Query Automation for Systematic Reviews

Harry Scells
Leipzig University
https://scells.me
How do clinicians become informed about how to treat their patients?

How do governments and institutions make health policy decisions?
How do clinicians become informed about how to treat their patients?

Systematic Reviews

How do governments and institutions make health policy decisions?
Systematic Reviews
Overview

- **Guide** clinical decisions
- **Inform** practice and policy
- **Provide** evidence
Systematic review creation is hard!

1. Protocol Definition
2. Search Strategy Development
3. Study Abstract Screening
4. Study Full-Text Screening
5. Study Synthesis & Results Preparation
6. Dissemination of Systematic Review
Why is systematic review creation hard?

1. Protocol Definition
2. Search Strategy Development (~30 million studies)
3. Study Abstract Screening (Between ~10 thousand to ~1 million studies)
4. Study Full-Text Screening (~100 studies)
5. Study Synthesis & Results Preparation (~10 studies)
6. Dissemination of Systematic Review
Why is systematic review creation hard?

- Protocol Definition
- Search Strategy Development (~30 million studies)
- Study Abstract Screening (Between ~10 thousand to ~1 million studies)
- Study Full-Text Screening (~100 studies)
- Study Synthesis & Results Preparation (~10 studies)
- Dissemination of Systematic Review

Cost >250,000 EUR and takes >2 years
Why is systematic review creation hard?

- Protocol Definition
- Search Strategy Development
- Study Abstract Screening
- Study Full-Text Screening
- Study Synthesis & Results Preparation
- Dissemination of Systematic Review

- Cost >250,000 EUR
- and takes >2 years

~30 million studies
Between ~10 thousand to ~1 million studies
~100 studies
~10 studies
Why is systematic review creation hard?

Protocol Definition

Search Strategy Development

- Research focus
- ~30 million studies

Study Abstract Screening

- Between ~10 thousand to ~1 million studies

Study Full-Text Screening

- ~100 studies

Study Synthesis & Results Preparation

- ~10 studies

Cost >250,000 EUR and takes >2 years

Dissemination of Systematic Review
Why such little research on queries?

Why such little research on queries?
Why are Boolean queries used?

Reproducibility \(\rightarrow\) double check screening

Understandability \(\rightarrow\) control set size
Query Formulation

The automatic creation of complex queries for the task of systematic review literature search

- Content covered

- Further reading
How humans formulate queries

Overview

Conceptual method [Clark 2013] \(\rightarrow\) Human expertise

Objective method [Hausner et al. 2012] \(\rightarrow\) More algorithmic

Both methods \(\rightarrow\) Seed studies
Automating the conceptual method

Systematic review statement + seed studies

Galactomannan detection for invasive aspergillosis in immunocompromised patients

1. Query Logic Composition

(Galactomannan AND (invasive OR aspergillosis) AND patients)

2. Entity Extraction

(C0060961 AND (C0205281 AND C0276652) AND C0030705)

3. Entity Expansion

(((C0060961 OR C1523094) AND (C0205281 OR C1156244 OR C0683468 OR C1334278) AND (C0276652 OR C0343856 OR C0348989) AND (C0030705 OR C0870602))

4. Keyword Mapping

(((galactomannan[Title/Abstract] OR galactomannan[MeSH Terms]) AND (invasive[Title/Abstract] OR ... OR “invasive lesion”[Title/Abstract]) AND (aspergillosis[Title/Abstract] OR ... OR “Tonsillaire aspergillose”[Title/Abstract]) AND (patients[Title/Abstract] OR patients[MeSH Terms]))

Final Boolean query

Evaluate

Validation

Development

Unseen

High level concepts ➔ broaden search ➔ iterate until satisfied
Automating the conceptual method

Galactomannan detection for invasive aspergillosis in immunocompromised patients

1. Query Logic Composition

\[(\text{Galactomannan AND (invasive OR aspergillosis) AND patients})\]

2. Entity Extraction

\[(\text{C0060961 AND (C0205281 AND C0276652) AND C0030705})\]

3. Entity Expansion

\[(\text{((C0060961 OR C1523094) AND (C0205281 OR C1156244 OR C0683468 OR C1334278) AND (C0276652 OR C0343856 OR C0348989) AND (C0030705 OR C0870602))})\]

4. Keyword Mapping

\[\text{((galactomannan[Title/Abstract] OR galactomannan[MeSH Terms]) AND (invasive[Title/Abstract] OR ... OR “invasive lesion”[Title/Abstract]) AND (aspergillosis[Title/Abstract] OR ... OR “Tonsillaire aspergillose”[Title/Abstract]) AND (patients[Title/Abstract] OR patients[MeSH Terms]))}\]

Evaluate

POS tagger ➔ parse grammar & segment words into noun phrases
Automating the conceptual method

Systematic review statement + seed studies

Galactomannan detection for invasive aspergillosis in immunocompromised patients

1. Query Logic Composition

(Galactomannan AND (invasive OR aspergillosis) AND patients)

2. Entity Extraction

(C0060961 AND (C0205281 AND C0276652) AND C0030705)

3. Entity Expansion

(((C0060961 OR C1523094) AND (C0205281 OR C1156244 OR C0683468 OR C1334278) AND (C0276652 OR C0343856 OR C0348989) AND (C0030705 OR C0870602)))

4. Keyword Mapping

(((galactomannan[Title/Abstract] OR galactomannan[MeSH Terms]) AND (invasive[Title/Abstract] OR ... OR “invasive lesion”[Title/Abstract]) AND (aspergillosis[Title/Abstract] OR ... OR “Tonsillaire aspergillose”[Title/Abstract]) AND (patients[Title/Abstract] OR patients[MeSH Terms]))

Final Boolean query

Unseen

Evaluate

MetaMap \(\rightarrow\) extract CUIs from UMLS ontology
Automating the conceptual method

Systematic review statement + seed studies

Galactomannan detection for invasive aspergillosis in immunocompromised patients

Development ➔ Validation

1. Query Logic Composition

(Galactomannan AND (invasive OR aspergillosis) AND patients)

