Figure-Inspired Text Retrieval

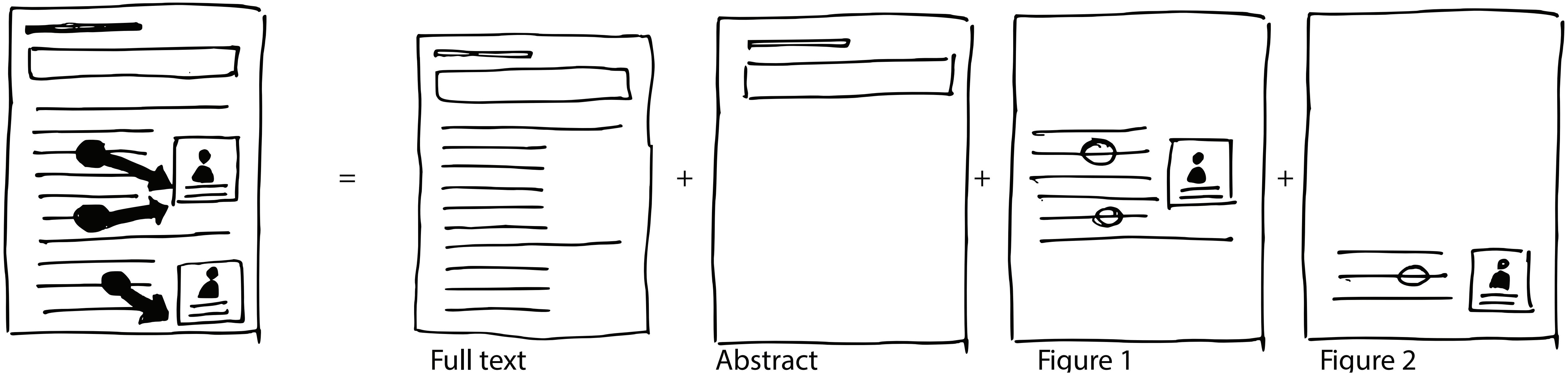
Jesse M Lingeman, Laura Dietz
BioNLP/CIIR Research Groups

UMASS
AMHERST

Task: Retrieve passages given question

- Participated in Task 2b Phase A of the BioASQ challenge
  - **Goal**: Given a natural language query, return documents and passages that ideally contain the answer to the query

Idea



- Important concepts in research publications merit a figure
- Use figure-associated text as coherent concepts
- Each Pubmed document broken into three parts: figure-related text, abstract text, and full text

Data

- Full text versions of the documents in Pubmed Central
- Queries provided by BioASQ
- Gold standard document/passage annotations provided by BioASQ

Basic Pipeline

- Preprocess Query
- Expand Query
- Retrieve Documents
- Generate features from retrieved documents
- Rerank documents

Query Preprocessing

- Example Query:

“~~Are there any~~ urine biomarkers ~~for~~ bladder cancer diagnosis~~?~~”

Sequential Dependence Retrieval Model (SDM):

0.8 * “urine” “biomarkers” “bladder” “cancer” “diagnosis”
0.15 * “urine biomarkers” “bladder cancer” “cancer diagnosis”
0.05 * “urine NEAR biomarkers” “bladder NEAR cancer” “cancer NEAR diagnosis”

