CODEC: Complex Document and Entity Collection
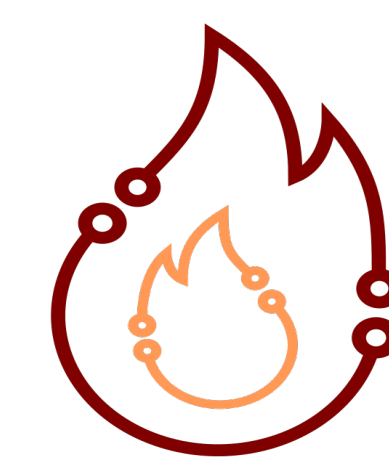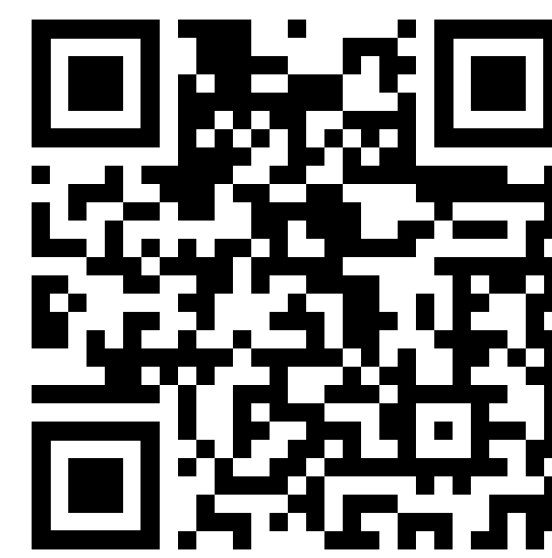
University of Glasgow

Iain Mackie, Paul Owoicho, Carlos Gemmell, Sophie Fischer, Sean MacAvaney, and Jeffrey Dalton

GRILL Lab

CODEC is a new resource for studying multifaceted complex information needs with entity-centric retrieval approaches.

Task: Adhoc document ranking and entity ranking

Contributions:
- New criteria to define complex topics
- 42 essay-style topics across history, economics and politics curated by domain experts
- 9.2 query reformulations per topic to explore multiple information needs (total 387 queries)
- New corpus of 730k documents across 24 websites
- 17,509 aligned judgments over documents and entities
- Entity links grounding documents to Wikipedia KB

Findings:
- Sparse and dense multi-stage retrieval methods perform poorly on complex topics
- Positive correlation between relevant documents and entities
  - ➤ Entity-centric methods improve document ranking (MAP+3%, Recall+7%)
- Exploration of complex topics requires multiple query reformulations covering different dimensions of the topic
  - ➤ Single and aggregate query reformulations improve document and entity ranking (MAP+3%, Recall+8%)

