Interpretability using Axiomatic IR

Explainable Information Retrieval

Procheta Sen

University of Liverpool

How it started

- Axiomatic Framework: Understanding Information Retrieval (Fang et al. SIGIR 2004)
- Given query Q, when would you prefer D_i over D_j?
- Formalised necessary (good) heuristics for retrieval effectiveness
- Relevance was defined as a set of formally defined constraints (axiom)
- Well known constraints to govern *term-weighting* schemes

Popular term weighting schemes

- Pivoted Normalisation (Vector Space Model)

$$\sum_{w \in D \cap Q} \frac{1 + \log(1 + \log(c(w, d)))}{(1 - s) + s\frac{|d|}{avdl}} \times \log(\frac{N + 1}{df(w)})$$

- BM25

$$\sum_{w \in D \cap Q} \log \frac{N - df(w) + 0.5}{df(w) + 0.5} \times \frac{(k_1 + 1) \times tf(w, d)}{k_1((1 - b) + b\frac{|d|}{avdl}) + tf(w, d)}$$

Research questions

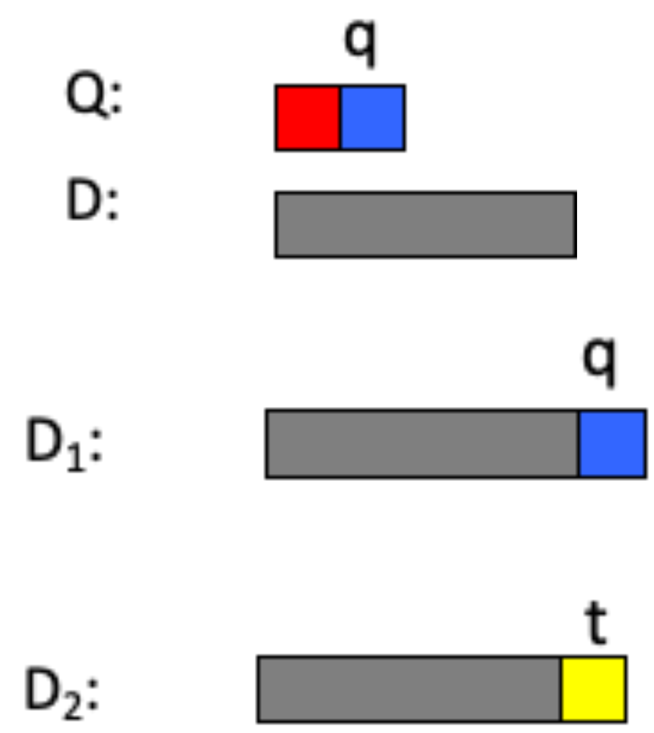
- M_1 and M_2 —Although derived differently, why do these two models perform similarly?
  - They share some common properties
- Why are they better than some other variants?
  - Other variants don't have "good" properties

Axiom structure (TFC1)

- Favour a document (higher score) with more occurrences of a query term

Let Q = {w} be a single term query, d_1 and d_2 be two documents having equal length.

If $count(q,d_1) > count(q,d_2)$ then $Score(q,d_1) > Score(q,d_2)$



Popular Axioms

Constraints	Intuitions
TFC1	To favor a document with more occurrences of a query term
TFC2	To ensure that the amount of increase in score due to adding a query term repeatedly must decrease as more terms are added
TFC3	To favor a document matching more distinct query terms
TDC	To penalize the words popular in the collection and assign higher weights to discriminative terms
LNC1	To penalize a long document (assuming equal TF)
LNC2, TF-LNC	To avoid over-penalizing a long document
TF-LNC	To regulate the interaction of TF and document length

Axiomatic Analysis and Optimisation of Information Retrieval Models, Fang and Zhai, *SIGIR Tutorial 14.*

