ago
Can multi-stage ranking systems do this?
THE VERY BEST OF

VERA LYNN

We'll Meet Again
TREC 2020 Health Misinformation Ad Hoc Retrieval Task

- Credible and correct information over incorrect information
- CommonCrawl News crawl - January 1\textsuperscript{st} to April 30\textsuperscript{th} 2020
- \~65 Million Articles
- Top 1K documents for a set of 46 COVID-19 related topics.
Before we rank: Topics

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 might create anxiety and/or cause people to avoid taking the drug.</narrative>
</topic>
<topic>

A sample topic

- Query$_{base}$: “[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

- \( \text{COMP}_{\text{HELP}} \)
- \( \text{COMP}_{\text{HARM}} \)
- \( \text{COMP}_\Delta = \text{COMP}_{\text{HELP}} - \text{COMP}_{\text{HARM}} \)
First Attempt: Mono-Duo-T5

multi-stage document retrieval

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

## Results 1 – Relevance Ranking

<table>
<thead>
<tr>
<th>Model</th>
<th>COMP\textsubscript{HELP}</th>
<th>COMP\textsubscript{HARM}</th>
<th>COMP\textsubscript{Δ}</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Median</td>
<td>0.334</td>
<td>0.075</td>
<td>0.259</td>
</tr>
<tr>
<td>(b) BM25</td>
<td>0.368</td>
<td>0.120</td>
<td>0.248</td>
</tr>
<tr>
<td>(c) + monoT5\textsubscript{base}</td>
<td>0.440</td>
<td>0.113</td>
<td>0.327</td>
</tr>
<tr>
<td>(d) + duoT5\textsubscript{base}</td>
<td>0.466</td>
<td>0.120</td>
<td>0.346</td>
</tr>
<tr>
<td>(e) + monoT5\textsubscript{NL}</td>
<td>0.511</td>
<td>0.075</td>
<td>0.436</td>
</tr>
<tr>
<td>(f) + duoT5\textsubscript{NL}</td>
<td>0.549</td>
<td>0.080</td>
<td>0.469</td>
</tr>
</tbody>
</table>
Vera(city) Prediction

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

Effective → true
Ineffective → false
No Info / Not Relevant → weak
We denote this as \( \text{Vera}(\lambda, z) \)

where \( z \) is one of mono or duo in this case.
### Results 2 – Do no harm!

<table>
<thead>
<tr>
<th>Model</th>
<th>( \text{COMP}_{\text{HELP}} )</th>
<th>( \text{COMP}_{\text{HARM}} )</th>
<th>( \text{COMP}_{\Delta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Median</td>
<td>0.334</td>
<td>0.075</td>
<td>0.259</td>
</tr>
<tr>
<td>(b) BM25</td>
<td>0.368</td>
<td>0.120</td>
<td>0.248</td>
</tr>
<tr>
<td>(c) + monoT5\text{base}</td>
<td>0.440</td>
<td>0.113</td>
<td>0.327</td>
</tr>
<tr>
<td>(d) + duoT5\text{base}</td>
<td>0.466</td>
<td>0.120</td>
<td>0.346</td>
</tr>
<tr>
<td>(e) + monoT5\text{NL}</td>
<td>0.511</td>
<td>0.075</td>
<td>0.436</td>
</tr>
<tr>
<td>(f) + duoT5\text{NL}</td>
<td>0.549</td>
<td>0.080</td>
<td>0.469</td>
</tr>
<tr>
<td>(g) Vera (( \lambda = 0.0, z = \text{mono} ))</td>
<td>0.449</td>
<td>0.015</td>
<td>0.434</td>
</tr>
</tbody>
</table>
Results 3 – (Help!) And do no harm!

<table>
<thead>
<tr>
<th>Model</th>
<th>COMP$_{\text{HELP}}$</th>
<th>COMP$_{\text{HARM}}$</th>
<th>COMP$_{\Delta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Median</td>
<td>0.334</td>
<td>0.075</td>
<td>0.259</td>
</tr>
<tr>
<td>(b) BM25</td>
<td>0.368</td>
<td>0.120</td>
<td>0.248</td>
</tr>
<tr>
<td>(c) + monoT5$_{\text{base}}$</td>
<td>0.440</td>
<td>0.113</td>
<td>0.327</td>
</tr>
<tr>
<td>(d) + duoT5$_{\text{base}}$</td>
<td>0.466</td>
<td>0.120</td>
<td>0.346</td>
</tr>
<tr>
<td>(e) + monoT5$_{\text{NL}}$</td>
<td>0.511</td>
<td>0.075</td>
<td>0.436</td>
</tr>
<tr>
<td>(f) + duoT5$_{\text{NL}}$</td>
<td>0.549</td>
<td>0.080</td>
<td>0.469</td>
</tr>
<tr>
<td>(g) Vera ($\lambda = 0.0, z = \text{mono}$)</td>
<td>0.449</td>
<td>0.015</td>
<td>0.434</td>
</tr>
<tr>
<td>(h) Vera ($\lambda = 0.5, z = \text{mono}$)</td>
<td>0.490</td>
<td>0.016</td>
<td>0.474</td>
</tr>
<tr>
<td>(i) Vera ($\lambda = 0.95, z = \text{mono}$)</td>
<td>0.507</td>
<td>0.019</td>
<td>0.488</td>
</tr>
<tr>
<td>(j) Vera ($\lambda = 0.95, z = \text{duo}$)</td>
<td>0.520</td>
<td>0.018</td>
<td>0.502</td>
</tr>
</tbody>
</table>
### Results 4 – Together We Stand