CODEC Resource Overview

Domain Experts

History
- History Teacher
- Historian

Economics
- FX Trader
- Accountant
- Investment Banker

Politics
- Political Scientist
- Politician

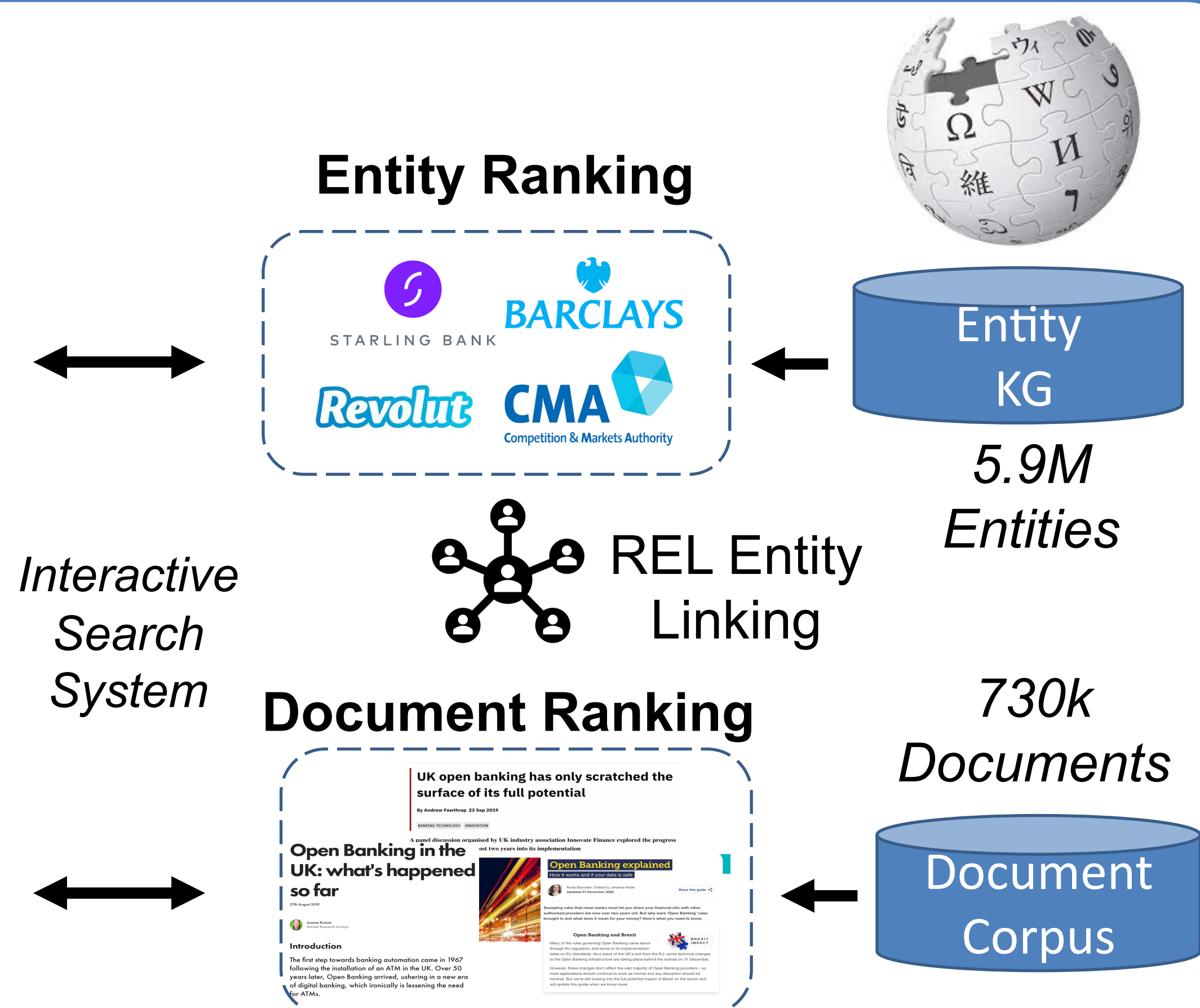**Economics-1: Open Banking**
☑ Entity-centric
☑ Complex topic

Annotator

300h of annotation

Judgment	Document Ranking	Entity Ranking
0	2,353	7,053
1	2,210	2,241
2	1,207	1,252
3	416	777
TOTAL	6,186	11,323

Manual Query Reformulation

United Kingdom open banking regulation
United Kingdom banking marketplace
UK Challenger banks market share
United kingdom open banking products
UK Challenger banks
UK Challenger banks (Atom Bank, Revolut, Starling, N26, and Tide)
United Kingdom new open banking products
PSD2 UK challenger banks
UK fintech startup funding
UK Open Banking impact on traditional banks (Barclays, HSBC, Lloyds Banking Group, and NatWest Group)

Interactive Search System

Entity Ranking

STARLING BANK BARCLAYS Revolut CMA

Entity KG

5.9M Entities

REL Entity Linking

Document Ranking

730k Documents

Document Corpus

Complex Topic Criteria

- Open-ended essay-style
- Natural language query
- Multiple points of view
- Concern multiple entities
- Requires domain knowledge

42 Essay-Style Topics

How has the UK's Open Banking Regulation benefited challenger banks?