Analysis

- Okapi aka BM25 performs poorly for verbose queries (Violates Constraints)
  - Modify formulae to satisfy constraints $\Longrightarrow$ Performs better!
- Relatively stable performance of BM25 compared to Pivoted Length Normalisation w.r.t parameter variation
- Empirical performance is related to how well they satisfy constraints

Axiomatic Result Reranking

- Turn any retrieval model to Axiom Compliant one [Hagen et al. CIKM 2016]
- **Step 1:** Start with any *top-k* ranking
- **Step 2:** Axiom aggregation:
  - For each axiom A_i compute preference/ordering of D_j and D_k
  - $M_{A_i}[j,k] = \begin{cases} 1, & \text{if } D_j > D_k \\ 0, & \text{otherwise} \end{cases}$

Axiomatic Result Re-Ranking

- Step 2: Axiom aggregation:
  - Set of 23 axioms
  - Relaxed version of some axioms
    - Extension (one query term to multiple query terms)
    - Relaxation (approximately fulfil the relationship)
  - Combined with learned aggregation function (retrieval-specific)
  - Classification Problem (Random Forest)

Axiom_1 : TFC1

$$M_1 \quad \begin{array}{c|ccccc} & 1 & 2 & 3 & \ldots & k \\ \hline 1 & . & 0 & 1 & & 1 \\ 2 & & . & 1 & & 0 \\ 3 & & & . & & 1 \\ \vdots & & & & & \vdots \\ k & & & & & . \end{array}$$

$\vdots$

Axiom_{23} : ORIG

$$M_{23} \quad \begin{array}{c|ccccc} & 1 & 2 & 3 & \ldots & k \\ \hline 1 & . & 1. & 1 & & 1 \\ 2 & & & 1 & & 1 \\ 3 & & & . & & 1 \\ \vdots & & & & & \vdots \\ k & & & & & . \end{array}$$