<table>
<thead>
<tr>
<th>Model</th>
<th>COMP$_{HELP}$</th>
<th>COMP$_{HARM}$</th>
<th>COMP$_{Δ}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Median</td>
<td>0.334</td>
<td>0.075</td>
<td>0.259</td>
</tr>
<tr>
<td>(b) cn-kq-td (Webis)</td>
<td>0.334</td>
<td>0.052</td>
<td>0.282</td>
</tr>
<tr>
<td>(c) adhoc_run3 (KU)</td>
<td>0.401</td>
<td>0.121</td>
<td>0.280</td>
</tr>
<tr>
<td>(d) Vera ($\lambda = 0.5$, $z =$ mono)</td>
<td>0.490</td>
<td>0.016</td>
<td>0.474</td>
</tr>
<tr>
<td>(e) Vera ($\lambda = 0.95$, $z =$ mono)</td>
<td>0.507</td>
<td>0.019</td>
<td>0.488</td>
</tr>
<tr>
<td>(f) Vera ($\lambda = 0.95$, $z =$ duo)</td>
<td>0.520</td>
<td>0.018</td>
<td>0.502</td>
</tr>
</tbody>
</table>
Clinical Trials Matching
Collections

Clinical Trials (Koopman et al. 2016)

TREC Clinical Trials 2021
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.
**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
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

Pyserini

Patient Description

H₀ keyword search

top-\(k₀\) passages

H₁ pointwise reranker

top-\(k₁\) passages

H₁' pointwise reranker prime

H₁ question decomposition

top-\(k₀\) texts
July is the hottest month in Washington DC with an average temperature of 27°C (80°F) and the coldest is January at 4°C (39°F) with the most daily sunshine hours at 9 in July. The wettest month is May with an average of 100mm of rain.
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.
1st stage retrieval (revisited)

H₁
question
synthesis

H₀
keyword
search

H₁
pointwise
reranker

Pyserini

RRF

Topic Description
(PD)
q₁
q₂
...
qᵢ

top-\(k₀\) passages

top-\(k₁\) passages

top-\(k₀\) texts

q₁
q₂
...
qᵢ
Results 1 – First Stage

Number of Synthetic Queries vs. Primary Metric

- $\text{nDCG@10}$
- Synthetic Queries

- $\text{NQSBM25}$
- $\text{NQS+PD}_{\text{BM25}}$
- $\text{NQSBM25+RM3}$
- $\text{NQS+PD}_{\text{BM25+RM3}}$
$H_1 - \text{monoT5}_{\text{MED}}$

- $H_1$ question synthesis
- $H_0$ keyword search
- $H_1$ pointwise reranker
- $H_1'$ pointwise reranker prime

Segments of top-$k_0$ texts

Query: [query] Document: [p$_i,j$] Relevant:

Encoder $\rightarrow$ Decoder

$e_i' = \max_{\text{eligibility segments}} s_{ij}$

$\max_{\text{description segments}} a_i' = \max_{\text{all segments}} s_{ij}$

$[p_{i,j}] = \text{title: } [t_{\text{title}}] \text{ condition: } [t_{\text{condition}}] \text{ eligibility: } [t_{\text{eligibility}}] \text{ or } [p_{i,j}] = \text{title: } [t_{\text{title}}] \text{ condition: } [t_{\text{condition}}] \text{ description: } [t_{\text{description}}]$
**Fine-tuning**
Judgements from Koopman et al. 2016.
Graded Relevance
≥ 1 → true
< 1 → false

**Three input templates for training**

\[
[p_{i,j}] = \text{title: } [t_{\text{title}}] \quad \text{condition: } [t_{\text{condition}}] \quad \text{eligibility: } [t_{\text{eligibility},j}]
\]

\[
[p_{i,j}] = \text{title: } [t_{\text{title}}] \quad \text{condition: } [t_{\text{condition}}] \quad \text{description: } [t_{\text{description},j}]
\]

\[
[p_{i,j}] = \text{title: } [t_{\text{title}}] \quad \text{condition: } [t_{\text{condition}}] \quad \text{eligibility: } [t_{\text{eligibility},j}] \quad \text{description: } [t_{\text{description},j}]
\]
Results 2 – Pointwise rerankers