Domain: Economics

Narrative: UK's Open Banking regulation, which has parallels to the EU's second payment service directive (PSD2), went live in January 2018. This piece of legislation …

Challenger banks are small, recently created retail banks that compete directly with the longer-established banks in the UK. Specifically, …

Challenger banks have used open banking to develop new products or capture market share ….
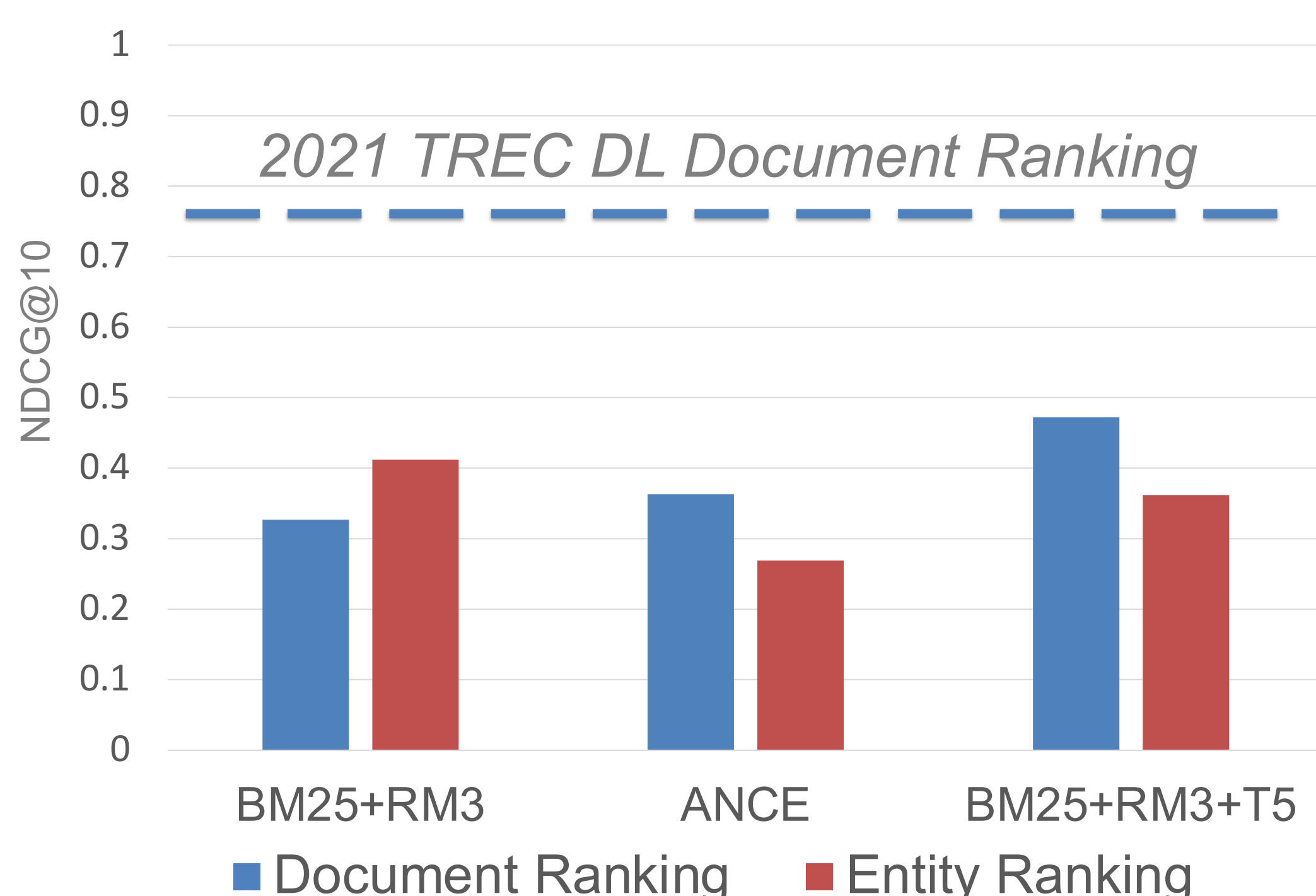
~13 words per query

