Information retrieval

Christopher Potts and Omar Khattab

Stanford Linguistics

CS224u: Natural language understanding
Guiding ideas
NLP is revolutionizing Information Retrieval (IR)

Understanding searches better than ever before

Bing delivers its largest improvement in search experience using Azure GPUs

Microsoft Stock Extends Gains As AI-Powered Dominant

“it’s a new reality” Satya Nadella

Bing search

MARTIN BACCARDAX

A leader in wealth management, Morgan Stanley maintains a content library with hundreds of thousands of pages of knowledge and insights spanning investment strategies, market research and commentary, and analyst insights. This vast amount of information is housed across many internal sites, largely in PDF form, requiring advisors to scan through a great deal of information to find answers to specific questions. Such searches can be time-consuming and cumbersome.

With the help of OpenAI’s GPT-4, Morgan Stanley is changing how its wealth management personnel locate relevant information.

Starting last year, the company began exploring how to harness its intellectual capital with GPT’s embeddings and retrieval capabilities—first GPT-3 and now GPT-4. The model will power an internal-facing chatbot that performs a comprehensive search of wealth management content and “effectively unlocks the cumulative knowledge of Morgan Stanley Wealth Management,” says Jeff McMillan, Head of Analytics, Data & Innovation, whose team is leading the initiative. GPT-4, his project lead notes, has finally put the ability to parse all that insight into a far more usable and actionable format.
IR is a hard NLU problem

what compounds protect the digestive system against viruses

In the **stomach**, gastric acid and proteases serve as powerful **chemical defenses** against ingested **pathogens**.
IR is revolutionizing NLP

<table>
<thead>
<tr>
<th>Standard QA</th>
<th>OpenQA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title:</strong> Bert</td>
<td><strong>Title:</strong> Sesame Street</td>
</tr>
<tr>
<td><strong>Context:</strong> Bert is a Muppet who lives with Ernie.</td>
<td><strong>Context:</strong> Bert and Ernie are Muppets who live together.</td>
</tr>
<tr>
<td><strong>Q:</strong> Who is Bert?</td>
<td><strong>Q:</strong> Who is Bert?</td>
</tr>
<tr>
<td><strong>A:</strong> Bert is a Muppet</td>
<td><strong>A:</strong> Bert is a Muppet</td>
</tr>
</tbody>
</table>

Title, Context, Question, and Answer given at train time. Title, Context, Question given a test time.

Only Question and Answer given at train time. Only Question given at test time. **Title/Context retrieved.**
Knowledge-intensive tasks

1. Question answering
2. Claim verification
3. Commonsense reasoning
4. Long-form reading comprehension
5. Information-seeking dialogue
6. Summarization
7. Natural language inference
Classical IR

When was Stanford University founded?

Term look-up

- founded
- fountain
- Stamford
- Stanford
- University

Document scoring

- doc39: A History of Stanford University
- doc64: Stanford University About Page
“LLMs for everything”

When was Stanford University founded?

Stanford University was founded in 1891.
Neural IR

When was Stanford University founded?

Scoring and extraction

“Stanford University was founded in 1885 by California senator Leland Stanford and his wife, Jane”
A History of Stanford University

“Stanford was founded in 1885 by Leland and Jane Stanford in memory of their only child, Leland Stanford Jr.”
Stanford University Wikipedia

“Opened in 1891”
Stanford University About Page
Retrieval-augmented in-context learning

Context: Kermit is one of the stars of Sesame Street.

Q: Who is Kermit?

A: Kermit is the one of the stars of Sesame Street.

Context: Bert is a Muppet who lives with Ernie.

Q: Who is Bert?

A: Bert is a Muppet.
IR is more important than ever!
IR is more important than ever!

Google's Bard AI bot mistake wipes $100bn off shares
Google's Bard AI bot mistake wipes $100bn off shares

Please answer the following question and provide references for your answer. Which web pages can back you up?

Q: Are professional baseball players allowed to glue small wings to their cap?

A: No, professional baseball players are not allowed to glue small wings to their cap. According to MLB rules, players are prohibited from altering their caps in any way (1). This includes gluing wings or any other objects to their caps. The only exception is that players may attach MLB-approved patches (2).

References:
IR is more important than ever!

Google's Bard AI bot mistake wipes $100bn off shares

Please answer the following question and provide references for your answer. Which web pages can back you up?

Q: Are professional baseball players allowed to glue small wings to their cap?

A: No, professional baseball players are not allowed to glue small wings to their cap. According to MLB rules, players are prohibited from altering their caps in any way (1). This includes gluing wings or any other objects to their caps. The only exception is that players may attach MLB-approved patches (2).

References:
IR is more important than ever!

Google's Bard AI bot mistake wipes $100bn off shares

Please answer the following question and provide references for your answer. Which web pages can back you up?

Q: Are professional baseball players allowed to glue small wings to their cap?

A: No, professional baseball players are not allowed to glue small wings to their cap. According to MLB rules, players are prohibited from altering their caps in any way (1). This includes gluing or any other objects to their caps. The only exception is that players may attach MLB-approved patches (2).

References:

How many storeys are in the castle David Gregory inherited?

Vanilla LM
LM: Castle Gregory has three storeys.

Retrieve-then-Read
RM: “St. Gregory Hotel is a nine-floor boutique hotel in D.C...”
LM: St. Gregory Hotel has nine storeys.

Multi-Hop DSP Program
LM: “Which castle did David Gregory inherit?”
RM: “David Gregory inherited Kinnaird Castle in 1664...”
LM: “How many storeys does Kinnaird Castle have?”
RM: “Kinnaird Castle is a tower house, having five storeys...”
LM: Kinnaird Castle has five storeys.

Fabricated!
IR is more important than ever!

Google's Bard AI bot mistake wipes $100bn off shares

Please answer the following question and provide references for your answer. Which web pages can back you up?

Q: Are professional baseball players allowed to glue small wings to their cap?