Meta learning of axiom impact

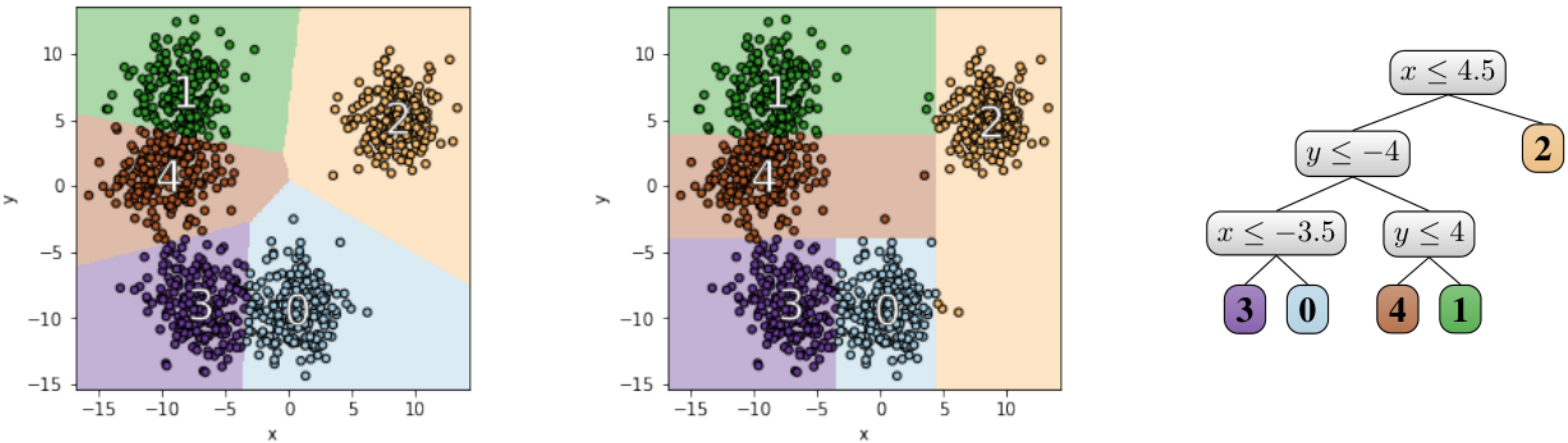
$$f_R\,(M_1, \ldots, M_{23})$$

Axiomatic Result Re-Ranking, Hagen et al., *CIKM 16.*

Axiomatic Result Re-Ranking

- Step 3: Combining preferences
  - Could contain conflict $D_j > D_k$, $D_k > D_l$, $D_l > D_j$
  - Translates to rank-aggregation problem
  - Objective: minimize distance function to the original m rankings (NP-Complete)
