An Axiomatic Approach to Diagnosing Neural IR Models

Daniël Rennings, Felipe Moraes, Claudia Hauff
Delft University of Technology, the Netherlands

Applied and Engineering Sciences

Why?

*Why **diagnose** neural IR models?*

*Why adopt an **axiomatic approach**?*

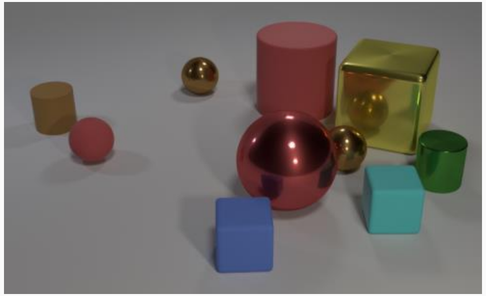
Why **diagnose** neural IR models?

- Neural IR has **not** (yet) achieved the progress seen in Computer Vision / NLP
- Issues:
  - Lack of large scale public datasets
  - Lack of shared public code repositories
  - **Lack of approaches to interpret and analyze neural IR models**
    - CV and NLP communities fare better, e.g. CLEVR, bAbI
- Can we create such an approach for analyzing neural **IR** models?

Why adopt an **axiomatic approach**?

Computer Vision

CLEVR instances diagnose aspects of **visual reasoning**, e.g. attribute identification and counting



Q: How many objects are small cubes?

NLP

bAbI tasks diagnose aspects of **reading comprehension**, e.g. counting and logical operations

Daniel picked up the football.
Daniel dropped the football.
Daniel got the milk.
Daniel took the apple.

Q: How many items is Daniel holding?

IR

we want to diagnose aspects of ***relevance*** …

… which are formalized in search heuristics or ***axioms***

?

Axiomatic Thinking in IR

*What are **axioms**?*

*How have they been **used**?*

What are **axioms?**

- **Axioms** are search heuristics that any reasonable retrieval function should satisfy

- **Term Frequency Constraint 1** (TFC1):
  - **Intuition**:
    A model must rank $\boldsymbol{d_1}$ higher than $\boldsymbol{d_2}$
    if $\boldsymbol{d_1}$ contains more query terms than $\boldsymbol{d_2}$
  - **Formally**:
    Assume $\boldsymbol{q} = \{w\}$ and $|\boldsymbol{d_1}| = |\boldsymbol{d_2}|$,
    If $c(w, \boldsymbol{d_1}) > c(w, \boldsymbol{d_2})$,
    Then $S(\boldsymbol{d_1}, \boldsymbol{q}) > S(\boldsymbol{d_2}, \boldsymbol{q})$

$\boldsymbol{q} = \{ \text{DIR} \}$

$\boldsymbol{d_1}$ $\boldsymbol{d_2}$

What are **axioms**?

- More than **twenty** axioms have been proposed

- We **explored**:
  - **TFC1** to favor a document with a **larger count of a query term**
  - **TFC2** to ensure that the **impact of TF** from 1 to 2 is larger than from 101 to 102
  - **M-TDC** to assign higher weights to **discriminative terms**
  - **LNC2** to **avoid over-penalizing** long documents

- Other constraints consider **semantic similarity**, **proximity**, ...

How can we **use** axioms?

- **TFC1**

Assume $\boldsymbol{q} = \{w\}$ and $|\boldsymbol{d_1}| = |\boldsymbol{d_2}|$,
If $c(w, \boldsymbol{d_1}) > c(w, \boldsymbol{d_2})$,
Then $S(\boldsymbol{d_1}, \boldsymbol{q}) > S(\boldsymbol{d_2}, \boldsymbol{q})$

- **BM25**

$$\sum_{w \in \boldsymbol{d} \cap \boldsymbol{q}} \left(ln \frac{|\boldsymbol{D}| - df(w) + 0.5}{df(w) + 0.5} \times \frac{(k_1 + 1) \times c(w, \boldsymbol{d})}{k_1((1-b) + b\frac{|\boldsymbol{d}|}{avdl}) + c(w, \boldsymbol{d})} \times \frac{(k_3 + 1) \times c(w, \boldsymbol{q})}{k_3 + c(w, \boldsymbol{q})} \right)$$

- TFC1 on BM25:
  - BM25 does **not always fulfill** TFC1
  - **Modified BM25** (always fulfills TFC1) leads to **higher retrieval effectiveness**