A: No, professional baseball players are not allowed to glue small wings to their cap. According to MLB rules, players are prohibited from altering their caps in any way. This includes gluing wings or any other objects to their caps. The only exception is that players may attach MLB-approved patches.

References:
IR is more important than ever!

Google's Bard AI bot mistake wipes $100bn off shares

Please answer the following question and provide references for your answer. Which web pages can back you up?

Q: Are professional baseball players allowed to glue small wings to their cap?

A: No, professional baseball players are not allowed to glue small wings to their cap. According to MLB rules, players are prohibited from altering their caps in any way. This includes gluing wings or any other objects to their caps. The only exception is that players may attach MLB-approved patches.

References:

How many storeys are in the castle David Gregory inherited?

- Searching for: castle David Gregory inherited
- Searching for: Kinnairdy Castle storeys
- Generating answers for you...

David Gregory inherited Kinnairdy Castle, a tower house in Aberdeenshire, Scotland. According to some sources, Kinnairdy Castle has five storeys and a garret. Is this what you were looking for?


Fabricated!

Our paper!
IR is more important than ever!

Google's Bard AI bot mistake wipes $100bn off stock

Google unveiled its new bot called Bard, which is the latest in a line of AI-powered tools from the tech giant. The tool, which is similar to ChatGPT, uses machine learning and natural language processing (NLP) to respond to queries. The tool has been widely used in various fields, including research, education, and entertainment. However, recent reports have suggested that the tool's performance is not as good as expected. Additionally, the tool has been criticized for its tendency to generate incorrect information. In this case, the tool incorrectly stated that Christopher Potts is a renowned linguist and cognitive scientist. The tool then went on to generate a biography of Christopher Potts that contained inaccuracies. For example, it stated that Potts was born on May 15, 1972, in New York City, USA. He developed a deep curiosity for language from a young age, and his passion for understanding the intricacies of human communication led him to pursue higher education in linguistics. He completed his Bachelor's degree in Linguistics from Yale University, where he graduated magna cum laude. He then went on to earn his Ph.D. in Linguistics from the University of Massachusetts, Amherst, under the supervision of renowned linguist Barbara Partee. However, these statements are factually incorrect. For example, Potts was actually born in New York City, USA, in 1972, and he completed his Bachelor's degree in Linguistics from Yale University, where he graduated magna cum laude. He then went on to earn his Ph.D. in Linguistics from the University of Massachusetts, Amherst, under the supervision of renowned linguist Barbara Partee. Additionally, it is important to note that the tool's tendency to generate incorrect information highlights the need for further research and development in the field of AI and NLP.
Blog posts

Building Scalable, Explainable, and Adaptive NLP Models with Retrieval
Omar Khattab, Christopher Potts, and Matei Zaharia
October 5, 2021

A Moderate Proposal for Radically Better AI-powered Web Search
Large language models could give us instant answers, but at a cost to trust. Stanford scholars propose an alternative.
Jul 6, 2021 | Omar Khattab, Christopher Potts, and Matei Zaharia
Classical IR
The term–document matrix

<table>
<thead>
<tr>
<th></th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
<th>d4</th>
<th>d5</th>
<th>d6</th>
<th>d7</th>
<th>d8</th>
<th>d9</th>
<th>d10</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>against</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>agent</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ages</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ago</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>agree</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ahead</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ain’t</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>air</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>aka</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
**TF-IDF**

For a corpus of documents $D$:

- Term frequency: $\text{TF}(w, \text{doc}) = \frac{\text{count}(w, \text{doc})}{|\text{doc}|}$

- Document frequency: $\text{df}(w, D) = |\{\text{doc} \in D : w \in \text{doc}\}|$

- Inverse document frequency: $\text{IDF}(w, D) = \log_e \left( \frac{|D|}{\text{df}(w, D)} \right)$

- $\text{TF-IDF}(w, \text{doc}, D) = \text{TF}(w, \text{doc}) \cdot \text{IDF}(w, D)$

### Example

<table>
<thead>
<tr>
<th></th>
<th>doc1</th>
<th>doc2</th>
<th>doc3</th>
<th>doc4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>$B$</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>$C$</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$D$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>0.00</td>
</tr>
<tr>
<td>$B$</td>
<td>0.29</td>
</tr>
<tr>
<td>$C$</td>
<td>0.69</td>
</tr>
<tr>
<td>$D$</td>
<td>1.39</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>0.33</td>
</tr>
<tr>
<td>$B$</td>
<td>0.33</td>
</tr>
<tr>
<td>$C$</td>
<td>0.33</td>
</tr>
<tr>
<td>$D$</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TF-IDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>0.00</td>
</tr>
<tr>
<td>$B$</td>
<td>0.10</td>
</tr>
<tr>
<td>$C$</td>
<td>0.23</td>
</tr>
<tr>
<td>$D$</td>
<td>0.00</td>
</tr>
</tbody>
</table>
IDF values

|D| = 10

$\text{IDF}(w, D) = \log \left( \frac{|D|}{\text{df}(w, D)} \right)$
Relevance scores

$$\text{RelevanceScore}(q, \text{doc}, D) = \sum_{w \in q} \text{Weight}(w, \text{doc}, D)$$

where **Weight** is often TF-IDF.
BM25

**Smoothed IDF**

\[
IDF_{BM25}(w, D) = \log_e \left( 1 + \frac{|D| - df(w, D) + 0.5}{df(w, D) + 0.5} \right)
\]

**Scoring**

With \( k = 1.2 \) and \( b = 0.75 \) (or thereabouts):

\[
Score_{BM25}(w, doc) = \frac{TF(w, doc) \cdot (k + 1)}{TF(w, doc) + k \cdot \left( 1 - b + b \cdot \frac{|doc|}{avgdoclen} \right)}
\]

**BM25 Weight**

\[
BM25(w, doc, D) = Score_{BM25}(w, doc) \cdot IDF_{BM25}(w, D)
\]