  - Apply KwikSort (Ailon et al. JACM 2008) on resulting matrix
- Observation: output is axiom compliant and effectiveness is better!

Similar analogy



(a) Optimal 5-means clusters (b) Tree based 5-means clusters (c) Threshold tree

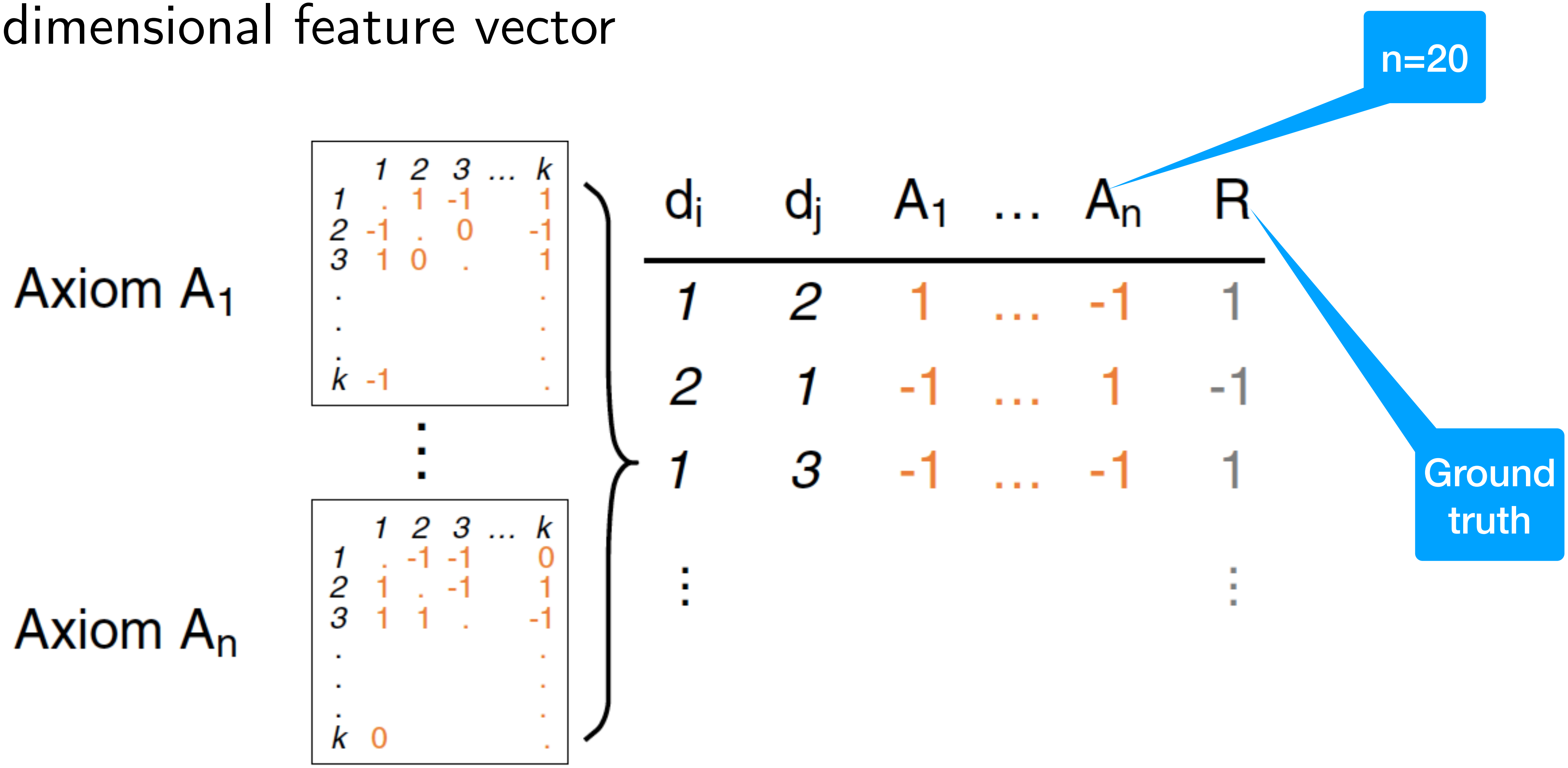
Explainable k-Means and k-Medians Clustering, Dasgupta et al., ICML 20.