~143 words per narrative

400+ judgments per topic

24 Document Sources

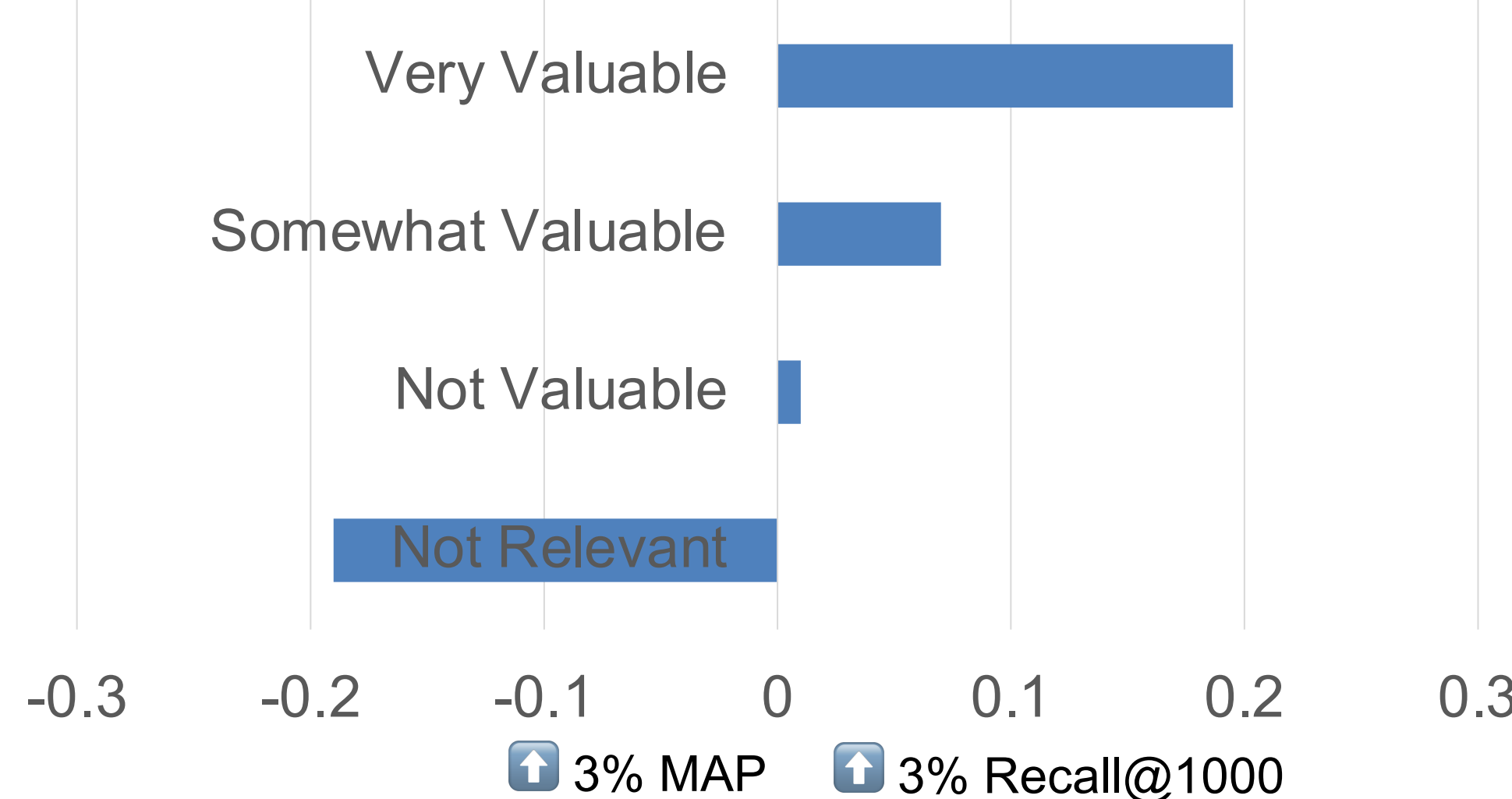REUTERS CNBC ENCYCLOPÆDIA Britannica ehistory Investopedia BBC Forbes HISTORY THE BROOKINGS INSTITUTE OFFICE OF THE HISTORIAN

Baselines

Existing methods fail on complex topics that require multi-document reasoning, external knowledge, or understanding of concepts

2021 TREC DL Document Ranking

NDCG@10

BM25+RM3   ANCE   BM25+RM3+T5

■ Document Ranking   ■ Entity Ranking

Entity Query Expansion

There's a positive correlation between document and entity relevance.