Best Match, Attempt #25; Robertson and Zaragoza 2009
BM25 IDF values

$$IDF_{BM25}(w, D) = \log_e\left(1 + \frac{|D| - df(w, D) + s}{df(w, D) + s}\right)$$
BM25 Scores: avgdoclen

\[
\text{Score}_{BM25}(w, \text{doc}) = \frac{\text{TF}(w, \text{doc}) \cdot (k + 1)}{\text{TF}(w, \text{doc}) + k \cdot (1 - b + b \cdot \frac{|\text{doc}|}{\text{avgdoclen}})}
\]

Penalizes long documents
BM25 Scores: $b$

$$\text{Score}_{BM25}(w, \text{doc}) = \frac{\text{TF}(w, \text{doc}) \cdot (k + 1)}{\text{TF}(w, \text{doc}) + k \cdot \left(1 - b + b \cdot \frac{|\text{doc}|}{\text{avgdoclen}}\right)}$$

$b$ controls the doc length penalty
**BM25 Scores: \( k \)**

**Formula:**

\[
\text{Score}_{BM25}(w, \text{doc}) = \frac{\text{TF}(w, \text{doc}) \cdot (k + 1)}{\text{TF}(w, \text{doc}) + k \cdot (1 - b + b \cdot \frac{|\text{doc}|}{\text{avgdoclen}})}
\]

**Flattens out higher frequencies**
Inverted indices

When was Stanford University founded?

Term look-up

- founded
  - doc47, doc39, doc41, ...
- fountain
  - doc21, doc64, doc16, ...
- Stamford
  - doc21, doc11, doc17, ...
- Stanford
  - doc47, doc39, doc68, ...
- University
  - doc21, doc39, doc68, ...

Document scoring

- doc39: A History of Stanford University
- doc64: Stanford University About Page
Inverted indices

When was Stanford University founded?

Term look-up

- founded
  - (doc47, 0.90), (doc39, 0.76), (doc41, 0.76), ...
  - 0.12
- fountain
  - (doc21, 0.65), (doc64, 0.60), (doc16, 0.10), ...
  - 0.88
- Stamford
  - (doc21, 0.91), (doc11, 0.89), (doc17, 0.50), ...
  - 0.01
- Stanford
  - (doc47, 0.29), (doc39, 0.01), (doc68, 0.10), ...
  - 0.56
- University
  - (doc21, 0.91), (doc39, 0.90), (doc68, 0.76), ...
  - 0.01

Document scoring

- doc39: A History of Stanford University
- doc64: Stanford University About Page
Beyond term matching

1. Query and document expansion
2. Phrase search
3. Term dependence
4. Different document fields (e.g., title, body)
5. Link analysis (e.g., PageRank)
6. Learning to rank
Tools for classical IR

1. Elasticsearch
   https://www.elastic.co

2. Pyserini:
   https://github.com/castorini/pyserini

3. PrimeQA
   https://github.com/primeqa/primeqa
IR metrics
Many dimensions

1. **Accuracy-style metrics**: These will be our focus.
2. **Latency**: Time to execute a single query.
3. **Throughput**: Total queries served in a fixed time, perhaps via batch processing.
4. **FLOPs**: Hardware agnostic measure of compute resources.
5. **Disk usage**: For the model, index, etc.
6. **Memory usage**: For the model, index, etc.
7. **Cost**: Total cost of deployment for a system.
Relevance data types

Given a query \( q \) and a collection of \( N \) documents \( D \):

1. A complete partial gold ranking \( \mathbf{D} = [\text{doc}_1, \ldots, \text{doc}_N] \) of \( D \) with respect to \( q \).
   - Unlikely unless \( \mathbf{D} \) was automatically generated.

2. An incomplete partial ranking of \( D \) with respect to \( q \).

3. Labels for which passages in \( D \) are relevant to \( q \).
   - Could be based in a weak supervision heuristic like whether each \( \text{doc}_i \) contains \( q \) as a substring.

4. A tuple consisting of one positive document \( \text{doc}^+ \) for \( q \) and one or more negatives \( \text{doc}^- \) for \( q \).
Success and Reciprocal Rank

**Rank**
For a ranking \( D = [\text{doc}_1, \ldots, \text{doc}_N] \), let

\[
\text{Rank}(q, D) \in \{1, 2, 3, \ldots\}
\]
be the position of the \textbf{first} relevant document for \( q \) in \( D \).

**Success**
\[
\text{Success@}_K(q, D) = \begin{cases} 
1 & \text{if } \text{Rank}(q, D) \leq K \\
0 & \text{otherwise}
\end{cases}
\]

**Reciprocal Rank**
\[
\text{RR@}_K(q, D) = \begin{cases} 
\frac{1}{\text{Rank}(q, D)} & \text{if } \text{Rank}(q, D) \leq K \\
0 & \text{otherwise}
\end{cases}
\]
MRR@\( K \) is the average of this over multiple queries.
Success and Reciprocal Rank: A comparison

<table>
<thead>
<tr>
<th>$D_1$ for $q$</th>
<th>$D_2$ for $q$</th>
<th>$D_3$ for $q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 doc$_C$ ★</td>
<td>1 doc$_A$</td>
<td>1 doc$_D$</td>
</tr>
<tr>
<td>2 doc$_E$ ★</td>
<td>2 doc$_C$ ★</td>
<td>2 doc$_B$</td>
</tr>
<tr>
<td>3 doc$_D$</td>
<td>3 doc$_G$</td>
<td>3 doc$_E$ ★</td>
</tr>
<tr>
<td>4 doc$_B$</td>
<td>4 doc$_B$</td>
<td>4 doc$_C$ ★</td>
</tr>
<tr>
<td>5 doc$_A$</td>
<td>5 doc$_E$ ★</td>
<td>5 doc$_F$ ★</td>
</tr>
<tr>
<td>6 doc$_F$ ★</td>
<td>6 doc$_F$ ★</td>
<td>6 doc$_A$</td>
</tr>
</tbody>
</table>

- $\text{Success@2}(q, D_1) = 1$
- $\text{RR@2}(q, D_1) = 1/1$
- $\text{Success@2}(q, D_2) = 1$
- $\text{RR@2}(q, D_2) = 1/2$
- $\text{Success@2}(q, D_3) = 0$
- $\text{RR@2}(q, D_3) = 0$
**Precision and Recall**

Ret\((D, K)\) is the set of documents at or above \(K\) in \(D\).