2. Entity Extraction

(C0060961 AND (C0205281 AND C0276652) AND C0030705)

3. Entity Expansion

(((C0060961 OR C1523094) AND (C0205281 OR C1156244 OR C0683468 OR C1334278) AND (C0276652 OR C0343856 OR C0348989) AND (C0030705 OR C0870602))

4. Keyword Mapping

(((galactomannan[Title/Abstract] OR galactomannan[MeSH Terms]) AND (invasive[Title/Abstract] OR ... OR “invasive lesion”[Title/Abstract]) AND (aspergillosis[Title/Abstract] OR ... OR “Tonsillaire aspergillose”[Title/Abstract]) AND (patients[Title/Abstract] OR patients[MeSH Terms])))

Final Boolean query ➔ Evaluate

Skipgram model ➔ broaden scope
Automating the conceptual method

1. Query Logic Composition
   
   (Galactomannan AND (invasive OR aspergillosis) AND patients)

2. Entity Extraction
   
   (C0060961 AND (C0205281 AND C0276652) AND C0030705)

3. Entity Expansion
   
   (((C0060961 OR C1523094) AND (C0205281 OR C1156244 OR C0683468 OR C1334278) AND (C0276652 OR C0343856 OR C0348989) AND (C0030705 OR C0870602))

4. Keyword Mapping
   
   (((galactomannan[Title/Abstract] OR galactomannan[MeSH Terms]) AND (invasive[Title/Abstract] OR ... OR “invasive lesion”[Title/Abstract]) AND (aspergillosis[Title/Abstract] OR ... OR “Tonsillaire aspergillose”[Title/Abstract]) AND (patients[Title/Abstract] OR patients[MeSH Terms]))

Map concepts (CUIs) to terms
Automating the objective method

Population set

seed studies

Split the Test set into Development, Validation, and Unseen

Development

Validation

Unseen

Analyse Development Set

Identify overrepresented population terms in the development set

Identify the most frequent MeSH terms

Add terms to the query

Final Boolean query

Evaluate

Find prominent terms from docs ➔ Add these terms to query
Automating the objective method

Population set

seed studies

Split the Test set into Development, Validation, and Unseen

Development

Analyse Development Set

Identify overrepresented population terms in the development set

Identify the most frequent MeSH terms

Unseen

Validate

Add terms to the query

Final Boolean query

Evaluate

Extract list of keywords from seed studies
Automating the objective method

Population set

seed studies

Split the Test set into Development, Validation, and Unseen

Analyse Development Set

Identify overrepresented population terms in the development set

Identify the most frequent MeSH terms

Add terms to the query

Final Boolean query

Evaluate

Rank documents using term frequency
Automating the objective method

Population set

seed studies

Split the Test set into Development, Validation, and Unseen

Development

Analyse Development Set

Identify overrepresented population terms in the development set

Validation

Identify the most frequent MeSH terms

Unseen

Add terms to the query

Final Boolean query

Evaluate

Add keywords from documents to query
# Conceptual versus objective results

**CLEF TAR** [Kanoulas et al. 2017, 2018]  
**Seed study collection** [Wang et al. 2022]

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual</td>
<td>0.0014</td>
<td>0.0027</td>
<td>0.6996</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0002</td>
<td>0.0005</td>
<td>0.9128</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual</td>
<td>0.0018</td>
<td>0.0036</td>
<td>0.4138</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0057</td>
<td>0.0113</td>
<td>0.5192</td>
</tr>
</tbody>
</table>

num. queries=72  num. queries=40
## Conceptual versus objective results

**CLEF TAR** [Kanoulas et al. 2017, 2018]

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.0207</td>
<td>0.0290</td>
<td>0.8317</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0014</td>
<td>0.0027</td>
<td>0.6996</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0002</td>
<td>0.0005</td>
<td>0.9128</td>
</tr>
</tbody>
</table>

Num. queries=72

**Seed study collection** [Wang et al. 2022]

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.0367</td>
<td>0.0651</td>
<td>0.7366</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0018</td>
<td>0.0036</td>
<td>0.4138</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0057</td>
<td>0.0113</td>
<td>0.5192</td>
</tr>
</tbody>
</table>

Num. queries=40
Conceptual versus objective results

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
<th></th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.0207</td>
<td>0.0290</td>
<td>0.8317</td>
<td>Objective</td>
<td>0.0367</td>
<td>0.0651</td>
<td>0.7366</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0014</td>
<td>0.0027</td>
<td>0.6996</td>
<td>Conceptual</td>
<td>0.0018</td>
<td>0.0036</td>
<td>0.4138</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0002</td>
<td>0.0005</td>
<td>0.9128</td>
<td>Objective</td>
<td>0.0057</td>
<td>0.0113</td>
<td>0.5192</td>
</tr>
</tbody>
</table>

num. queries=72  num. queries=40

Objective provides highest recall, but larger trade-offs than conceptual
## Conceptual versus objective results

**CLEF TAR** [Kanoulas et al. 2017, 2018]

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.0207</td>
<td>0.0290</td>
<td>0.8317</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0014</td>
<td>0.0027</td>
<td>0.6996</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0002</td>
<td>0.0005</td>
<td>0.9128</td>
</tr>
</tbody>
</table>

**Seed study collection** [Wang et al. 2022]

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.0367</td>
<td>0.0651</td>
<td>0.7366</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0018</td>
<td>0.0036</td>
<td>0.4138</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0057</td>
<td>0.0113</td>
<td>0.5192</td>
</tr>
</tbody>
</table>

| num. queries=72 | num. queries=40 |

Objective provides highest recall, but larger trade-offs than conceptual

Both methods are highly sensitive to initial seed studies
## Conceptual versus objective results

**CLEF TAR** [Kanoulas et al. 2017, 2018]

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.0207</td>
<td>0.0290</td>
<td>0.8317</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0014</td>
<td>0.0027</td>
<td>0.6996</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0002</td>
<td>0.0005</td>
<td>0.9128</td>
</tr>
</tbody>
</table>

**Seed study collection** [Wang et al. 2022]

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0367</td>
<td>0.0651</td>
<td>0.7366</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0018</td>
<td>0.0036</td>
<td>0.4138</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0057</td>
<td>0.0113</td>
<td>0.5192</td>
</tr>
</tbody>
</table>

num. queries=72

### Observations

- **Objective** provides highest recall, but larger trade-offs than conceptual.
- Both methods are highly sensitive to initial seed studies.
- Conceptual method better starting point for precision, and objective for recall.
Can ChatGPT formulate effective queries?