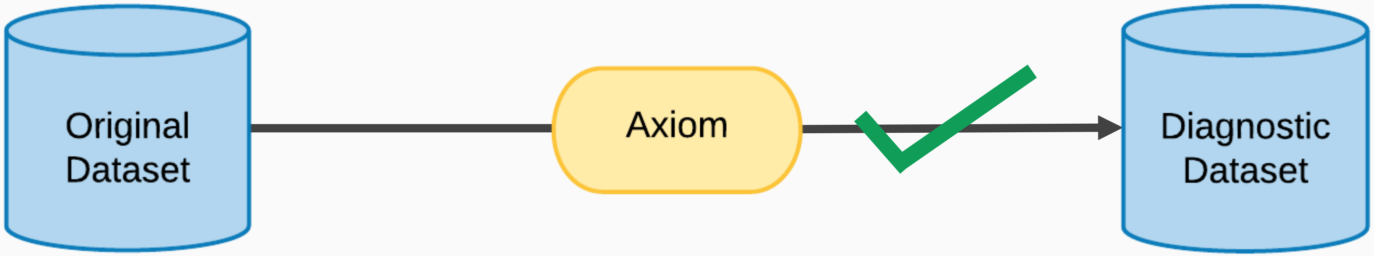
- **Problem**: not feasible for neural approaches with potentially millions of parameters!

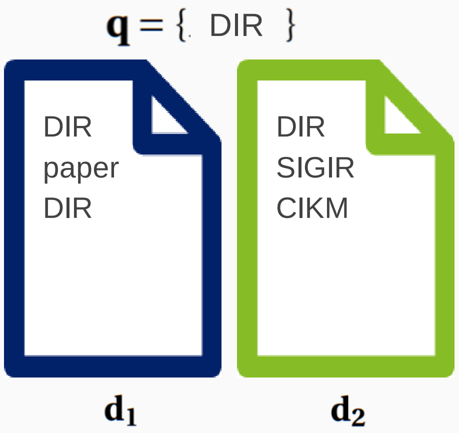
Diagnosing neural IR models

*How to **obtain** diagnostic datasets?*

*How to **use** such diagnostic datasets?*

How to **obtain** diagnostic datasets



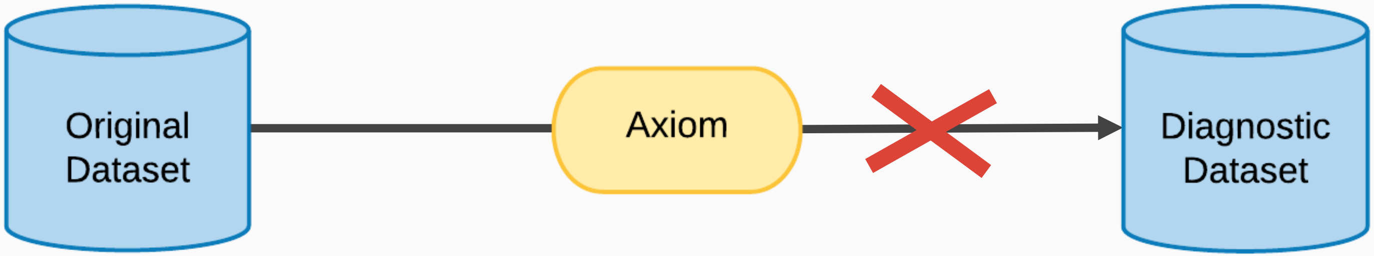


TFC1

Assume	$\mathbf{q} = \{w\}$ and $\lvert\mathbf{d}_1\rvert = \lvert\mathbf{d}_2\rvert$,
If	$c(w, \mathbf{d}_1) > c(w, \mathbf{d}_2)$,
Then	$S(\mathbf{d}_1, \mathbf{q}) > S(\mathbf{d}_2, \mathbf{q})$

Observation:
all conditions hold,
add $(\mathbf{q}, \mathbf{d}_1, \mathbf{d}_2)$ to our diagnostic dataset