Rel\((D, q)\) is the set of all documents that are relevant to \(q\).

**Precision**

\[
\text{Prec@K}(q, D) = \frac{|\text{Ret}(D, K) \cap \text{Rel}(D, q)|}{K}
\]

**Recall**

\[
\text{Rec@K}(q, D) = \frac{|\text{Ret}(D, K) \cap \text{Rel}(D, q)|}{|\text{Rel}(D, q)|}
\]
## Precision and Recall examples

<table>
<thead>
<tr>
<th>( D_1 ) for ( q )</th>
<th>( D_2 ) for ( q )</th>
<th>( D_3 ) for ( q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1    ( \text{doc}_C ) ★</td>
<td>1    ( \text{doc}_A )</td>
<td>1    ( \text{doc}_D )</td>
</tr>
<tr>
<td>2    ( \text{doc}_E ) ★</td>
<td>2    ( \text{doc}_C ) ★</td>
<td>2    ( \text{doc}_B )</td>
</tr>
<tr>
<td>3    ( \text{doc}_D )</td>
<td>3    ( \text{doc}_G )</td>
<td>3    ( \text{doc}_E ) ★</td>
</tr>
<tr>
<td>4    ( \text{doc}_B )</td>
<td>4    ( \text{doc}_B )</td>
<td>4    ( \text{doc}_C ) ★</td>
</tr>
<tr>
<td>5    ( \text{doc}_A )</td>
<td>5    ( \text{doc}_E ) ★</td>
<td>5    ( \text{doc}_F ) ★</td>
</tr>
<tr>
<td>6    ( \text{doc}_F ) ★</td>
<td>6    ( \text{doc}_F ) ★</td>
<td>6    ( \text{doc}_A )</td>
</tr>
</tbody>
</table>

- \( \text{Prec}@2(q, D_1) = 2/2 \)
- \( \text{Rec}@2(q, D_1) = 2/3 \)
- \( \text{Prec}@2(q, D_2) = 1/2 \)
- \( \text{Rec}@2(q, D_2) = 1/3 \)
- \( \text{Prec}@2(q, D_3) = 0/2 \)
- \( \text{Rec}@2(q, D_3) = 0/3 \)
## Precision and Recall examples

<table>
<thead>
<tr>
<th>$D_1$ for $q$</th>
<th>$D_2$ for $q$</th>
<th>$D_3$ for $q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 $doc_C$ ★</td>
<td>1 $doc_A$</td>
<td>1 $doc_D$</td>
</tr>
<tr>
<td>2 $doc_E$ ★</td>
<td>2 $doc_C$ ★</td>
<td>2 $doc_B$</td>
</tr>
<tr>
<td>3 $doc_D$</td>
<td>3 $doc_G$</td>
<td>3 $doc_E$ ★</td>
</tr>
<tr>
<td>4 $doc_B$</td>
<td>4 $doc_B$</td>
<td>4 $doc_C$ ★</td>
</tr>
<tr>
<td>5 $doc_A$</td>
<td>5 $doc_E$ ★</td>
<td>5 $doc_F$ ★</td>
</tr>
<tr>
<td>6 $doc_F$ ★</td>
<td>6 $doc_F$ ★</td>
<td>6 $doc_A$</td>
</tr>
</tbody>
</table>

- Prec@5($q$, $D_1$) = 2/5
- Rec@5($q$, $D_1$) = 2/3
- Prec@5($q$, $D_2$) = 2/5
- Rec@5($q$, $D_2$) = 2/3
- Prec@5($q$, $D_3$) = 3/5
- Rec@5($q$, $D_3$) = 3/3
### Average Precision

$$\text{AvgPrec}(q, D) = \frac{\sum_{i=1}^{\mid D \mid} \text{Prec@i}(q, D) \cdot \text{Rel}(q, \text{doc}_i)}{|\text{Rel}(D, q)|}$$

<table>
<thead>
<tr>
<th>$D_1$ for $q$</th>
<th>$D_2$ for $q$</th>
<th>$D_3$ for $q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 doc$_C$ *</td>
<td>1 doc$_A$</td>
<td>1 doc$_D$</td>
</tr>
<tr>
<td>2 doc$_E$ *</td>
<td>2 doc$_C$ *</td>
<td>2 doc$_B$</td>
</tr>
<tr>
<td>3 doc$_D$</td>
<td>3 doc$_G$</td>
<td>3 doc$_E$ *</td>
</tr>
<tr>
<td>4 doc$_B$</td>
<td>4 doc$_B$</td>
<td>4 doc$_C$ *</td>
</tr>
<tr>
<td>5 doc$_A$</td>
<td>5 doc$_E$ *</td>
<td>5 doc$_F$ *</td>
</tr>
<tr>
<td>6 doc$_F$ *</td>
<td>6 doc$_F$ *</td>
<td>6 doc$_A$</td>
</tr>
</tbody>
</table>

- $\text{Prec@1}(q, D) = 1/1 + 2.5/3$
- $\text{Prec@2}(q, D) = 1/2 + 1.4/3$
- $\text{Prec@3}(q, D) = 1/3 + 1.43/3$
Which metric? There is no single answer!