Hey ChatGPT, formulate a query for me!

Sure, no problem! ;-)

- Translation task: prompt $\Rightarrow$ query

- Two methods to formulate queries:
  - Unguided prompt *one shot to generate query*
  - Guided prompt *follows objective approach*

- Prevalence of Differentiated Thyroid Cancer in Autopsy Studies Over Six Decades: A Meta-Analysis
For a systematic review titled "{review_title}" , can you generate a systematic review Boolean query to find all included studies on PubMed for the review topic?

("Prevalence of Differentiated Thyroid Cancer" AND "Autopsy Studies" AND "Six Decades") AND (systematic review OR meta-analysis)
You are an information specialist who develops Boolean queries for systematic reviews. You have extensive experience developing highly effective queries for searching the medical literature. Your specialty is developing queries that retrieve as few irrelevant documents as possible and retrieve all relevant documents for your information need. Now you have your information need to conduct research on "{review_title}". Please construct a highly effective systematic review Boolean query that can best serve your information need.

("Differentiated Thyroid Cancer" OR "Papillary Thyroid Cancer" OR "Follicular Thyroid Cancer") AND "Autopsy Studies" AND "Prevalence" AND (("1960/01/01"[PDAT] : "2023/01/30"[PDAT]) OR ("1960"[PDAT]:"2023"[PDAT]))
You are an information specialist who develops Boolean queries for systematic reviews. You have extensive experience developing highly effective queries for searching the medical literature. Your specialty is developing queries that retrieve as few irrelevant documents as possible and retrieve all relevant documents for your information need. You are able to take an information need such as: "{example_review_title}" and generate valid pubmed queries such as: "{example_review_query}". Now you have the information need to conduct research on "{review_title}", please generate a highly effective systematic review Boolean query for the information need.

## Previous query formulation results

**CLEF TAR** [Kanoulas et al. 2017, 2018]

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.0207</td>
<td>0.0290</td>
<td>0.8317</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0014</td>
<td>0.0027</td>
<td>0.6996</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0002</td>
<td>0.0005</td>
<td>0.9128</td>
</tr>
</tbody>
</table>

**Seed study collection** [Wang et al. 2022]

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0367</td>
<td>0.0651</td>
<td>0.7366</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0018</td>
<td>0.0036</td>
<td>0.4138</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0057</td>
<td>0.0113</td>
<td>0.5192</td>
</tr>
</tbody>
</table>
## Unguided query formulation results

### CLEF TAR [Kanoulas et al. 2017, 2018]

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.0207</td>
<td>0.0290</td>
<td>0.8317</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0014</td>
<td>0.0027</td>
<td>0.6996</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0002</td>
<td>0.0005</td>
<td>0.9128</td>
</tr>
<tr>
<td>Simple</td>
<td>0.0543</td>
<td>0.0500</td>
<td>0.1293</td>
</tr>
<tr>
<td>Detailed</td>
<td>0.1166</td>
<td>0.0654</td>
<td>0.1310</td>
</tr>
<tr>
<td>Example</td>
<td>0.0752</td>
<td>0.0642</td>
<td>0.5035</td>
</tr>
</tbody>
</table>

### Seed study collection [Wang et al. 2022]

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0367</td>
<td>0.0651</td>
<td>0.7366</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0018</td>
<td>0.0036</td>
<td>0.4138</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0057</td>
<td>0.0113</td>
<td>0.5192</td>
</tr>
<tr>
<td>Simple</td>
<td>0.0501</td>
<td>0.0274</td>
<td>0.0528</td>
</tr>
<tr>
<td>Detailed</td>
<td>0.0983</td>
<td>0.0310</td>
<td>0.0394</td>
</tr>
<tr>
<td>Example</td>
<td>0.0283</td>
<td>0.0274</td>
<td>0.1290</td>
</tr>
</tbody>
</table>
### Unguided query formulation results

**CLEF TAR** [Kanoulas et al. 2017, 2018]

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.0207</td>
<td>0.0290</td>
<td>0.8317</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0014</td>
<td>0.0027</td>
<td>0.6996</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0002</td>
<td>0.0005</td>
<td>0.9128</td>
</tr>
<tr>
<td>Simple</td>
<td>0.0543</td>
<td>0.0500</td>
<td>0.1293</td>
</tr>
<tr>
<td>Detailed</td>
<td>0.1166</td>
<td>0.0654</td>
<td>0.1310</td>
</tr>
<tr>
<td>Example</td>
<td>0.0752</td>
<td>0.0642</td>
<td>0.5035</td>
</tr>
</tbody>
</table>

**Seed study collection** [Wang et al. 2022]

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0367</td>
<td>0.0651</td>
<td>0.7366</td>
</tr>
<tr>
<td></td>
<td>0.0018</td>
<td>0.0036</td>
<td>0.4138</td>
</tr>
<tr>
<td></td>
<td>0.0057</td>
<td>0.0113</td>
<td>0.5192</td>
</tr>
<tr>
<td></td>
<td>0.0501</td>
<td>0.0274</td>
<td>0.0528</td>
</tr>
<tr>
<td></td>
<td>0.0983</td>
<td>0.0310</td>
<td>0.0394</td>
</tr>
<tr>
<td></td>
<td>0.0283</td>
<td>0.0274</td>
<td>0.1290</td>
</tr>
</tbody>
</table>

Collection contains **seed studies**  
→ Use ChatGPT for objective method
Guided prompt examples

Step 1

Follow my instructions precisely to develop a highly effective Boolean query for a medical systematic review literature search. Do not explain or elaborate. First, Given the following statement and text from a relevant study, identify **50 terms or phrases that are relevant**. The terms you identify should be used to retrieve more relevant studies, so be careful that the terms you choose are not too broad. You are not allowed to have duplicates in your list. Statement: {statement}. Text: {text}

1. Differentiated thyroid cancer
2. Prevalence
3. Autopsy studies
4. Meta-analysis
5. Occult carcinomas
...
50. Parameters.
For each item in the list you created in Step 1, **classify it into three categories**: terms relating to health conditions (A), terms relating to a treatment (B), terms relating to types of study design (C). When an item does not fit one of these categories, mark it as (N/A). Each item needs to be categorised into (A), (B), (C), or (N/A).