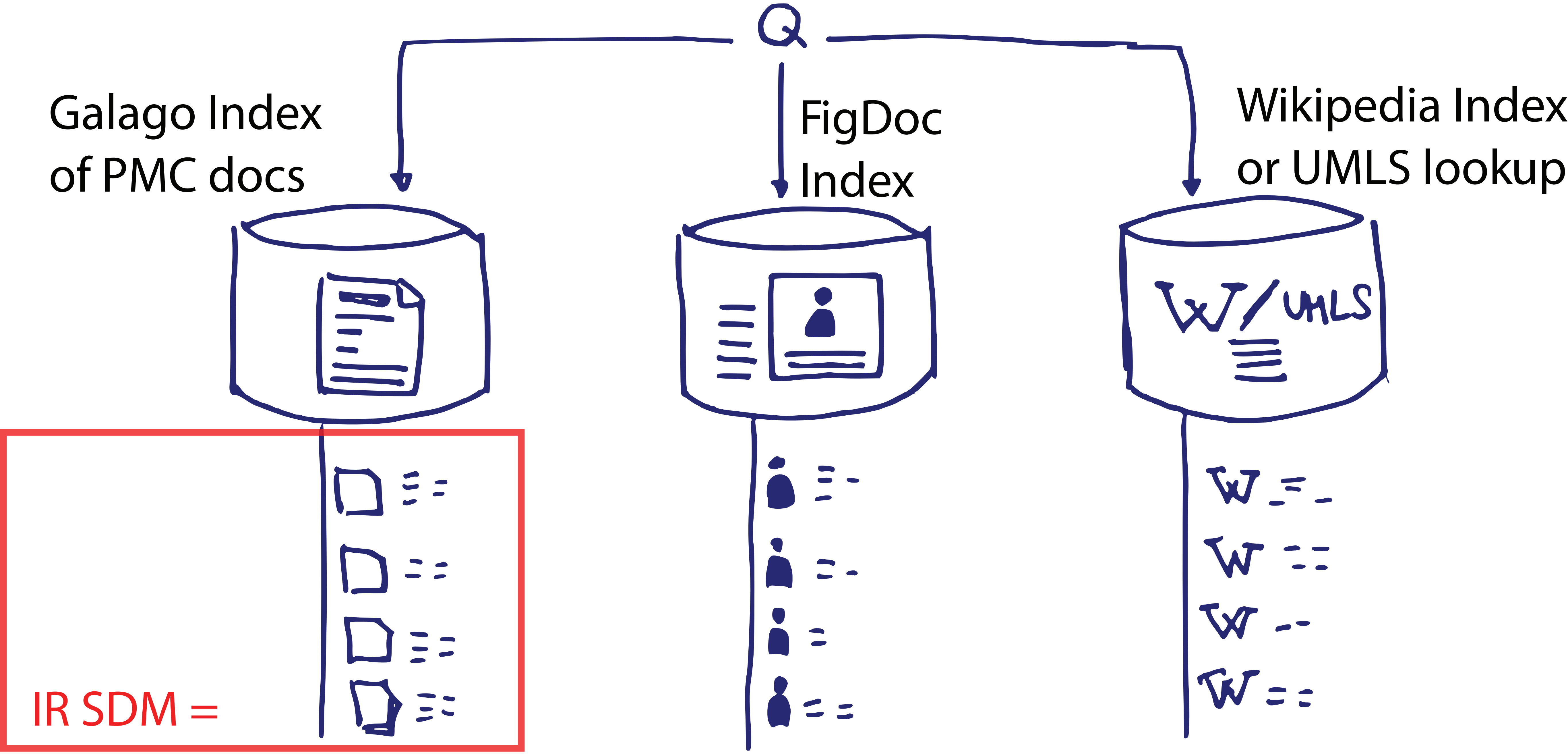
Pseudo-Document Creation

- In order to expand the queries, we create “pseudo-document” corpora that we can query alongside PMC documents.

- We do this for Wikipedia, UMLS, PMC Abstracts, and “Figure Documents”.

- **Idea:** These allow us to get term distributions from special parts of documents and from external resources for query expansion

Query Expansion and Retrieval



Galago Index
of PMC docs

FigDoc
Index

Wikipedia Index
or UMLS lookup

IR SDM =

Query Expansion and Retrieval

"~~Are there any~~ urine biomarkers ~~for~~ bladder cancer diagnosis~~?~~"