1. Is the cost of scrolling through K passages low? Then perhaps Success@K is fine-grained enough.

2. Are there multiple relevant documents per query? If so, Success@K and RR@K may be too coarse-grained.

3. Is it more important to find every relevant document? If so, favor Recall.

4. Is it more important to review only relevant documents? If so, favor Precision.

5. F1@K is the harmonic mean of Prec@K and Recall@K. It can be used where there are multiple relevant documents but their relative order above K doesn’t matter that much.

6. AvgPrec will give the finest-grained distinctions of the metrics discussed here: it is sensitive to rank, precision, and recall.
## Beyond accuracy

<table>
<thead>
<tr>
<th>Guiding ideas</th>
<th>Classical IR</th>
<th>IR metrics</th>
<th>Neural IR</th>
<th>Datasets</th>
<th>Conclusion</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPU</strong></td>
<td><strong>CPU</strong></td>
</tr>
<tr>
<td>BM25 (Mackenzie et al., 2021)</td>
<td>0</td>
</tr>
<tr>
<td>BM25 (Lassance and Clinchant, 2022)</td>
<td>0</td>
</tr>
<tr>
<td>SPLADEv2-distil (Mackenzie et al., 2021)</td>
<td>0</td>
</tr>
<tr>
<td>SPLADEv2-distil (Lassance and Clinchant, 2022)</td>
<td>0</td>
</tr>
<tr>
<td>BT-SPLADE-S (Lassance and Clinchant, 2022)</td>
<td>0</td>
</tr>
<tr>
<td>BT-SPLADE-M (Lassance and Clinchant, 2022)</td>
<td>0</td>
</tr>
<tr>
<td>BT-SPLADE-L (Lassance and Clinchant, 2022)</td>
<td>0</td>
</tr>
<tr>
<td>ANCE (Xiong et al., 2020)</td>
<td>1</td>
</tr>
<tr>
<td>RocketQAv2 (Ren et al., 2021)</td>
<td>-</td>
</tr>
<tr>
<td>coCondenser (Gao and Callan, 2021)</td>
<td>-</td>
</tr>
<tr>
<td>CoT-MAE (Wu et al., 2022)</td>
<td>-</td>
</tr>
<tr>
<td>ColBERTv1 (Khattab and Zaharia, 2020)</td>
<td>4</td>
</tr>
<tr>
<td>PLAID ColBERTv2 (Santhanam et al., 2022a)</td>
<td>4</td>
</tr>
<tr>
<td>PLAID ColBERTv2 (Santhanam et al., 2022a)</td>
<td>4</td>
</tr>
<tr>
<td>DESSERT (Engels et al., 2022)</td>
<td>0</td>
</tr>
</tbody>
</table>

Santhanam et al. 2022c
Beyond accuracy

Santhanam et al. 2022c
Neural IR
Cross-encoders

1. Examples: \(\langle q_i, \text{doc}^+_i, \{\text{doc}^-_{i,k}\} \rangle\)

2. For a BERT-style encoder with \(N\) layers:

\[
\text{Rep}(q, \text{doc}) = \text{Dense}(\text{Enc}([q; \text{doc}]_{N,0}))
\]

3. Loss: negative log-likelihood of the positive passage

\[
- \log \frac{\exp\left(\text{Rep}(q_i, \text{doc}^+_i)\right)}{\exp\left(\text{Rep}(q_i, \text{doc}^+_i)\right) + \sum_{j=1}^{n} \exp\left(\text{Rep}(q_i, \text{doc}^-_{i,j})\right)}
\]

Incredibly rich, but won’t scale!
DPR

1. Examples: \( \langle q_i, \text{doc}_i^+, \{\text{doc}_{i,k}^- \} \rangle \)

2. For a BERT-style encoder with \( N \) layers:
\[
\text{Sim}(q, \text{doc}) = \text{Enc}_Q(q)_{N,0}^T \text{Enc}_D(\text{doc})_{N,0}
\]

3. Loss: negative log-likelihood of the positive passage
\[
-\log \frac{\exp(\text{Sim}(q_i, \text{doc}_i^+))}{\exp(\text{Sim}(q_i, \text{doc}_i^+)) + \sum_{j=1}^n \exp(\text{Sim}(q_i, \text{doc}_{i,j}^-))}
\]

Highly scalable, but limited query/doc interactions!

Karpukhin et al. 2020
Shared loss function

The negative log-likelihood of the positive passage:

Cross encoders

$$- \log \frac{\exp(\text{Rep}(q_i, doc_i^+))}{\exp(\text{Rep}(q_i, doc_i^+)) + \sum_{j=1}^{n} \exp(\text{Rep}(q_i, doc_{i,j}^-))}$$

DPR

$$- \log \frac{\exp(\text{Sim}(q_i, doc_i^+))}{\exp(\text{Sim}(q_i, doc_i^+)) + \sum_{j=1}^{n} \exp(\text{Sim}(q_i, doc_{i,j}^-))}$$

General form

$$- \log \frac{\exp(\text{Cmp}(q_i, doc_i^+))}{\exp(\text{Cmp}(q_i, doc_i^+)) + \sum_{j=1}^{n} \exp(\text{Cmp}(q_i, doc_{i,j}^-))}$$
ColBERT

**MaxSim** = .97 + .84 + .85

1. Examples:
   \( \langle q_i, \text{doc}_i^+, \{\text{doc}^-_{i,k} \} \rangle \)

2. Loss: negative log-likelihood of the positive passage, with MaxSim as the basis.

Highly scalable with late, contextual interactions!

For a BERT-style encoder with \( N \) layers:

\[
\text{MaxSim}(q, \text{doc}) = \sum_i^L \max_j^M \text{Enc}(q)_{N,i}^T \text{Enc}(\text{doc})_{N,j}
\]

with \( L \) is the length of \( q \), \( M \) the length of \( \text{doc} \).