1. **(A) Differentiated thyroid cancer**
2. (N/A) Prevalence
3. (C) Autopsy studies
4. (C) Meta-analysis
5. **(A) Occult carcinomas**
...
50. (N/A) Parameters.
Step 3

Using the categorised list you created in step 2, create a Boolean query that can be submitted to PubMed which groups together items from each category. For example: ((itemA1 OR itemA2 or itemA3) AND (itemB1 OR itemB2 OR itemB3) AND (itemC1 OR itemC2 OR itemC3))

Use your expert knowledge to **refine the query**, making it retrieve as many relevant documents as possible while minimising the total number of documents retrieved. Also add relevant MeSH terms into the query where necessary, e.g., MeSHTerm[MeSH]. Retain the general structure of the query, however, with each main clause of the query corresponding to a PICO element. The final query still needs to be executable on PubMed, so it should be a valid query.

### Unguided query formulation results

Seed study collection [Wang et al. 2022]

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.0367</td>
<td>0.0651</td>
<td>0.7366</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0018</td>
<td>0.0036</td>
<td>0.4138</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0057</td>
<td>0.0113</td>
<td>0.5192</td>
</tr>
<tr>
<td>Simple</td>
<td>0.0501</td>
<td>0.0274</td>
<td>0.0528</td>
</tr>
<tr>
<td>Detailed</td>
<td>0.0983</td>
<td>0.0310</td>
<td>0.0394</td>
</tr>
<tr>
<td>Example</td>
<td>0.0283</td>
<td>0.0274</td>
<td>0.1290</td>
</tr>
</tbody>
</table>
Guided query formulation results
Seed study collection [Wang et al. 2022]

<table>
<thead>
<tr>
<th>Type</th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.0367</td>
<td>0.0651</td>
<td>0.7366</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0018</td>
<td>0.0036</td>
<td>0.4138</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0057</td>
<td>0.0113</td>
<td>0.5192</td>
</tr>
<tr>
<td>Simple</td>
<td>0.0501</td>
<td>0.0274</td>
<td>0.0528</td>
</tr>
<tr>
<td>Detailed</td>
<td>0.0983</td>
<td>0.0310</td>
<td>0.0394</td>
</tr>
<tr>
<td>Example</td>
<td>0.0283</td>
<td>0.0274</td>
<td>0.1290</td>
</tr>
<tr>
<td>Guided</td>
<td>0.0993</td>
<td>0.0492</td>
<td>0.5171</td>
</tr>
</tbody>
</table>
Guided query formulation results
Seed study collection [Wang et al. 2022]

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.0367</td>
<td>0.0651</td>
<td>0.7366</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0018</td>
<td>0.0036</td>
<td>0.4138</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0057</td>
<td>0.0113</td>
<td>0.5192</td>
</tr>
<tr>
<td>Simple</td>
<td>0.0501</td>
<td>0.0274</td>
<td>0.0528</td>
</tr>
<tr>
<td>Detailed</td>
<td>0.0983</td>
<td>0.0310</td>
<td>0.0394</td>
</tr>
<tr>
<td>Example</td>
<td>0.0283</td>
<td>0.0274</td>
<td>0.1290</td>
</tr>
<tr>
<td>Guided</td>
<td>0.0993</td>
<td>0.0492</td>
<td>0.5171</td>
</tr>
</tbody>
</table>

ChatGPT is more effective than automatic conceptual and objective methods
Guided query formulation results
Seed study collection [Wang et al. 2022]

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.0367</td>
<td>0.0651</td>
<td>0.7366</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.0018</td>
<td>0.0036</td>
<td>0.4138</td>
</tr>
<tr>
<td>Objective</td>
<td>0.0057</td>
<td>0.0113</td>
<td>0.5192</td>
</tr>
<tr>
<td>Simple</td>
<td>0.0501</td>
<td>0.0274</td>
<td>0.0528</td>
</tr>
<tr>
<td>Detailed</td>
<td>0.0983</td>
<td>0.0310</td>
<td>0.0394</td>
</tr>
<tr>
<td>Example</td>
<td>0.0283</td>
<td>0.0274</td>
<td>0.1290</td>
</tr>
<tr>
<td>Guided</td>
<td>0.0993</td>
<td>0.0492</td>
<td>0.5171</td>
</tr>
</tbody>
</table>

ChatGPT is more effective than automatic conceptual and objective methods

ChatGPT is highly dependent on prompt and prone to hallucination
Query Augmentation

The modification or extension of complex queries in order to improve their effectiveness at the task of systematic review literature search

- Content covered

- Further reading
Smooth operators

Intuition

Boolean AND

Boolean OR

Boolean NOT
Smooth operators

Intuition

- Boolean AND
- Boolean OR
- Boolean NOT

Smoothed AND Equivalent
Smooth operators

Intuition

Boolean AND

Boolean OR

Boolean NOT

Smoothed AND Equivalent

Smoothed OR Equivalent

Smoothed NOT Equivalent
Smooth operators

Intuition

Boolean AND

Boolean OR

Boolean NOT

Smoothed AND Equivalent

Smoothed OR Equivalent

Smoothed NOT Equivalent
Smooth operators

Intuitive example

\[
\text{AND} \quad q_1 \quad q_2 \quad q_3 \quad \text{query}
\]
Smooth operators