Axiomatic Explanations of Neural Models

- RQ(s): To what extent can we explain neural models with Axiomatic Framework? (Völske et al. ICTIR 2021)
- Post-hoc explanations of IR models
- 20 axioms were considered
- Simple classification model (Random Forest) to make pairwise decision

Intuitive diagram

- Objective is to classify the preferences: based on 20 dimensional feature vector



Towards Axiomatic Explanations for Neural Ranking Models, Volske et al., *ICTIR 21*.

Observations

- Large difference in retrieval score can be well explained
- Pairs with more similar retrieval scores are difficult to explain

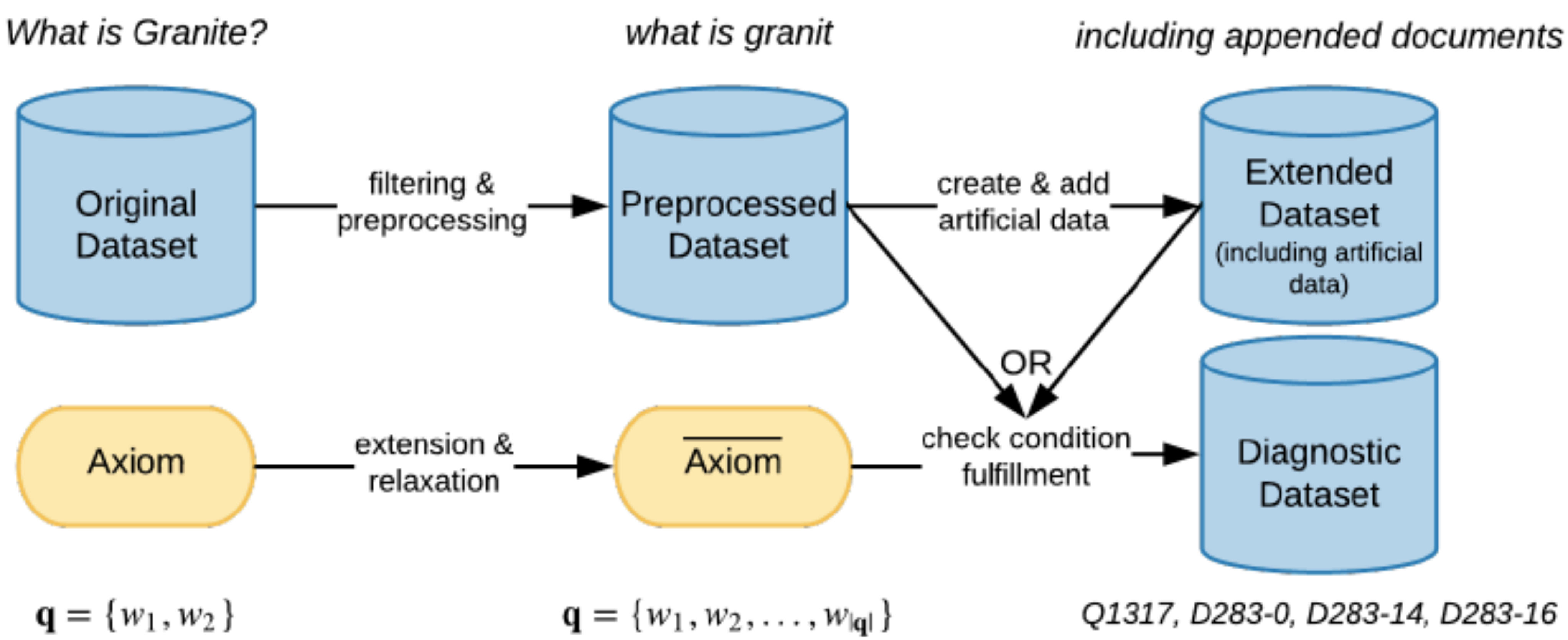
Diagnosing Neural IR Models

- Diagnostic Dataset (Renning et al. ECIR 2019)
- RQ: To what extent do neural IR models fulfil the axioms?
- Relaxed and Extended version of TFC1, TFC2, TDC, LNC2
- Diagnosed models : BM25, LMDIR, DRMM, aNMM, Duet, MatchPyramid

Diagnostic Datasets

- Originally inspired from NLP, Computer Vision domain
- For NLP fine grained linguistic Tasks: anaphora resolution, entailment,...
- Answer-Passage retrieval dataset WikiPassageQA
- Sample <query, document pairs> triplets
- If it satisfies axioms put It in the diagnostic set

Pipeline and objective



Objective : Given a tuned model how well they can predict the axiomatic preferences?

An Axiomatic Approach to Diagnosing Neural IR Models, Rennings et al., *ECIR 19.*

Dataset statistics

	TFC1	TFC2	M-TDC	LNC2^{Test}	LNC2^{All}
Parameters				$k = \{2, 3, 4\}$, $doc\_len_{max} = 240$	
Train	2,758,223	837,838	33,509	0	82,785
Dev	376,902	50,772	3,958	0	10,485
Test	353,621	183,898	4,497	10,074	10,074
Total	3,488,746	1,072,508	41,964	10,074	103,344

An Axiomatic Approach to Diagnosing Neural IR Models, Rennings et al., *ECIR 19.*

Observation

	MAP	MRR	P@5	TFC1	TFC2	TDC	LNC2(T)	LNC2(A)
BM25	0,52	0,60	0,18	0,73	0,98	1,00	0,80	0,80
LMDIR	0,54	0,62	0,19	0,87	0,63	0,94	0,68	0,68
Duet	0,25	0,29	0,10	0,69	0,56	0,48	0,19	0,47
MatchPyramid	0,44	0,51	0,18	0,79	0,58	0,63	0,00	0,19
DRMM	0,55	0,64	0,20	0,84	0,60	0,76	0,05	0,12
aNMM	0,57	0,66	0,21	0,85	0,56	0,69	0,38	0,47

Observations

- Fulfilment of axioms is not a good indicator for NRM
- NRMs did (not) learn some patterns
- Could fix Duet model with additional triplets

Diagnosing Distill BERT Model

- RQ: Why BERT based model is so powerful?
- Diagnosing dataset from TREC 2019 Deep Learning track
- Using 9 axioms (TFC1, TFC2, TDC, LNC1, LNC2, STMC1, STMC2, STMC3, TP)

- Retrieve top-k (100) with LMDIR
- Add pair of documents D_j, D_k if they satisfy constraints
- For LNC2 create duplicate documents for test set only
  - Recall that LNC2 says we should avoid over-penalizing long relevant documents.