Khattab and Zaharia 2020
Soft alignment with ColBERT

The animated Transformers was released in August 1986.
ColBERT as a reranker

Given query $q = [w^1, \ldots, w^M]$:

1. Get the top K documents for $q$ using a fast, term-based model like BM25.
2. Score each of those top K documents using ColBERT.
Beyond reranking for ColBERT

Given query \( q \) encoded as vectors \([w^1, \ldots, w^M]\), for each query vector \( w^i \):

1. Retrieve the \( \rho \) most similar token vectors \( w^k_j \) to \( w^i \).

2. Score each \( \text{doc}_j \) using ColBERT.
Centroid-based ranking

Given $q$ encoded as $[w_1, \ldots, w^M]$, for each vector $w^i$:

1. Retrieve the $p$ centroids closest to $w^i$.
2. Retrieve the $t$ most similar token vectors $w^k_j$ to any of the centroids.
3. Score each doc$^j$ using ColBERT.
ColBERT latency analysis

Santhanam et al. 2022a
ColBERT latency analysis

Santhanam et al. 2022a
ColBERT latency analysis

Initial use of centroids for pruning

Santhanam et al. 2022a
ColBERT latency analysis

Memory overhead from centroid and residual retrieval over a huge index.

Santhanam et al. 2022a
Additional ColBERT optimizations

PLAID generates many more candidates and then filters them extremely efficiently.

Santhanam et al. 2022a
1. $S_{ij} = \text{transform} (\text{Enc}(t)_N, i)^T \text{Emb}(w_j) + b_j$

where
\[
\text{transform}(x) = \text{LayerNorm}(\text{GeLU}(xW + b))
\]
and $\text{Emb}(w)$ is the embedding for $w$.

2. $\text{SPLADE}(t, w_j) = \sum_{i=1}^{M} \log(1 + \text{ReLU}(S_{ij}))$

3. $\text{Sim}_{\text{SPLADE}}(q, \text{doc}) = \text{SPLADE}(q)^T \text{SPLADE}(\text{doc})$

4. Loss: Usual negative log-likelihood plus a regularization term that leads to sparse, balanced scores.

Formal et al. 2021
Additional recent developments

This is an incredibly fast-moving field, but here are some selected developments that caught my attention. I confess that these are heavily biased towards ColBERT:

1. CITADEL (Li et al. 2022) is a lightning fast ColBERT-style model.
2. Lassance and Clinchant (2022) developed lightning fast SPLADE variants.
3. DESSERT (Engels et al. 2022) offer ultra-efficient approximate embedding search.
4. Lin et al. (2020) distill ColBERT into a single-vector model akin to DPR.
5. DR.DECR Li et al. (2021) distills multilingual ColBERT models.
6. Choi et al. (2021) distill cross-encoders into ColBERT models.
7. Lee et al. (2023) rework the standard ColBERT objective so that important tokens are retrieved first for blazing fast retrieval.
## Multidimensional benchmarking

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPU</strong></td>
<td><strong>Latency</strong></td>
</tr>
<tr>
<td><strong>BM25</strong></td>
<td>0</td>
</tr>
<tr>
<td>BM25</td>
<td>0</td>
</tr>
<tr>
<td>DPR</td>
<td>146</td>
</tr>
<tr>
<td>ColBERTv2-S</td>
<td>206</td>
</tr>
<tr>
<td>ColBERTv2-M</td>
<td>321</td>
</tr>
<tr>
<td>ColBERTv2-L</td>
<td>459</td>
</tr>
<tr>
<td>BT-SPLADE-L</td>
<td>46</td>
</tr>
<tr>
<td>BM25</td>
<td>1</td>
</tr>
<tr>
<td>DPR</td>
<td>18</td>
</tr>
<tr>
<td>ColBERTv2-S</td>
<td>27</td>
</tr>
<tr>
<td>ColBERTv2-M</td>
<td>36</td>
</tr>
<tr>
<td>ColBERTv2-L</td>
<td>55</td>
</tr>
<tr>
<td>BT-SPLADE-L</td>
<td>33</td>
</tr>
</tbody>
</table>

Selected MS MARCO results form Santhanam et al. 2022c
# Multidimensional benchmarking

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU</td>
<td>CPU</td>
</tr>
<tr>
<td>BM25</td>
<td>0</td>
</tr>
<tr>
<td>BM25</td>
<td>0</td>
</tr>
<tr>
<td>DPR</td>
<td>146</td>
</tr>
<tr>
<td>ColBERTv2-S</td>
<td>206</td>
</tr>
<tr>
<td>ColBERTv2-M</td>
<td>321</td>
</tr>
<tr>
<td>ColBERTv2-L</td>
<td>459</td>
</tr>
<tr>
<td>BT-SPLADE-L</td>
<td>46</td>
</tr>
<tr>
<td>BM25</td>
<td>1</td>
</tr>
<tr>
<td>DPR</td>
<td>18</td>
</tr>
<tr>
<td>ColBERTv2-S</td>
<td>27</td>
</tr>
<tr>
<td>ColBERTv2-M</td>
<td>36</td>
</tr>
<tr>
<td>ColBERTv2-L</td>
<td>55</td>
</tr>
<tr>
<td>BT-SPLADE-L</td>
<td>33</td>
</tr>
</tbody>
</table>

Selected MS MARCO results form Santhanam et al. 2022c
Multidimensional benchmarking

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPU</strong></td>
<td><strong>CPU</strong></td>
</tr>
<tr>
<td>BM25</td>
<td>0</td>
</tr>
<tr>
<td>BM25</td>
<td>0</td>
</tr>
<tr>
<td>DPR</td>
<td></td>
</tr>
<tr>
<td>ColBERTv2-S</td>
<td></td>
</tr>
<tr>
<td>ColBERTv2-M</td>
<td></td>
</tr>
<tr>
<td>ColBERTv2-L</td>
<td></td>
</tr>
<tr>
<td>BT-SPLADE-L</td>
<td></td>
</tr>
</tbody>
</table>

Selected MS MARCO results form Santhanam et al. 2022c
## Multidimensional benchmarking