Intuitive example

\[
\text{AND} \quad q_1 \quad q_2 \quad q_3
\]

\[
d_1 \quad d_2 \quad d_3 \quad \text{documents}
\]
Smooth operators

Intuitive example

AND

query

q₁ q₂ q₃

d₁ d₂ d₃

documents
Smooth operators

Intuitive example

AND

query

$q_1$  $q_2$  $q_3$

$d_1$  $d_2$  $d_3$

documents
Smooth operators

Intuitive example

AND

$q_1$  $q_2$  $q_3$

$d_1$  $d_2$  $d_3$

query

documents
Smoothing result sets

\[ P(d|q) \rightarrow \text{extent to which } d \text{ should belong to } q \]

\[ P(d|q) = \frac{P(d)P(q|d)}{P(q)} \]
Theory
Smoothing result sets

\[
P(d|q_1, \ldots, q_k) = \frac{P(d) \prod P(q_i|d)}{P(d) \prod P(q_i|d) + P(\bar{d}) \prod P(q_i|\bar{d})}
\]
Theory
Smoothing result sets

\[
P(d|q_1, \ldots, q_k) = \frac{P(d) \prod P(q_i|d)}{P(d) \prod P(q_i|d) + P(\bar{d}) \prod P(q_i|\bar{d})}
\]

Leaves two estimations:

- \( P(d) \rightarrow \text{Probability of a document} \)
- \( P(q_i|d) \rightarrow \text{Probability of a query given a document} \)
Theory

Smoothing result sets

\[ P(d) \rightarrow \text{ratio of queries that retrieve } d \]

\[ P(d) = \frac{|\{\forall q_i \in q : d \in D_{q_i}\}|}{|q|} \]
Theory
Smoothing result sets

\[ P(d) \rightarrow \text{ratio of queries that retrieve } d \]

\[ P(d) = \frac{\left| \{ \forall q_i \in q : d \in D_{q_i} \} \right|}{|q|} \]

\[ P(q_i|d) \rightarrow \text{relevance between } q_i \text{ and } d \]

\[ P(q_i|d) = 1 - \frac{pos(q_i, d)}{|D_{q_i}|} \]
Theory
Implementing smooth operators

AND@1.0

$q_1$ $q_2$ $q_3$

query

$d_1$ $d_2$ $d_3$
documents
Theory
Implementing smooth operators

\[ P(d_3|q_1, q_2, q_3) = 0.95 \]
Theory
Implementing smooth operators

\[ P(d_3|q_1, q_2, q_3) = 0.95 \]
## Results

### Seed study collection [Wang et al. 2022]

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>$F_1$</th>
<th>Precision</th>
<th>nDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean operators</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>-</td>
</tr>
<tr>
<td>BM25 Title</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.0972</td>
</tr>
<tr>
<td>smooth Boolean equivalents</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.2486</td>
</tr>
</tbody>
</table>
### Results

**Seed study collection** [Wang et al. 2022]

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>$F_1$</th>
<th>Precision</th>
<th>nDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean operators</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>-</td>
</tr>
<tr>
<td>BM25 Title</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.0972</td>
</tr>
<tr>
<td>smooth Boolean</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.2486</td>
</tr>
<tr>
<td>equivalents</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Only using smooth operators for ranking is already considerably better than using BM25**
## Results

### Seed study collection [Wang et al. 2022]

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>$F_1$</th>
<th>Precision</th>
<th>nDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean operators</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>-</td>
</tr>
<tr>
<td>BM25 Title</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.0972</td>
</tr>
<tr>
<td>smooth Boolean</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.2486</td>
</tr>
<tr>
<td>equivalents AND@0.99</td>
<td>0.7206</td>
<td>0.0038</td>
<td>0.0019</td>
<td>0.2331</td>
</tr>
<tr>
<td>AND@0.9</td>
<td>0.7658</td>
<td>0.0003</td>
<td>0.0002</td>
<td>0.2180</td>
</tr>
</tbody>
</table>
### Results

**Seed study collection** [Wang et al. 2022]

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>$F_1$</th>
<th>Precision</th>
<th>nDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean operators</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>-</td>
</tr>
<tr>
<td>BM25 Title</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.0972</td>
</tr>
<tr>
<td>smooth Boolean equivalents</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.2486</td>
</tr>
<tr>
<td>AND@0.99</td>
<td>0.7206</td>
<td>0.0038</td>
<td>0.0019</td>
<td>0.2331</td>
</tr>
<tr>
<td>AND@0.9</td>
<td>0.7658</td>
<td>0.0003</td>
<td>0.0002</td>
<td>0.2180</td>
</tr>
</tbody>
</table>

Smoothing AND operators increases recall at the cost of precision and ranking effectiveness.
## Results

Seed study collection [Wang et al. 2022]

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>$F_1$</th>
<th>Precision</th>
<th>nDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean operators</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>-</td>
</tr>
<tr>
<td>BM25 Title</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.0972</td>
</tr>
<tr>
<td>smooth Boolean equivalents</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.2486</td>
</tr>
<tr>
<td>AND@0.99</td>
<td>0.7206</td>
<td>0.0038</td>
<td>0.0019</td>
<td>0.2331</td>
</tr>
<tr>
<td>AND@0.9</td>
<td>0.7658</td>
<td>0.0003</td>
<td>0.0002</td>
<td>0.2180</td>
</tr>
<tr>
<td>OR@0.01</td>
<td>0.3078</td>
<td>0.0582</td>
<td>0.0366</td>
<td>0.1604</td>
</tr>
<tr>
<td>OR@0.1</td>
<td>0.0612</td>
<td>0.0417</td>
<td>0.0486</td>
<td>0.0620</td>
</tr>
</tbody>
</table>
## Results