<table>
<thead>
<tr>
<th>Run</th>
<th>nDCG@10</th>
<th>P@10</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Median</td>
<td>0.3040</td>
<td>0.1613</td>
<td>0.2942</td>
</tr>
<tr>
<td>(b) damoebrtog</td>
<td>0.5953</td>
<td>0.4093</td>
<td>0.6083</td>
</tr>
<tr>
<td>(c) CSIROmed_inc</td>
<td>0.5320</td>
<td>0.3173</td>
<td>-</td>
</tr>
<tr>
<td>(d) BM25</td>
<td>0.2923</td>
<td>0.1680</td>
<td>0.3015</td>
</tr>
<tr>
<td>(e) BM25+RM3</td>
<td>0.3539</td>
<td>0.2040</td>
<td>0.3659</td>
</tr>
<tr>
<td>(f) NQS+PD_{BM25+RM3}</td>
<td>0.4726</td>
<td>0.2760</td>
<td>0.4304</td>
</tr>
<tr>
<td>(g) monoT5^A_{MED} +</td>
<td>0.2994</td>
<td>0.1973</td>
<td>0.3560</td>
</tr>
<tr>
<td>(h) monoT5^D_{MED} +</td>
<td>0.2311</td>
<td>0.1507</td>
<td>0.3223</td>
</tr>
<tr>
<td>(i) monoT5^E_{MED} +</td>
<td>0.4715</td>
<td>0.2987</td>
<td>0.4830</td>
</tr>
<tr>
<td>(j) monoT5^A_{CT} +</td>
<td>0.6763</td>
<td>0.5480</td>
<td>0.7253</td>
</tr>
<tr>
<td>(k) monoT5^D_{CT} +</td>
<td>0.4493</td>
<td>0.3267</td>
<td>0.6260</td>
</tr>
<tr>
<td>(l) monoT5^E_{CT} +</td>
<td>0.6792</td>
<td>0.5493</td>
<td>0.7161</td>
</tr>
<tr>
<td>(m) monoT5_{CT} +</td>
<td>0.7118</td>
<td>0.5933</td>
<td>0.8162</td>
</tr>
</tbody>
</table>
Segments of top-$k_1$ texts

$H_{1}$
question decomposition

$H_0$
keyword search

$H_1$
pointwise reranker

$H_{1}'$
pointwise reranker prime

$[P_{i,j}] = \text{title: } [t_{\text{title}}] \text{ condition: } [t_{\text{condition}}] \text{ eligibility: } [t_{\text{eligibility,best}}] \text{ description: } [t_{\text{description,best}}]$
## Results 3 – State of the Art!