How to **obtain** diagnostic datasets



$\mathbf{q} = \{$ DIR, **paper, ...** $\}$

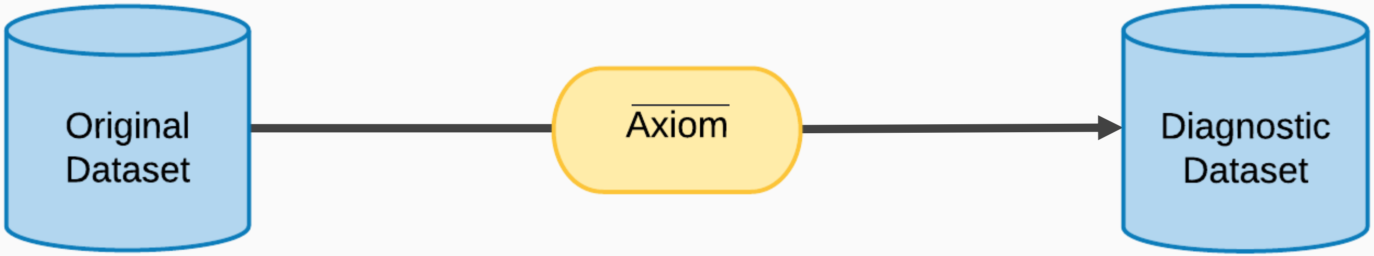
$\mathbf{d}_1$ $\mathbf{d}_2$

TFC1

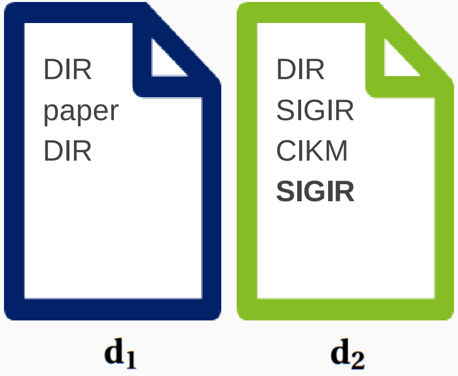
Assume	$\mathbf{q} = \{w\}$ and $\lvert\mathbf{d}_1\rvert = \lvert\mathbf{d}_2\rvert$,
If	$c(w, \mathbf{d}_1) > c(w, \mathbf{d}_2)$,
Then	$S(\mathbf{d}_1, \mathbf{q}) > S(\mathbf{d}_2, \mathbf{q})$

Problem:
$\mathbf{q} \neq \{w\}$ and $\lvert\mathbf{d}_1\rvert \neq \lvert\mathbf{d}_2\rvert$

How to **obtain** diagnostic datasets



$\mathbf{q} = \{$ DIR, **paper, ...** $\}$

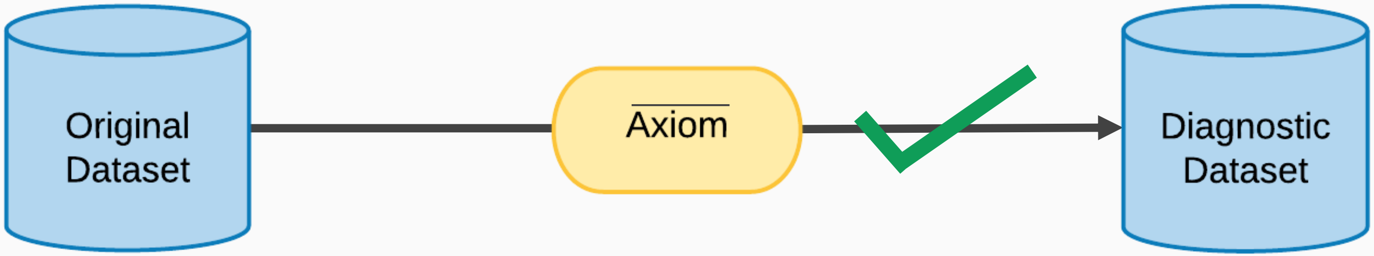


TFC1

Assume	$\mathbf{q} = \{w_1, \mathbf{w_2}, ...\}$ and $\lvert\mathbf{d_1}\rvert \sim \lvert\mathbf{d_2}\rvert$,
If	... and ...,
Then	$S(\mathbf{d_1}, \mathbf{q}) > S(\mathbf{d_2}, \mathbf{q})$

Solution:
adapt axiom to match real-world instances (e.g. multi-term queries, roughly equal length docs)