**Seed study collection [Wang et al. 2022]**

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>$F_1$</th>
<th>Precision</th>
<th>nDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean operators</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>-</td>
</tr>
<tr>
<td>BM25 Title</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.0972</td>
</tr>
<tr>
<td>smooth Boolean</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.2486</td>
</tr>
<tr>
<td>equivalents AND@0.99</td>
<td>0.7206</td>
<td>0.0038</td>
<td>0.0019</td>
<td>0.2331</td>
</tr>
<tr>
<td>AND@0.9</td>
<td>0.7658</td>
<td>0.0003</td>
<td>0.0002</td>
<td>0.2180</td>
</tr>
<tr>
<td>OR@0.01</td>
<td>0.3078</td>
<td>0.0582</td>
<td>0.0366</td>
<td>0.1604</td>
</tr>
<tr>
<td>OR@0.1</td>
<td>0.0612</td>
<td>0.0417</td>
<td>0.0486</td>
<td>0.0620</td>
</tr>
</tbody>
</table>

Smoothing OR operators increases precision at the cost of recall and ranking effectiveness.
## Results

**Seed study collection [Wang et al. 2022]**

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>$F_1$</th>
<th>Precision</th>
<th>nDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean operators</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>-</td>
</tr>
<tr>
<td>BM25 Title</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.0972</td>
</tr>
<tr>
<td>smooth Boolean equivalents</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.2486</td>
</tr>
<tr>
<td>AND@0.99</td>
<td>0.7206</td>
<td>0.0038</td>
<td>0.0019</td>
<td>0.2331</td>
</tr>
<tr>
<td>AND@0.9</td>
<td>0.7658</td>
<td>0.0003</td>
<td>0.0002</td>
<td>0.2180</td>
</tr>
<tr>
<td>OR@0.01</td>
<td>0.3078</td>
<td>0.0582</td>
<td>0.0366</td>
<td>0.1604</td>
</tr>
<tr>
<td>OR@0.1</td>
<td>0.0612</td>
<td>0.0417</td>
<td>0.0486</td>
<td>0.0620</td>
</tr>
<tr>
<td>Predictor</td>
<td>0.3876</td>
<td>0.0566</td>
<td>0.0364</td>
<td>0.1651</td>
</tr>
</tbody>
</table>
### Results

#### Seed study collection [Wang et al. 2022]

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>$F_1$</th>
<th>Precision</th>
<th>nDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean operators</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>-</td>
</tr>
<tr>
<td>BM25 Title</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.0972</td>
</tr>
<tr>
<td>smooth Boolean equivalents</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.2486</td>
</tr>
<tr>
<td>AND@0.99</td>
<td>0.7206</td>
<td>0.0038</td>
<td>0.0019</td>
<td>0.2331</td>
</tr>
<tr>
<td>AND@0.9</td>
<td>0.7658</td>
<td>0.0003</td>
<td>0.0002</td>
<td>0.2180</td>
</tr>
<tr>
<td>OR@0.01</td>
<td>0.3078</td>
<td>0.0582</td>
<td>0.0366</td>
<td>0.1604</td>
</tr>
<tr>
<td>OR@0.1</td>
<td>0.0612</td>
<td>0.0417</td>
<td>0.0486</td>
<td>0.0620</td>
</tr>
<tr>
<td>Predictor</td>
<td>0.3876</td>
<td>0.0566</td>
<td>0.0364</td>
<td>0.1651</td>
</tr>
</tbody>
</table>

**Predicting the smoothness using features struck middleground between smooth OR and AND**
## Results

Seed study collection [Wang et al. 2022]

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>$F_1$</th>
<th>Precision</th>
<th>nDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean operators</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>-</td>
</tr>
<tr>
<td>BM25 Title</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.0972</td>
</tr>
<tr>
<td>smooth Boolean equivalents</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.2486</td>
</tr>
<tr>
<td>AND@0.99</td>
<td>0.7206</td>
<td>0.0038</td>
<td>0.0019</td>
<td>0.2331</td>
</tr>
<tr>
<td>AND@0.9</td>
<td>0.7658</td>
<td>0.0003</td>
<td>0.0002</td>
<td>0.2180</td>
</tr>
<tr>
<td>OR@0.01</td>
<td>0.3078</td>
<td>0.0582</td>
<td>0.0366</td>
<td>0.1604</td>
</tr>
<tr>
<td>OR@0.1</td>
<td>0.0612</td>
<td>0.0417</td>
<td>0.0486</td>
<td>0.0620</td>
</tr>
<tr>
<td>Predictor</td>
<td>0.3876</td>
<td>0.0566</td>
<td>0.0364</td>
<td>0.1651</td>
</tr>
<tr>
<td>Oracle</td>
<td>0.7437</td>
<td>0.0769</td>
<td>0.0440</td>
<td>0.2885</td>
</tr>
</tbody>
</table>
## Results

Seed study collection [Wang et al. 2022]

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>$F_1$</th>
<th>Precision</th>
<th>nDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean operators</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>-</td>
</tr>
<tr>
<td>BM25 Title</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.0972</td>
</tr>
<tr>
<td>smooth Boolean equivalents</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.2486</td>
</tr>
<tr>
<td>AND@0.99</td>
<td>0.7206</td>
<td>0.0038</td>
<td>0.0019</td>
<td>0.2331</td>
</tr>
<tr>
<td>AND@0.9</td>
<td>0.7658</td>
<td>0.0003</td>
<td>0.0002</td>
<td>0.2180</td>
</tr>
<tr>
<td>OR@0.01</td>
<td>0.3078</td>
<td>0.0582</td>
<td>0.0366</td>
<td>0.1604</td>
</tr>
<tr>
<td>OR@0.1</td>
<td>0.0612</td>
<td>0.0417</td>
<td>0.0486</td>
<td>0.0620</td>
</tr>
<tr>
<td>Predictor</td>
<td>0.3876</td>
<td>0.0566</td>
<td>0.0364</td>
<td>0.1651</td>
</tr>
<tr>
<td>Oracle</td>
<td>0.7437</td>
<td>0.0769</td>
<td>0.0440</td>
<td>0.2885</td>
</tr>
</tbody>
</table>