<table>
<thead>
<tr>
<th>Hardware</th>
<th>BM25</th>
<th>0</th>
<th>1</th>
<th>4</th>
<th>m6gd.med</th>
<th>11</th>
<th>$0.14</th>
<th>38.6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BM25</td>
<td>0</td>
<td>1</td>
<td>32</td>
<td>x2gd.lrg</td>
<td>10</td>
<td>$0.48</td>
<td>38.6</td>
</tr>
<tr>
<td></td>
<td>DPR</td>
<td>146</td>
<td></td>
<td></td>
<td></td>
<td>146</td>
<td>$6.78</td>
<td>52.1</td>
</tr>
<tr>
<td></td>
<td>ColIBERTv2-S</td>
<td>206</td>
<td></td>
<td></td>
<td></td>
<td>206</td>
<td>$9.58</td>
<td>68.8</td>
</tr>
<tr>
<td></td>
<td>ColIBERTv2-M</td>
<td>321</td>
<td></td>
<td></td>
<td></td>
<td>321</td>
<td>$14.90</td>
<td>69.6</td>
</tr>
<tr>
<td></td>
<td>ColIBERTv2-L</td>
<td>459</td>
<td></td>
<td></td>
<td></td>
<td>459</td>
<td>$21.30</td>
<td>69.7</td>
</tr>
<tr>
<td></td>
<td>BT-SPLADE-L</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td>46</td>
<td>$2.15</td>
<td>66.3</td>
</tr>
<tr>
<td></td>
<td>BM25</td>
<td>1</td>
<td>16</td>
<td>32</td>
<td>p3.8xl</td>
<td>9</td>
<td>$29.94</td>
<td>38.6</td>
</tr>
<tr>
<td></td>
<td>DPR</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>18</td>
<td>$61.06</td>
<td>52.1</td>
</tr>
<tr>
<td></td>
<td>ColIBERTv2-S</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td>27</td>
<td>$90.41</td>
<td>68.8</td>
</tr>
<tr>
<td></td>
<td>ColIBERTv2-M</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td>36</td>
<td>$123.35</td>
<td>69.6</td>
</tr>
<tr>
<td></td>
<td>ColIBERTv2-L</td>
<td>55</td>
<td></td>
<td></td>
<td></td>
<td>55</td>
<td>$187.24</td>
<td>69.7</td>
</tr>
<tr>
<td></td>
<td>BT-SPLADE-L</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td>33</td>
<td>$112.87</td>
<td>66.3</td>
</tr>
</tbody>
</table>

### Selected MS MARCO results form Santhanam et al. 2022c
## Multidimensional benchmarking

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU</td>
<td>CPU</td>
</tr>
<tr>
<td>BM25</td>
<td>0</td>
</tr>
<tr>
<td>BM25</td>
<td>0</td>
</tr>
<tr>
<td>DPR</td>
<td></td>
</tr>
<tr>
<td>ColBERTv2-S</td>
<td></td>
</tr>
<tr>
<td>ColBERTv2-M</td>
<td></td>
</tr>
<tr>
<td>ColBERTv2-L</td>
<td></td>
</tr>
<tr>
<td>BT-SPLADE-L</td>
<td></td>
</tr>
<tr>
<td>BM25</td>
<td>1</td>
</tr>
<tr>
<td>DPR</td>
<td></td>
</tr>
<tr>
<td>ColBERTv2-S</td>
<td></td>
</tr>
<tr>
<td>ColBERTv2-M</td>
<td></td>
</tr>
<tr>
<td>ColBERTv2-L</td>
<td></td>
</tr>
<tr>
<td>BT-SPLADE-L</td>
<td></td>
</tr>
</tbody>
</table>

Selected MS MARCO results form Santhanam et al. 2022c
Datasets
TREC

1. **Text REtrieval Conference** (TREC) has annual competitions for comparing IR systems.
2. The 2023 iteration has a number of tracks: https://trec.nist.gov/pubs/call2023.html
3. TREC tends to emphasize careful evaluation with a very small set of queries (e.g., 50 queries, each with >100 annotated documents).
4. Having few test queries does not imply few documents!
**MS MARCO ranking tasks**

1. MS MARCO Ranking is the largest public IR benchmark.
2. It is adapted from a Question Answering dataset.
3. It consists of more than 500k Bing search queries.
4. Sparse labels: approx. one relevance label per query!
5. Fantastic for training IR models!
6. Passage Ranking: 9M short passages; sparse labels.
BEIR: Benchmarking IR

For testing models in zero-shot scenarios:

<table>
<thead>
<tr>
<th>Split (→)</th>
<th>Task (↓)</th>
<th>Domain (↓)</th>
<th>Dataset (↓)</th>
<th>Title</th>
<th>Relevancy</th>
<th>Train #Pairs</th>
<th>Dev #Query</th>
<th>Test #Query</th>
<th>Test #Corpus</th>
<th>Avg. D / Q</th>
<th>Avg. Word Lengths</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Passage-Retrieval</td>
<td>Misc.</td>
<td>MS MARCO [45]</td>
<td>X</td>
<td>Binary</td>
<td>532,761</td>
<td>6,980</td>
<td>8,841,823</td>
<td>1.1</td>
<td>5.96</td>
<td>55.98</td>
</tr>
<tr>
<td></td>
<td>Bio-Medical Information Retrieval (IR)</td>
<td>Bio-Medical</td>
<td>TREC-COVID [65]</td>
<td>✓</td>
<td>3-level</td>
<td>100,575</td>
<td>324</td>
<td>7,633</td>
<td>38.2</td>
<td>3.20</td>
<td>232.26</td>
</tr>
<tr>
<td></td>
<td>Question Answering (QA)</td>
<td>Wikipedia</td>
<td>NQ [34]</td>
<td>✓</td>
<td>Binary</td>
<td>132,803</td>
<td>5,447</td>
<td>5,233,329</td>
<td>2.0</td>
<td>17.61</td>
<td>46.30</td>
</tr>
<tr>
<td></td>
<td>Tweet-Retrieval</td>
<td>Twitter</td>
<td>Signal-1M (RT) [59]</td>
<td>X</td>
<td>3-level</td>
<td>1,146</td>
<td>49</td>
<td>57,638</td>
<td>2.6</td>
<td>10.77</td>
<td>132.32</td>
</tr>
<tr>
<td></td>
<td>News Retrieval</td>
<td>News</td>
<td>TREC-NEWS [58]</td>
<td>✓</td>
<td>5-level</td>
<td>97,000</td>
<td>5,478</td>
<td>7,095</td>
<td>3.2</td>
<td>9.16</td>
<td>78.88</td>
</tr>
<tr>
<td></td>
<td>Argument Retrieval</td>
<td>Misc.</td>
<td>ArguAna [67]</td>
<td>✓</td>
<td>Binary</td>
<td>14,085</td>
<td>6,666</td>
<td>6,666</td>
<td>5,416,568</td>
<td>1.2</td>
<td>8.13</td>
</tr>
<tr>
<td></td>
<td>Duplicate-Question Retrieval</td>
<td>StackEx. Quora</td>
<td>CQADupStack [25]</td>
<td>✓</td>
<td>Binary</td>
<td>13,145</td>
<td>5,000</td>
<td>5,000</td>
<td>457,199</td>
<td>1.4</td>
<td>9.35</td>
</tr>
<tr>
<td></td>
<td>Entity-Retrieval</td>
<td>Wikipedia</td>
<td>DBPedia [21]</td>
<td>✓</td>
<td>3-level</td>
<td>67</td>
<td>400</td>
<td>4,635,922</td>
<td>38.2</td>
<td>5.39</td>
<td>49.68</td>
</tr>
<tr>
<td></td>
<td>Citation-Prediction</td>
<td>Scientific</td>
<td>SCIDOCS [9]</td>
<td>✓</td>
<td>Binary</td>
<td>1,000</td>
<td>25,657</td>
<td>4.9</td>
<td>9.38</td>
<td>176.19</td>
<td></td>
</tr>
</tbody>
</table>

Thakur et al. 2021
## LoTTE: Long-Tail, Topic-stratified Evaluation

<table>
<thead>
<tr>
<th>Topic</th>
<th>Question Set</th>
<th>Dev</th>
<th>Test</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Writing</td>
<td>Search Forum</td>
<td>497/2003</td>
<td>277k</td>
<td>ESL, Linguistics,</td>
<td>1071/2000</td>
<td>200k English</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Worldbuilding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recreation</td>
<td>Search Forum</td>
<td>563/2002</td>
<td>263k</td>
<td>Sci-Fi, RPGs,</td>
<td>924/2002</td>
<td>167k Gaming,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Photography</td>
<td></td>
<td>Anime, Movies</td>
</tr>
<tr>
<td>Science</td>
<td>Search Forum</td>
<td>538/2013</td>
<td>344k</td>
<td>Chemistry,</td>
<td>617/2017</td>
<td>1.694M Math,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Statistics,</td>
<td></td>
<td>Physics, Biology</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Academia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technology</td>
<td>Search Forum</td>
<td>916/2003</td>
<td>1.276M</td>
<td>Web Apps,</td>
<td>596/2004</td>
<td>639k Apple,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ubuntu, SysAdmin</td>
<td></td>
<td>Android, UNIX,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Security</td>
</tr>
<tr>
<td>Lifestyle</td>
<td>Search Forum</td>
<td>496/2076</td>
<td>269k</td>
<td>DIY, Music,</td>
<td>661/2002</td>
<td>119k Cooking,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bicycles, Car</td>
<td></td>
<td>Sports, Travel</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Maintenance</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Topic-aligned dev-test pairings

Search queries are from GooAQ linked to StackExchange. Forum queries are from questions-like StackExchange titles

Santhanam et al. 2022b
XOR-TyDi

Information-seeking QA, OpenQA, and multilingual QA

XOR-TyDi v1.1 Leaderboard

Task 1: XOR-Retrieve

XOR-Retrieve is a cross-lingual retrieval task where a question is written in a target language (e.g., Japanese) and a system is required to retrieve English paragraphs that answer the question. The scores are macro-average over the 7 target languages. Although we see the effectiveness of blackbox systems (e.g., Google Translate), we encourage the community to use white-box systems so that all experimental details can be understood. The systems using external blackbox APIs are highlighted in gray and ranked in the table of "Systems using external APIs" for reference.

Metrics: $R@5kt$, $R@2kt$ (the recall by computing the fraction of the questions for which the minimal answer is contained in the top 5,000 / 2,000 tokens selected.)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>$R@5kt$</th>
<th>$R@2kt$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PrimeQA (DrDecr-large with PLAID + Colbert V2)</td>
<td>74.7</td>
<td>69.2</td>
</tr>
<tr>
<td></td>
<td>IBM Research AI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

https://nlp.cs.washington.edu/xorqa/
Other topics

1. There is a large literature on different techniques for sampling negatives.

2. Weak supervision can often create effective retrieval labels. For example, Khattab et al. (2021) say a passage is relevant in a QA context if it contains the answer as a substring anywhere in the passage.

3. Santhanam et al. (2022c) use Dynascores (Ma et al. 2021) to create unified leaderboards measuring diverse IR metrics, including cost, latency and performance. We will discuss Dynascores in detail later in the course.
Conclusion
NLU and IR are back together again, with profound implications for research and technology development!


Keshav Santhanam, Jon Saad-Falcon, Martin Franz, Omar Khattab, Avirup Sil, Radu Florian, Md Arafat Sultan, Salim Roukos, Matei Zaharia, and Christopher Potts. 2022c. *Moving beyond downstream task accuracy for information retrieval benchmarking*. Ms., Stanford University and IBM Research AI.