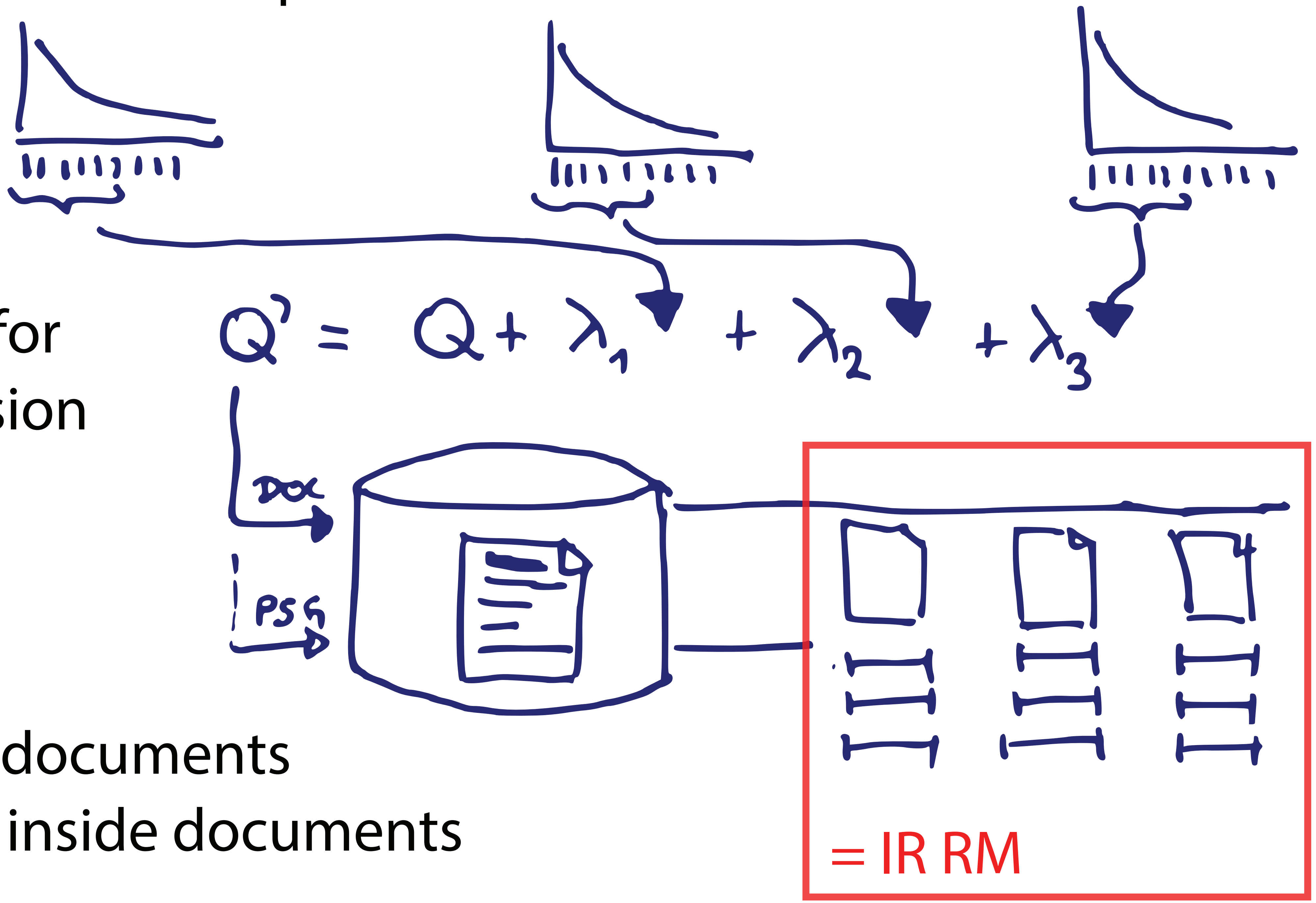
FigDoc RM	UMLS	Wiki	Abstract
biomarker	cellular	cell	patients
patients	neoplasm	carcinoma	protein
panel	malignant	urinary	tumor
cells	carcinoma	cells	test
levels	urinary	frequent	expression
samples	stage	urination	urinary
serum	cell	transitional	detection

Query Expansion and Retrieval

Word distributions of top k results

Top k' words for query expansion

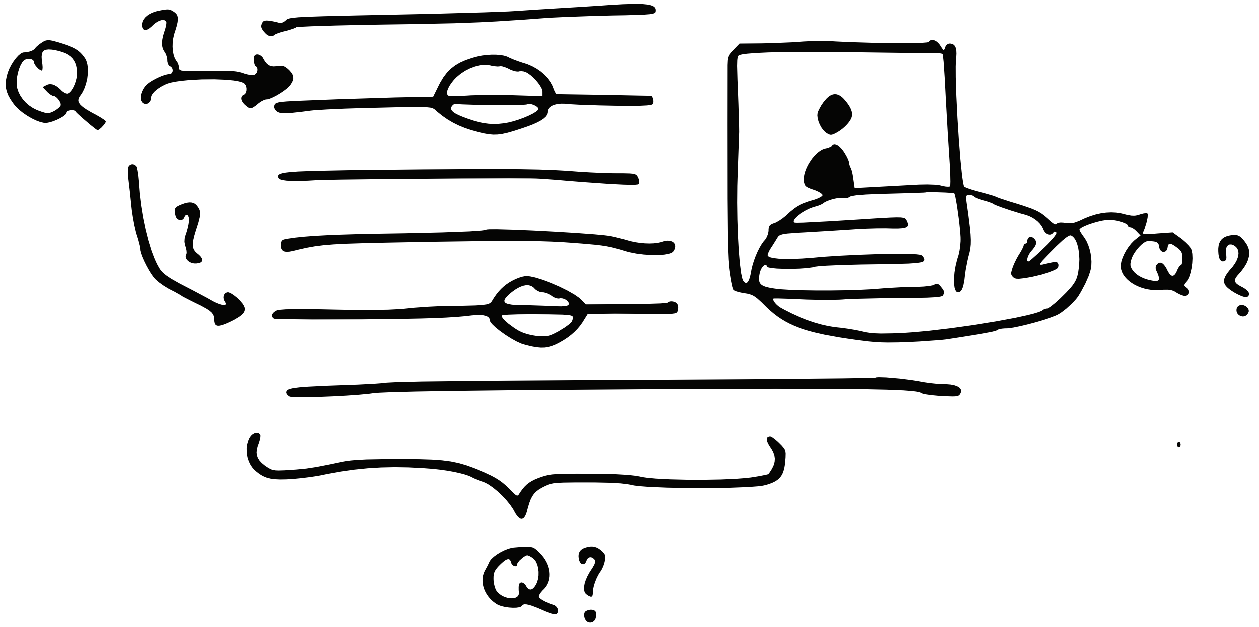
Retrieve PMC documents and passages inside documents