However: using ground truth information, more effective queries are possible.
## Results

**Seed study collection** [Wang et al. 2022]

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>$F_1$</th>
<th>Precision</th>
<th>nDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean operators</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>-</td>
</tr>
<tr>
<td>BM25 Title</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.0972</td>
</tr>
<tr>
<td>smooth Boolean equivalents</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.2486</td>
</tr>
<tr>
<td>AND@0.99</td>
<td>0.7206</td>
<td>0.0038</td>
<td>0.0019</td>
<td>0.2331</td>
</tr>
<tr>
<td>AND@0.9</td>
<td>0.7658</td>
<td>0.0003</td>
<td>0.0002</td>
<td>0.2180</td>
</tr>
<tr>
<td>OR@0.01</td>
<td>0.3078</td>
<td>0.0582</td>
<td>0.0366</td>
<td>0.1604</td>
</tr>
<tr>
<td>OR@0.1</td>
<td>0.0612</td>
<td>0.0417</td>
<td>0.0486</td>
<td>0.0620</td>
</tr>
<tr>
<td>Predictor</td>
<td>0.3876</td>
<td>0.0566</td>
<td>0.0364</td>
<td>0.1651</td>
</tr>
<tr>
<td>Oracle</td>
<td>0.7437</td>
<td>0.0769</td>
<td>0.0440</td>
<td>0.2885</td>
</tr>
<tr>
<td>PubmedBERT</td>
<td>0.7148</td>
<td>0.0643</td>
<td>0.0363</td>
<td>0.2252</td>
</tr>
<tr>
<td>BERT</td>
<td>0.7149</td>
<td>0.0644</td>
<td>0.0363</td>
<td>0.2447</td>
</tr>
<tr>
<td>DistilBERT</td>
<td>0.7118</td>
<td>0.0641</td>
<td>0.0362</td>
<td>0.2449</td>
</tr>
</tbody>
</table>
## Results

**Seed study collection [Wang et al. 2022]**

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>$F_1$</th>
<th>Precision</th>
<th>nDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean operators</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>-</td>
</tr>
<tr>
<td>BM25 Title</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.0972</td>
</tr>
<tr>
<td>smooth Boolean equivalents</td>
<td>0.7149</td>
<td>0.0642</td>
<td>0.0362</td>
<td>0.2486</td>
</tr>
<tr>
<td>AND@0.99</td>
<td>0.7206</td>
<td>0.0038</td>
<td>0.0019</td>
<td>0.2331</td>
</tr>
<tr>
<td>AND@0.9</td>
<td>0.7658</td>
<td>0.0003</td>
<td>0.0002</td>
<td>0.2180</td>
</tr>
<tr>
<td>OR@0.01</td>
<td>0.3078</td>
<td>0.0582</td>
<td>0.0366</td>
<td>0.1604</td>
</tr>
<tr>
<td>OR@0.1</td>
<td>0.0612</td>
<td>0.0417</td>
<td>0.0486</td>
<td>0.0620</td>
</tr>
<tr>
<td>Predictor</td>
<td>0.3876</td>
<td>0.0566</td>
<td>0.0364</td>
<td>0.1651</td>
</tr>
<tr>
<td>Oracle</td>
<td>0.7437</td>
<td>0.0769</td>
<td>0.0440</td>
<td>0.2885</td>
</tr>
<tr>
<td>PubmedBERT</td>
<td>0.7148</td>
<td>0.0643</td>
<td>0.0363</td>
<td>0.2252</td>
</tr>
<tr>
<td>BERT</td>
<td>0.7149</td>
<td>0.0644</td>
<td>0.0363</td>
<td>0.2447</td>
</tr>
<tr>
<td>DistilBERT</td>
<td>0.7118</td>
<td>0.0641</td>
<td>0.0362</td>
<td>0.2449</td>
</tr>
</tbody>
</table>

More advanced neural ranking models did not have any effect on ranking effectiveness.
## Results

**CLEF TAR** [Kanoulas et al. 2018]

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>$F_1$</th>
<th>Precision</th>
<th>nDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean operators</td>
<td>0.8344</td>
<td>0.0385</td>
<td>0.0204</td>
<td>-</td>
</tr>
<tr>
<td>BM25 Title</td>
<td>0.8344</td>
<td>0.0385</td>
<td>0.0204</td>
<td>0.0232</td>
</tr>
<tr>
<td>smooth Boolean equivalents</td>
<td>0.8344</td>
<td>0.0385</td>
<td>0.0204</td>
<td>0.1995</td>
</tr>
<tr>
<td>Predictor</td>
<td>0.6205</td>
<td>0.0372</td>
<td>0.0206</td>
<td>0.1698</td>
</tr>
<tr>
<td>Oracle</td>
<td>0.8487</td>
<td>0.0397</td>
<td>0.0211</td>
<td>0.2125</td>
</tr>
<tr>
<td>BERT</td>
<td>0.8344</td>
<td>0.0385</td>
<td>0.0204</td>
<td>0.2191</td>
</tr>
<tr>
<td>ECNU_RUN1</td>
<td>0.5147</td>
<td>0.0806</td>
<td>0.0490</td>
<td>0.2440</td>
</tr>
<tr>
<td>ECNU_RUN2</td>
<td>0.3831</td>
<td>0.0823</td>
<td>0.0539</td>
<td>0.1368</td>
</tr>
<tr>
<td>ECNU_RUN3</td>
<td>0.5147</td>
<td>0.0806</td>
<td>0.0490</td>
<td>0.2438</td>
</tr>
<tr>
<td>sheffield-bm25</td>
<td>0.4525</td>
<td>0.0180</td>
<td>0.0095</td>
<td>0.1197</td>
</tr>
<tr>
<td>sheffield-boolean</td>
<td>0.3048</td>
<td>0.0116</td>
<td>0.0061</td>
<td>0.0562</td>
</tr>
<tr>
<td>sheffield-tfidf</td>
<td>0.2572</td>
<td>0.0112</td>
<td>0.0059</td>
<td>0.0154</td>
</tr>
</tbody>
</table>