<table>
<thead>
<tr>
<th>Run</th>
<th>nDCG@10</th>
<th>P@10</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>0.3040</td>
<td>0.1613</td>
<td>0.2942</td>
</tr>
<tr>
<td>damoebrtog</td>
<td>0.5953</td>
<td>0.4093</td>
<td>0.6083</td>
</tr>
<tr>
<td>CSIROmed_med</td>
<td>0.5320</td>
<td>0.3173</td>
<td>-</td>
</tr>
<tr>
<td>BM25</td>
<td>0.2923</td>
<td>0.1680</td>
<td>0.3015</td>
</tr>
<tr>
<td>BM25+RM3</td>
<td>0.3539</td>
<td>0.2040</td>
<td>0.3659</td>
</tr>
<tr>
<td>NQS+PD&lt;sub&gt;BM25+RM3&lt;/sub&gt;</td>
<td>0.4726</td>
<td>0.2760</td>
<td>0.4304</td>
</tr>
<tr>
<td>+ monoT5&lt;sub&gt;A_MED&lt;/sub&gt;</td>
<td>0.2994</td>
<td>0.1973</td>
<td>0.3560</td>
</tr>
<tr>
<td>+ monoT5&lt;sub&gt;D_MED&lt;/sub&gt;</td>
<td>0.2311</td>
<td>0.1507</td>
<td>0.3223</td>
</tr>
<tr>
<td>+ monoT5&lt;sub&gt;E_MED&lt;/sub&gt;</td>
<td>0.4715</td>
<td>0.2987</td>
<td>0.4830</td>
</tr>
<tr>
<td>+ monoT5&lt;sub&gt;A_CT&lt;/sub&gt;</td>
<td>0.6763</td>
<td>0.5480</td>
<td>0.7253</td>
</tr>
<tr>
<td>+ monoT5&lt;sub&gt;D_CT&lt;/sub&gt;</td>
<td>0.4493</td>
<td>0.3267</td>
<td>0.6260</td>
</tr>
<tr>
<td>+ monoT5&lt;sub&gt;E_CT&lt;/sub&gt;</td>
<td>0.6792</td>
<td>0.5493</td>
<td>0.7161</td>
</tr>
<tr>
<td>+ monoT5′&lt;sub&gt;CT&lt;/sub&gt;</td>
<td>0.7118</td>
<td>0.5933</td>
<td>0.8162</td>
</tr>
</tbody>
</table>
Not the friendliest Muppet in Sesame Street!
Lack of Biomedical Knowledge
<table>
<thead>
<tr>
<th>Round 4: 45 topics</th>
<th>Run</th>
<th>Type</th>
<th>nDCG@20</th>
<th>P@20</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>(4a) unique.ptr</td>
<td>UPrrf38rff3-r4†</td>
<td>feedback</td>
<td>0.7843</td>
<td>0.8211</td>
<td>0.4681</td>
</tr>
<tr>
<td>(4b) covidex</td>
<td>covidex.r4.duot5.1r</td>
<td>feedback</td>
<td>0.7745</td>
<td>0.7967</td>
<td>0.3846</td>
</tr>
<tr>
<td></td>
<td>(= expando + monoT5 + duoT5 + LR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4c) covidex</td>
<td>covidex.r4.d2q.duot5</td>
<td>automatic</td>
<td>0.7219</td>
<td>0.7267</td>
<td>0.3122</td>
</tr>
<tr>
<td></td>
<td>(= expando + monoT5 + duoT5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4d) covidex</td>
<td>covidex.r4.duot5</td>
<td>automatic</td>
<td>0.6877</td>
<td>0.6922</td>
<td>0.3283</td>
</tr>
<tr>
<td></td>
<td>(= monoT5 + duoT5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4e) uogTr</td>
<td>uogTrDPH_QE_SCB1</td>
<td>automatic</td>
<td>0.6820</td>
<td>0.7144</td>
<td>0.3457</td>
</tr>
<tr>
<td>(4f) anserini</td>
<td>r4.fusion2</td>
<td>automatic</td>
<td>0.6089</td>
<td>0.6589</td>
<td>0.3088</td>
</tr>
<tr>
<td>(4g) anserini</td>
<td>r4.fusion1</td>
<td>automatic</td>
<td>0.5244</td>
<td>0.5611</td>
<td>0.2666</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Round 5: 50 topics</th>
<th>Run</th>
<th>Type</th>
<th>nDCG@20</th>
<th>P@20</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5a) unique.ptr</td>
<td>UPrrf93-wt-r5†</td>
<td>feedback</td>
<td>0.8496</td>
<td>0.8760</td>
<td>0.4718</td>
</tr>
<tr>
<td>(5b) covidex</td>
<td>covidex.r5.2s.1r</td>
<td>feedback</td>
<td>0.8311</td>
<td>0.8460</td>
<td>0.3922</td>
</tr>
<tr>
<td></td>
<td>(= monoT5 + duoT5 + LR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5c) covidex</td>
<td>covidex.r5.d2q.2s.1r</td>
<td>feedback</td>
<td>0.8304</td>
<td>0.8380</td>
<td>0.3875</td>
</tr>
<tr>
<td></td>
<td>(= expando + monoT5 + duoT5 + LR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5d) covidex</td>
<td>covidex.r5.2s.2s</td>
<td>automatic</td>
<td>0.7539</td>
<td>0.7700</td>
<td>0.3227</td>
</tr>
<tr>
<td></td>
<td>(= monoT5 + duoT5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5e) covidex</td>
<td>covidex.r5.2s</td>
<td>automatic</td>
<td>0.7457</td>
<td>0.7610</td>
<td>0.3212</td>
</tr>
<tr>
<td></td>
<td>(= monoT5 + duoT5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5f) uogTr</td>
<td>uogTrDPH_QE_SB_CB</td>
<td>automatic</td>
<td>0.7427</td>
<td>0.7910</td>
<td>0.3305</td>
</tr>
<tr>
<td>(5g) covidex</td>
<td>covidex.r5.d2q.1s</td>
<td>automatic</td>
<td>0.7121</td>
<td>0.7320</td>
<td>0.3150</td>
</tr>
<tr>
<td></td>
<td>(= expando + monoT5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5h) anserini</td>
<td>r5.fusion2</td>
<td>automatic</td>
<td>0.6007</td>
<td>0.6440</td>
<td>0.2734</td>
</tr>
<tr>
<td>(5i) anserini</td>
<td>r5.fusion1</td>
<td>automatic</td>
<td>0.5313</td>
<td>0.5840</td>
<td>0.2314</td>
</tr>
</tbody>
</table>
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.