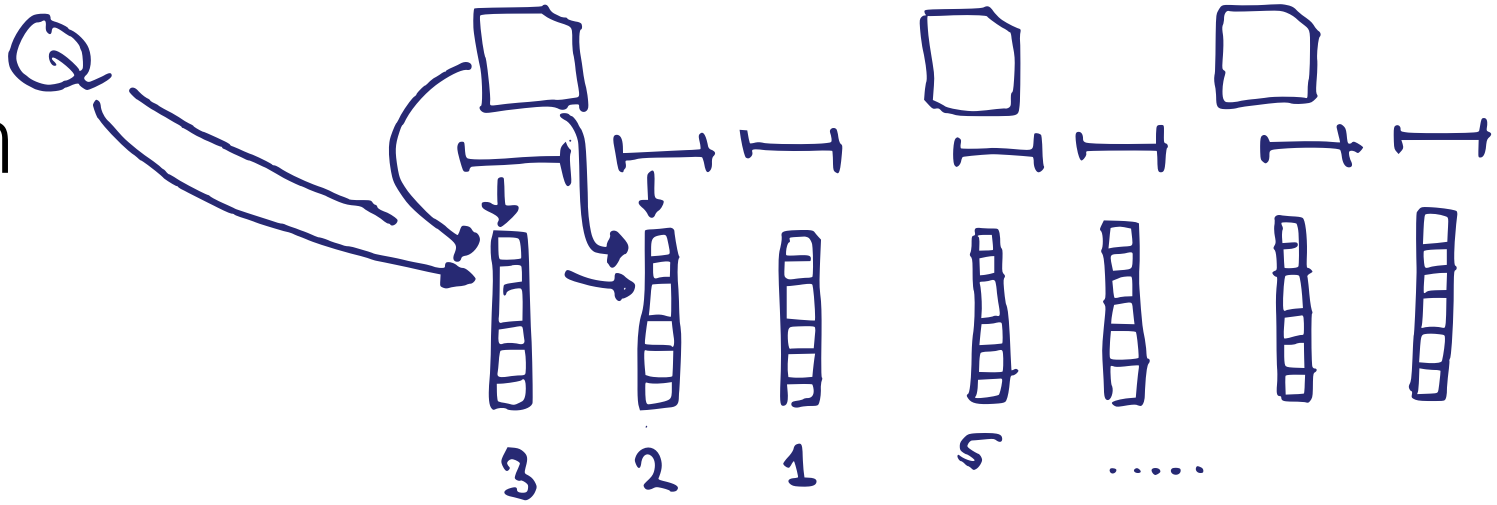
= IR RM

Feature Extraction

Features extracted from figures and figure-related text



Features extracted from documents, passages, and IR ranking



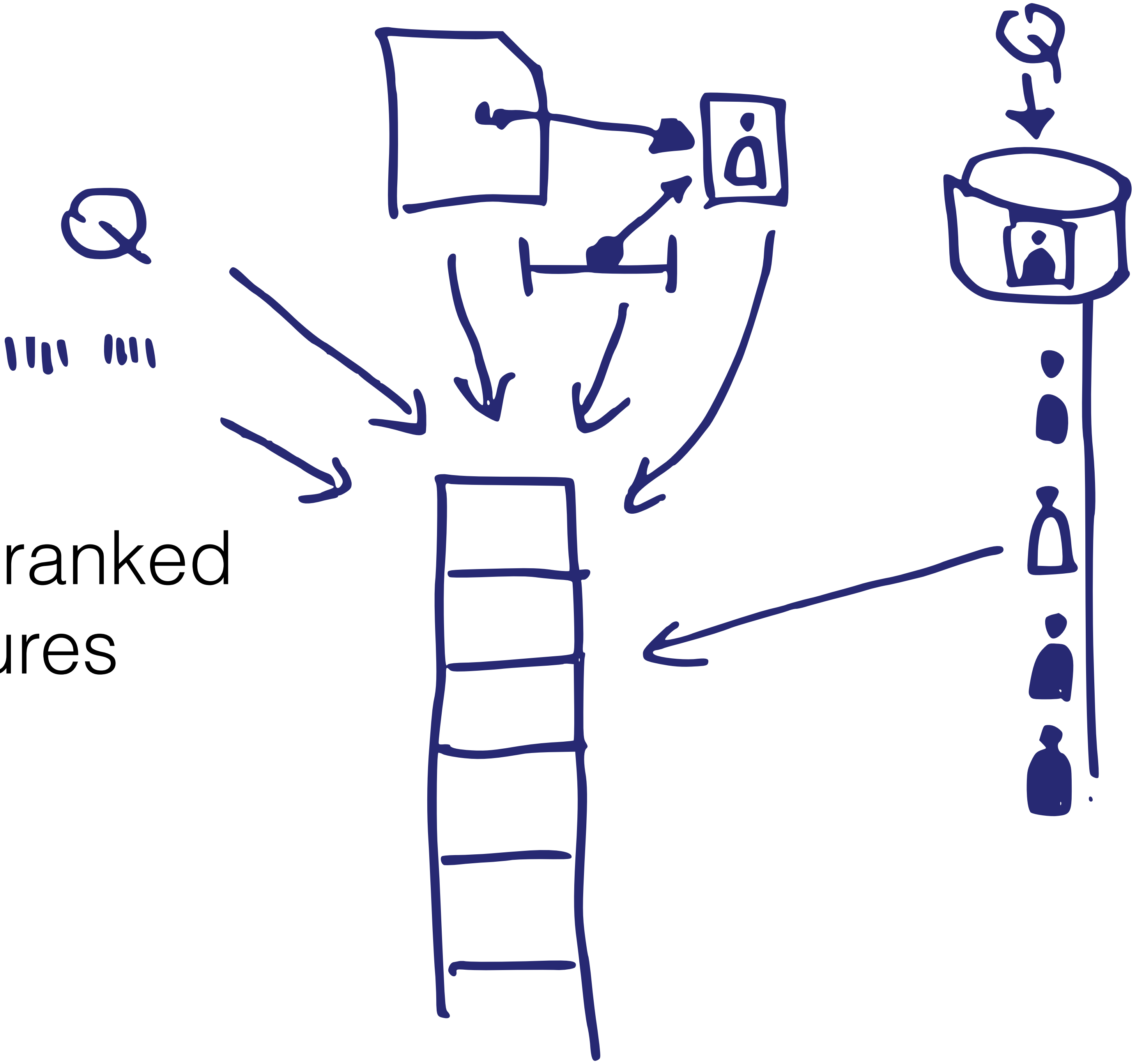
Feature Extraction

- **IR and Doc Features:**
    - IR Features:
        - Retrieval score and rank under: SDM, unigram, and expansion (of document and passage)
    - Document Features:
        - Query cover and TF-IDF of passage, full text, title, abstract, citations, and tables

Feature Extraction: Figures

- **Figure Features:**
  - Figure Document Retrieval Features:
    - Number of figures contained in top *K* of FigDoc ranking, rank of contained figures in FigDoc ranking
  - Figure Text Features:
    - Query cover in figure captions, figure references, and sentences neighboring figure references

Reranking



Retrieved documents reranked
from generated features

Results?

- Terrible!
  - Almost 0 for documents!
  - 0 for passages!
- ***So what happened?!***