Pearson Correlation: Relevant Entities and Document Relevance

Very Valuable
Somewhat Valuable
Not Valuable
Not Relevant

-0.3 -0.2 -0.1 0 0.1 0.2 0.3

⬆ 3% MAP   ⬆ 3% Recall@1000

Entity-QE+T5: Expanding queries with relevant entities improves document ranking

Query Reformulations

For each topic there are manual query reformulations that improve effectiveness

Document Ranking

Entity Ranking

Recall@1000

Original Query   Best Reformulation   Reform-QE

Original Query   Best Reformulation   Reform-QE
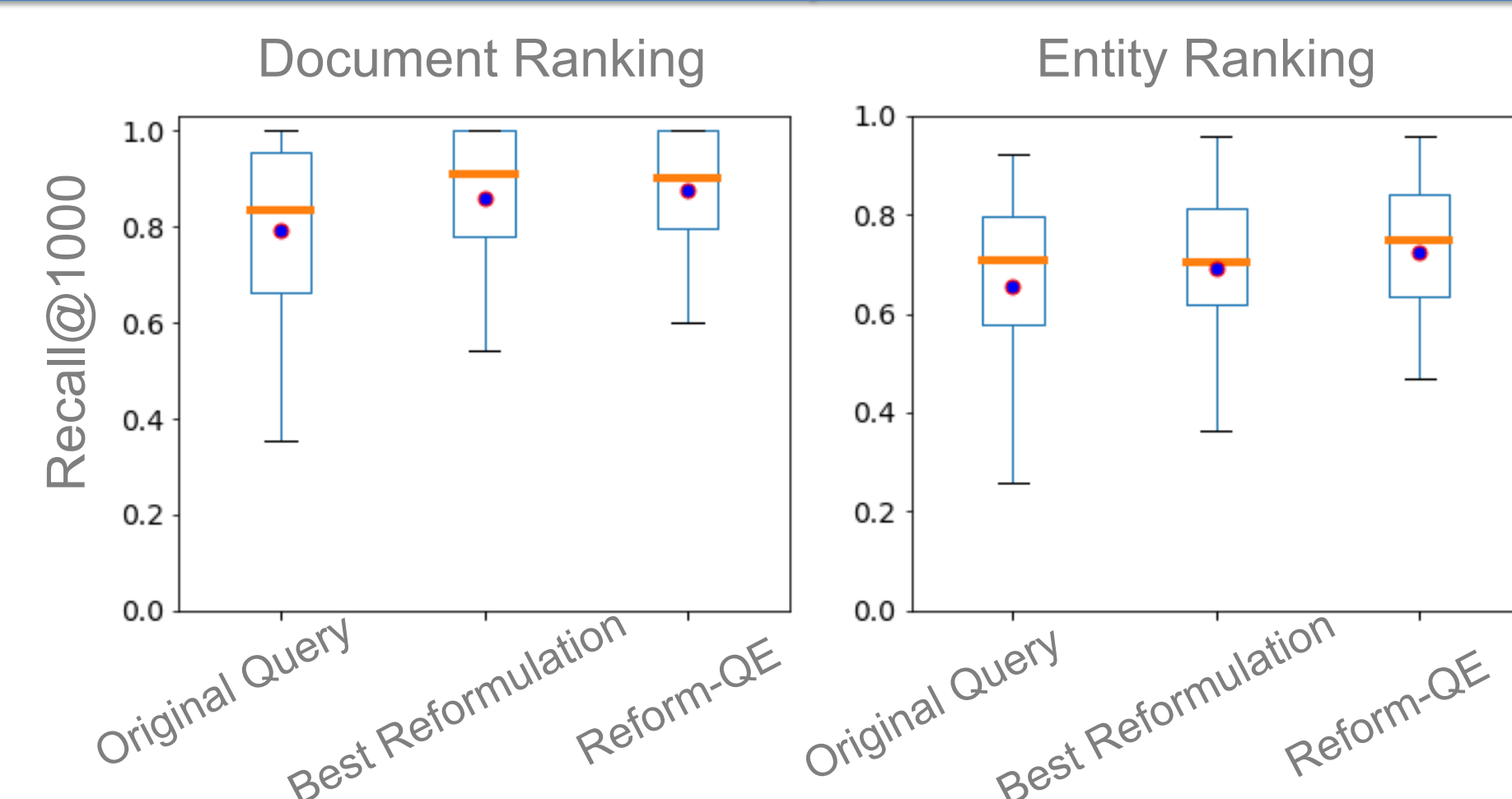
Reform-QE+T5: Multiple queries improves effectiveness with query expansion

Document Ranking

⬆ 3%   ⬆ 8%

MAP   Recall@1000

■ BM25+RM3+T5   ■ Reform-QE+T5