- Retrieval effectiveness wise DistilBERT > QL

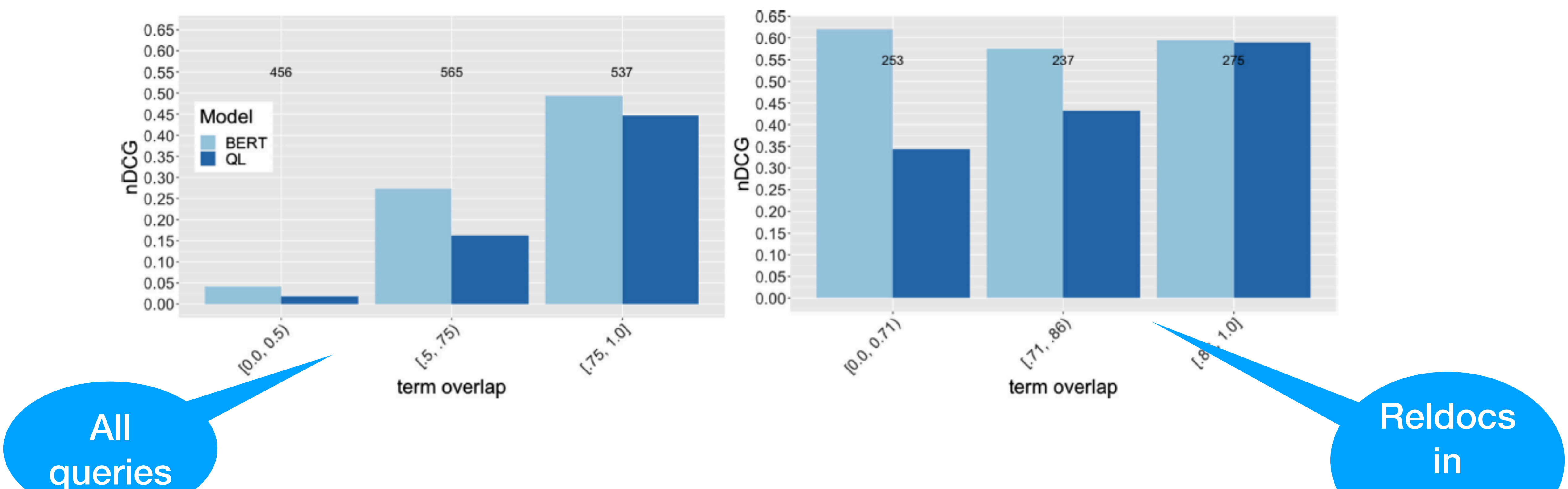
- Axioms are "not applicable" or "not sufficient"

	nDCG	MRR	TFC1	TFC2	M-TDC	LNC1	LNC2	TP	STMC1	STMC2	STMC3
QL	0.2627	0.3633	**0.99**	**0.70**	**0.88**	**0.50**	**1.00**	0.39	0.49	**0.70**	**0.70**
DistilBERT	**0.3633**	**0.4537**	0.61	0.39	0.51	**0.50**	0.00	**0.41**	**0.50**	0.51	0.51

Diagnosing BERT with Retrieval Heuristics, Camara and Hauff, *ECIR 20.*

Further Investigation(s)

- Divide Q, D_R pair into three buckets
- Query/Document pair has few, moderate and large overlap



Diagnosing BERT with Retrieval Heuristics, Camara and Hauff, *ECIR 20.*

Question(s)

- Axioms are not complete yet!
- BERT models fail to adhere to many constraints still perform really well...
- We need more (better) axioms to explain them

References

- "A formal study of information retrieval heuristics", Fang et al., SIGIR 2004.
- "Axiomatic Result Re-Ranking", Hagen et al., CIKM 2016.
- "Aggregating inconsistent information: Ranking and clustering", Ailon et al., JACM 2008.
- "Explainable k-Means and k-Medians Clustering", Dasgupta et al., ICML 2020.
- "Towards Axiomatic Explanations for Neural Ranking Models", Volske et al., ICTIR 2021.
- "An Axiomatic Approach to Diagnosing Neural IR Models", Rennings et al., ECIR 2019.
- "Diagnosing BERT with Retrieval Heuristics", Camara and Hauff, ECIR 2020.

Thank you