Issue: Predicted passages outside of gold std

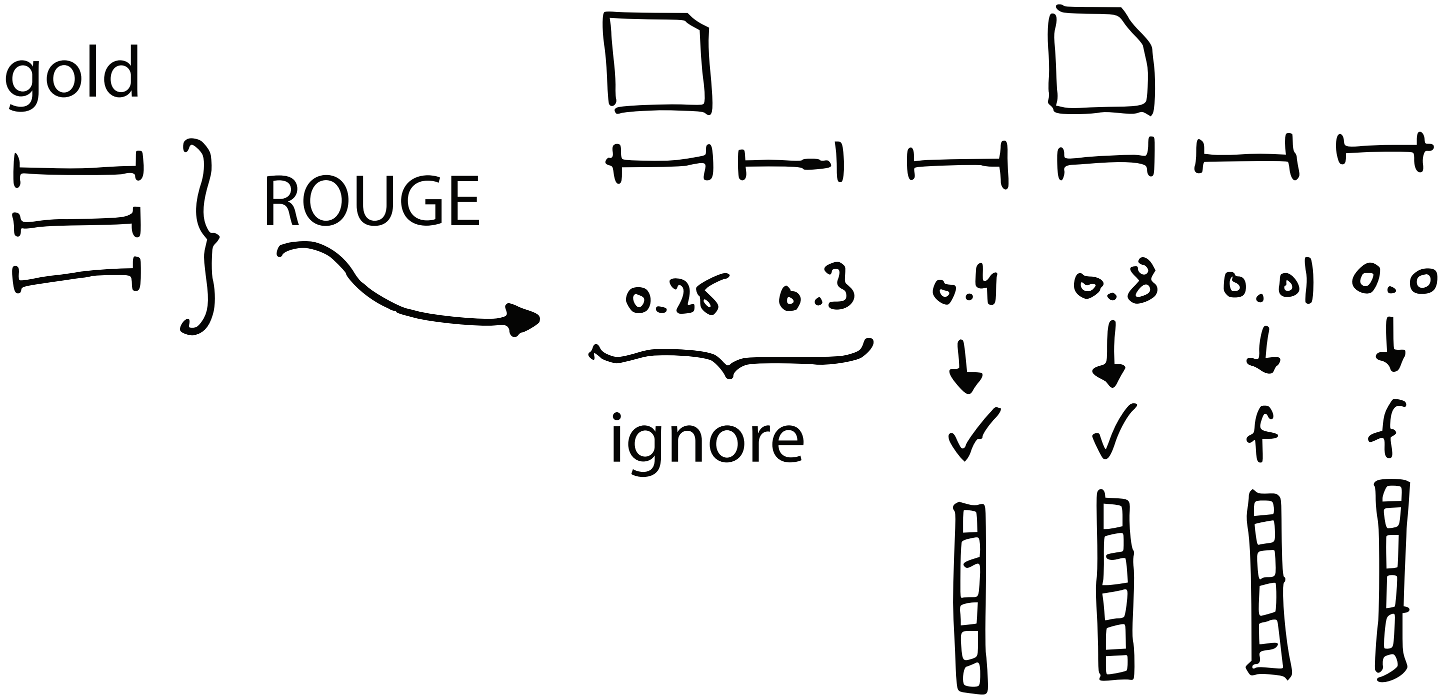
- Most of the documents provided only contained abstracts and no figures! So we used versions that did have the figures.

- We manually annotated a query to confirm that we were actually retrieving relevant answer passages (Precision @ 20 = 0.5 on the bladder cancer query)

- “Silver standard” was created to analyze results

Solution: ROUGE-based Silver Standard

Training:

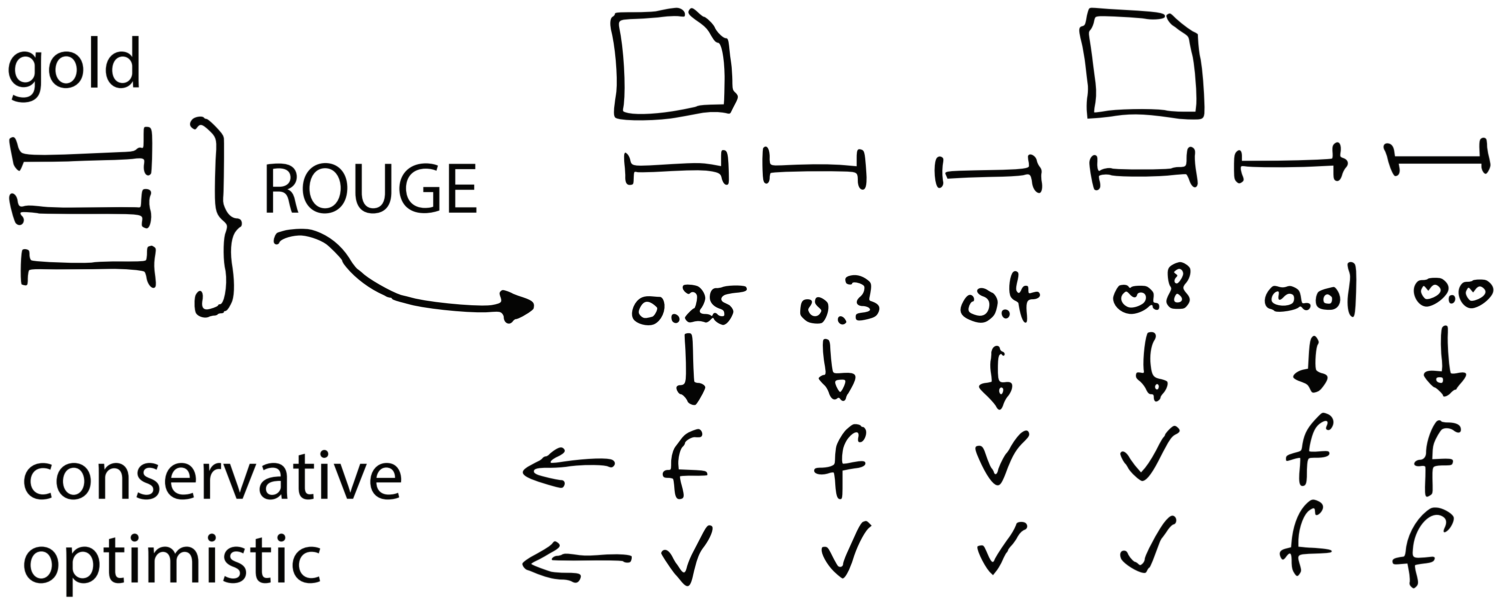


gold

ROUGE

ignore

Test:



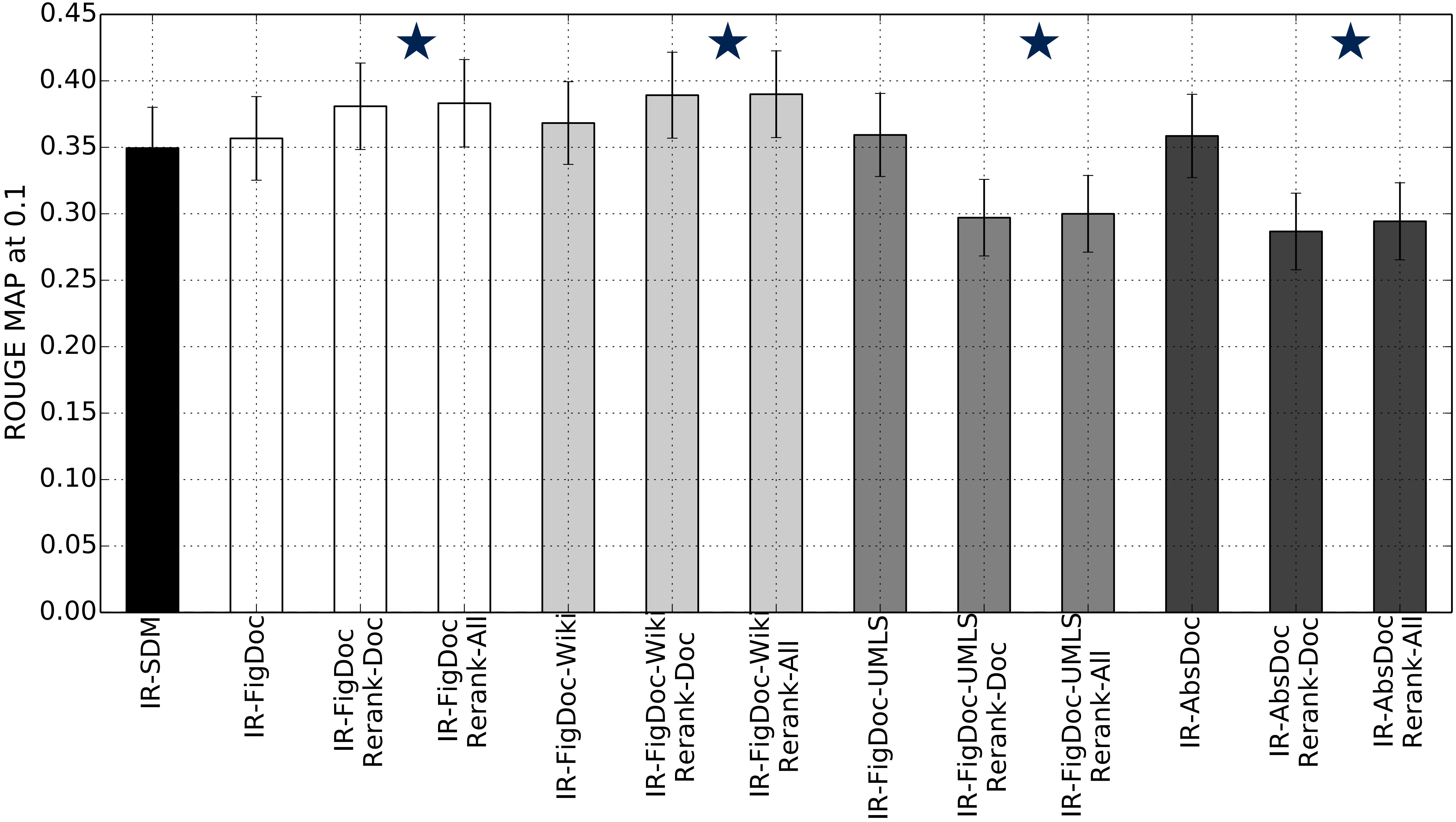
gold

ROUGE

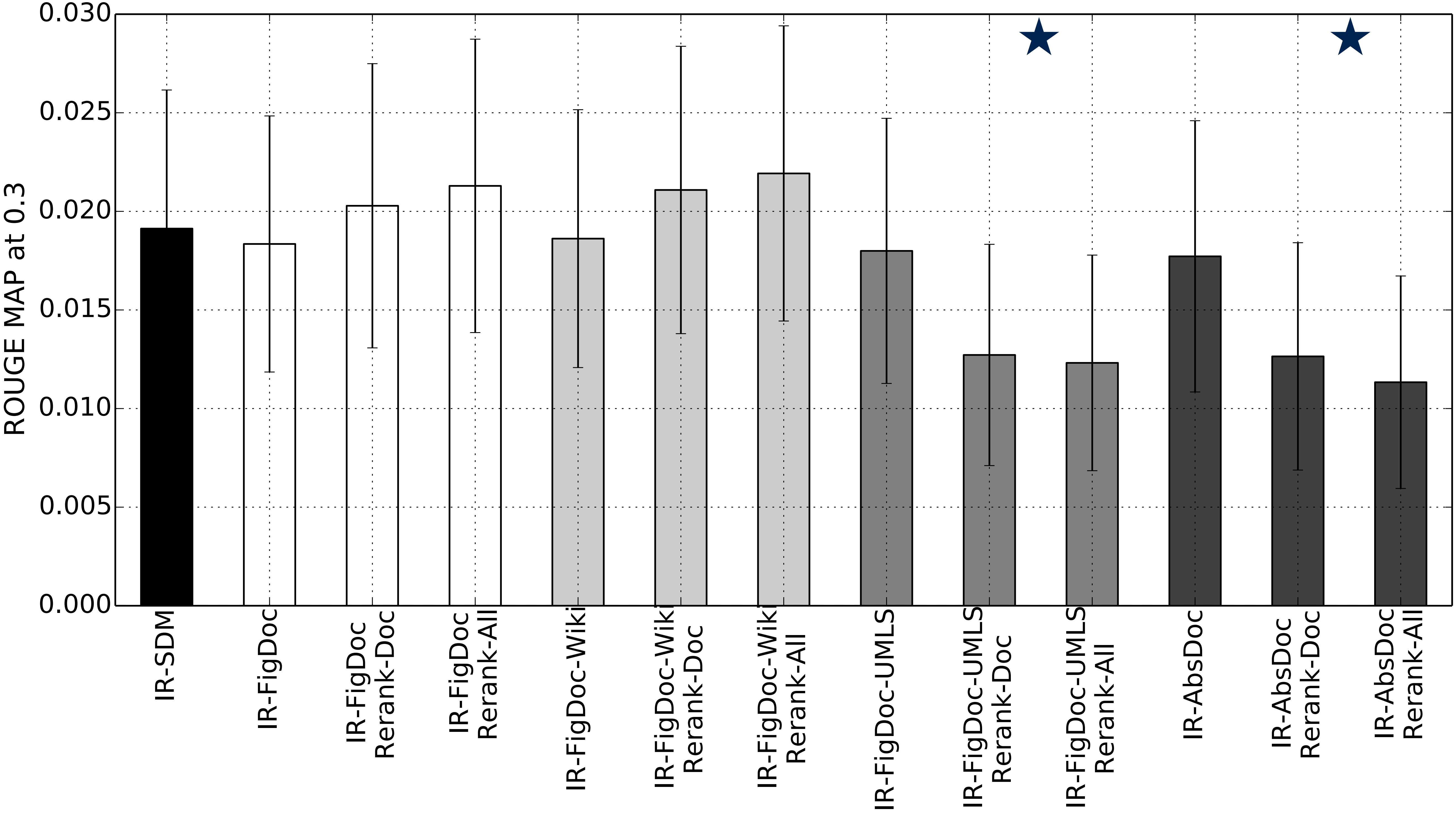
conservative

optimistic

Passage Results - Optimistic



Passage Results - Conservative



Conclusions

- So it turns out that figures, at least the way we used them, were not particularly helpful compared to baseline.

- But the ROUGE-based annotation allows us to analyze our results when we are working with out-of-corpus data that was not judged manually judged.

- We hypothesize our real results lie somewhere between the optimistic and conservative ROUGE thresholds.

Thanks!