How to **obtain** diagnostic datasets



$\mathbf{q} = \{$ DIR, **paper, ...** $\}$

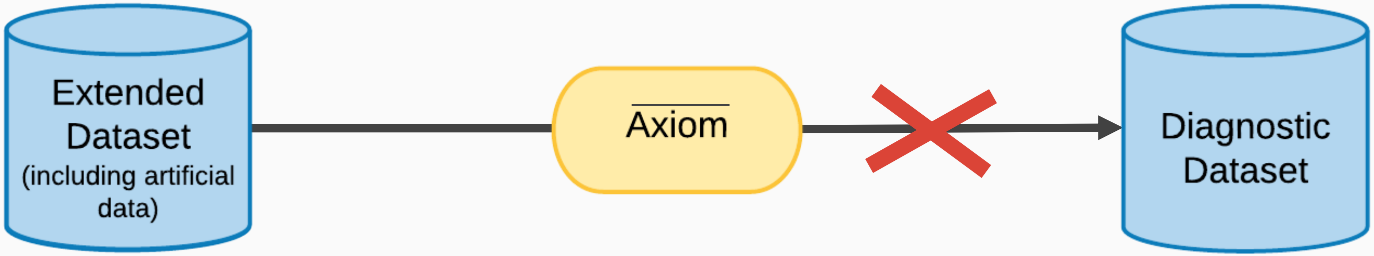
$\mathbf{d}_1$ $\mathbf{d}_2$

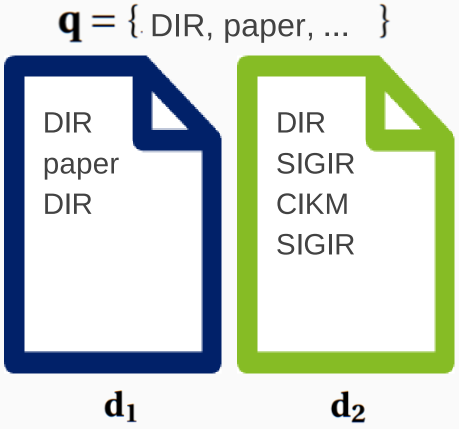
TFC1

Assume	$\mathbf{q} = \{w_1, \mathbf{w}_2, ...\}$ and $\lvert\mathbf{d}_1\rvert \sim \lvert\mathbf{d}_2\rvert$,
If	... and ...,
Then	$S(\mathbf{d}_1, \mathbf{q}) > S(\mathbf{d}_2, \mathbf{q})$

Observation:
all conditions hold,
add $(\mathbf{q}, \mathbf{d}_1, \mathbf{d}_2)$ to our diagnostic dataset

How to **obtain** diagnostic datasets



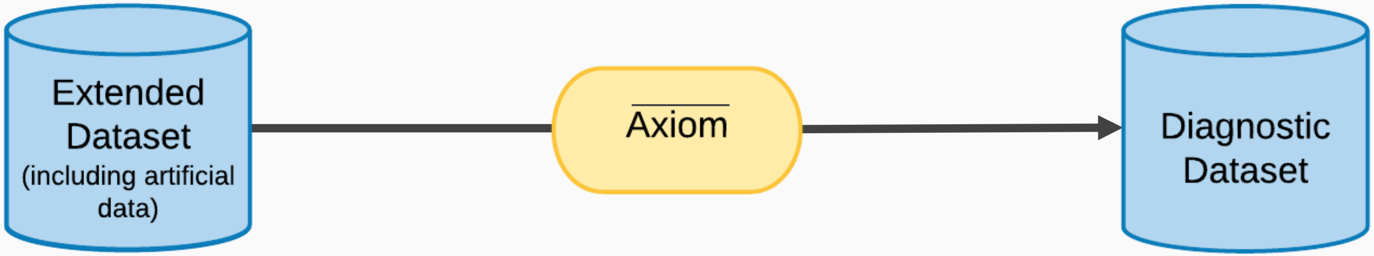


LNC2

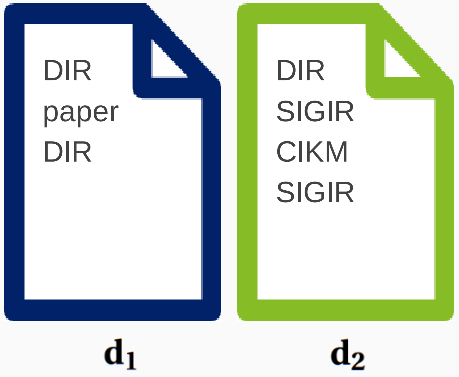
Assume	... and ...,
If	... $\mathbf{d}_1 = k \times \mathbf{d}_2$ ...,
Then	$S(\mathbf{d}_1, \mathbf{q}) \geq S(\mathbf{d}_2, \mathbf{q})$