[Wu et al. 2018, Alharbi et al. 2018]
Query Automation

Formulation

Refinement

Tools

Augmentation
Tools

Software to support systematic review and information retrieval practitioners in undertaking query automation

- Content covered

- Further reading
Domain-specific search has high barrier to entry

- Slow/limited search APIs
- Specific indexing and document processing
- Complex query languages
Domain-specific search has high barrier to entry

- Slow/limited search APIs
- Specific indexing and document processing
- Complex query languages

```python
from pybool_ir.experiments.collections import load_collection
from pybool_ir.experiments.retrieval import RetrievalExperiment
from ir_measures import *
import ir_measures

# Automatically downloads, then loads this collection.
col = load_collection("ielab/sysrev-seed-collection")

# Point the experiment to your index, your collection.
with RetrievalExperiment(indexer=PubmedIndexer("./pubmed"),
                          collection=col) as experiment:
    # Get the run of the experiment.
    # This automatically executes the queries.
    run = experiment.run

# Evaluate the run using ir_measures.
ir_measures.calc_aggregate([SetP, SetR, SetF], col.qrels, run)
```
Domain-specific search has high barrier to entry

- Slow/limited search APIs
- Specific indexing and document processing
- Complex query languages

```python
from pybool_ir.experiments.collections import load_collection
from pybool_ir.experiments.retrieval import RetrievalExperiment
from ir_measures import *
import ir_measures

# Automatically downloads, then loads this collection.
col = load_collection("ielab/sysrev-seed-collection")

# Point the experiment to your index, your collection.
with RetrievalExperiment(indexer=PubmedIndexer("./pubmed"),
                          collection=col) as experiment:
    # Get the run of the experiment.
    # This automatically executes the queries.
    run = experiment.run

# Evaluate the run using ir_measures.
ir_measures.calc_aggregate([SetP, SetR, SetF], col.qrels, run)
```

![Graph showing precision and recall for Pubmed and pybool_ir]
Domain-specific search has high barrier to entry

- Slow/limited search APIs
- Specific indexing and document processing
- Complex query languages

```python
from pybool.ir.experiments.collections import load_collection
from pybool.ir.experiments.retrieval import RetrievalExperiment
from ir_measures import *
import ir_measures

# Automatically downloads, then loads this collection.
collection = load_collection("ieliab/sysrev-seed-collection")

# Point the experiment to your index, your collection.
with RetrievalExperiment(indexer=PubmedIndexer("./pubmed"),
collection=collection) as experiment:
    # Get the run of the experiment.
    # This automatically executes the queries.
    run = experiment.run

# Evaluate the run using ir_measures.
ir_measures.calc_aggregate([SetP, SetR, SetF], col.qrels, run)
```

- Domain-specific indexing
  - `ir_datasets`
  - Arbitrary JSON
- Extend query syntaxes
  - Smooth operators
  - Faster demo prototyping
- Compatibility with pyserini
  - index ➔ search
Domain-specific search has high barrier to entry

- Slow/limited search APIs
- Specific indexing and document processing
- Complex query languages

```python
from pybool_ir.experiments.collections import load_collection
from pybool_ir.experiments.retrieval import RetrievalExperiment
from ir_measures import *
import ir_measures

# Automatically downloads, then loads this collection.
col = load_collection("ielab/sysrev-seed-collection")

# Point the experiment to your index, your collection.
with RetrievalExperiment(indexer=PubmedIndexer("./pubmed"),
                          collection=col) as experiment:
    # Get the run of the experiment.
    # This automatically executes the queries.
    run = experiment.run

# Evaluate the run using ir_measures.
ir_measures.calc_aggregate([SetP, SetR, SetF], col.qrels, run)
```

- Domain-specific indexing
  - ir_datasets
  - Arbitrary JSON
- Extend query syntaxes
  - Smooth operators
  - Faster demo prototyping
- Compatibility with pyserini
  - index ➔ search
Domain-specific search has high barrier to entry

- Slow/limited search APIs
- Specific indexing and document processing
- Complex query languages

```python
from pybool_ir.experiments.collections import load_collection
from pybool_ir.experiments.retrieval import RetrievalExperiment
from ir_measures import *
import ir_measures

# Automatically downloads, then loads this collection.
col = load_collection("ielab/sysrev-seed-collection")

# Point the experiment to your index, your collection.
with RetrievalExperiment(indexer=PubmedIndexer("./pubmed"),
collection=col) as experiment:
    # Get the run of the experiment.
    # This automatically executes the queries.
    run = experiment.run

# Evaluate the run using ir_measures.
ir_measures.calc_aggregate([SetP, SetR, SetF], col.qrels, run)
```

- Domain-specific indexing
  - `ir_datasets`
  - Arbitrary JSON
- Extend query syntaxes
  - Smooth operators
  - Faster demo prototyping
- Compatibility with pyserini
  - index ➔ search
Query Automation

Formulation

Refinement

Test Collections

Screening Prioritisation

Tools

Augmentation
Outlook

- **Money** ➔ Can cost upwards of 250,000 Euros
- **Time** ➔ Can take over a year

[McGowan and Sampson, 2005]
Outlook

- **Money** ➜ Can cost upwards of 250,000 Euros
- **Time** ➜ Can take over a year

[McGowan and Sampson, 2005]

**Does hydrocloroquin treat COVID-19?**

- Priority for decision making?
- Level of uncertainty in the literature?
- Frequency of new studies is high?

**Should everyone wear a mask during the COVID-19 pandemic?**
Conclusion

Next steps

- Evaluating generative IR
  Is information relevant/correct/readable?
- Making query development easier
  Formulation, refinement, augmentation
- Furthering tools to enable research
  Reproduction with pybool_ir

Envisioned outcomes

- Faster and less expensive systematic reviews
- Fully automated evidence synthesis
- Tools for librarians and researchers to automate evidence creation

Stay in touch

- @hscells
- https://scells.me
- harry.scells@uni-leipzig.de