Problem:
conditions are not likely fulfilled in regular dataset instances

How to **obtain** diagnostic datasets



$\mathbf{q} = \{$ DIR, paper, ... $\}$

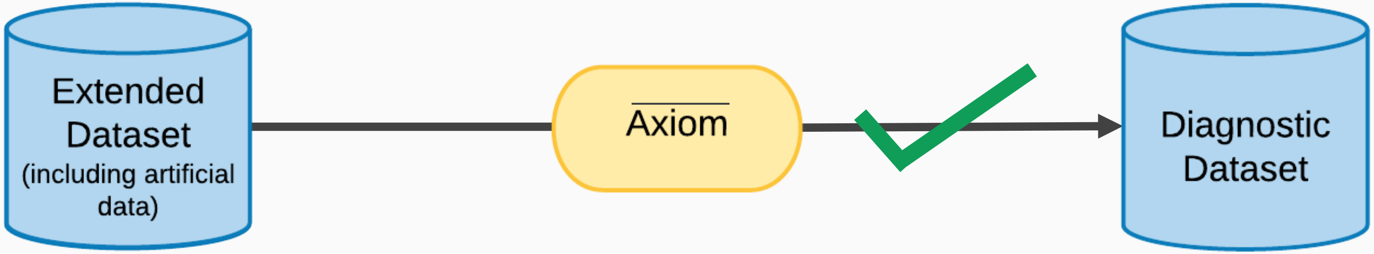


$\overline{\text{LNC2}}$

Assume	... and ...,
If	... $\mathbf{d}_1 = k \times \mathbf{d}_2$ ...,
Then	$S(\mathbf{d}_1, \mathbf{q}) \geq S(\mathbf{d}_2, \mathbf{q})$

Solution:
slightly adapt documents

How to **obtain** diagnostic datasets



$\mathbf{q} = \{$ DIR, paper, ... $\}$

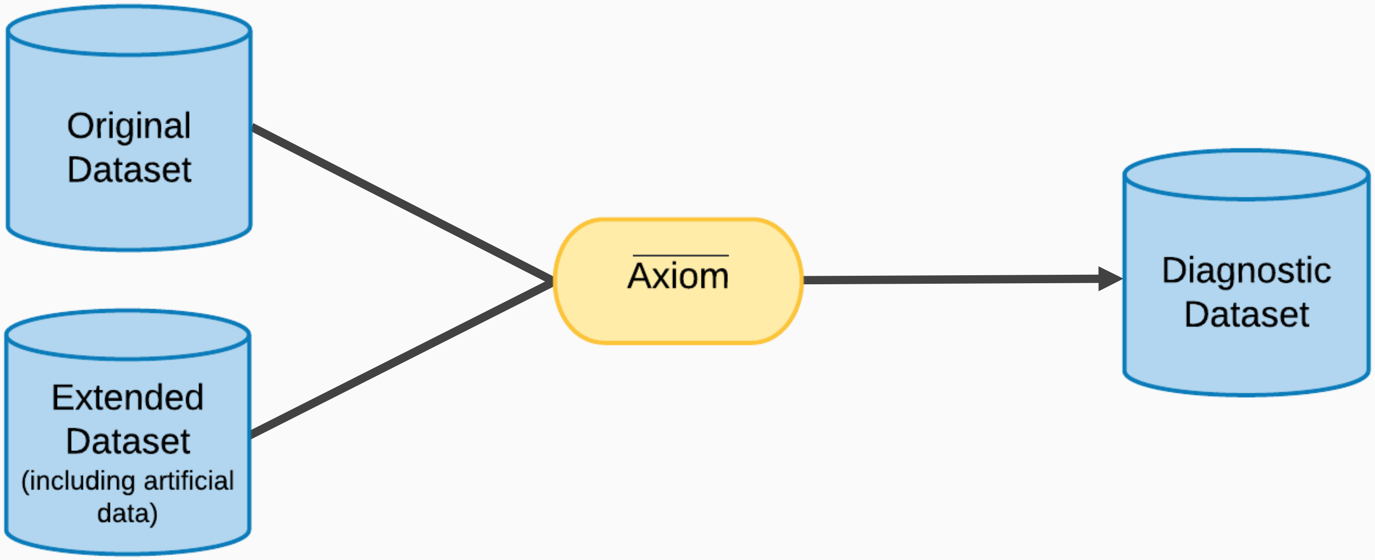
$\mathbf{d}_1$ $\mathbf{d}_2$

$\overline{\text{LNC2}}$

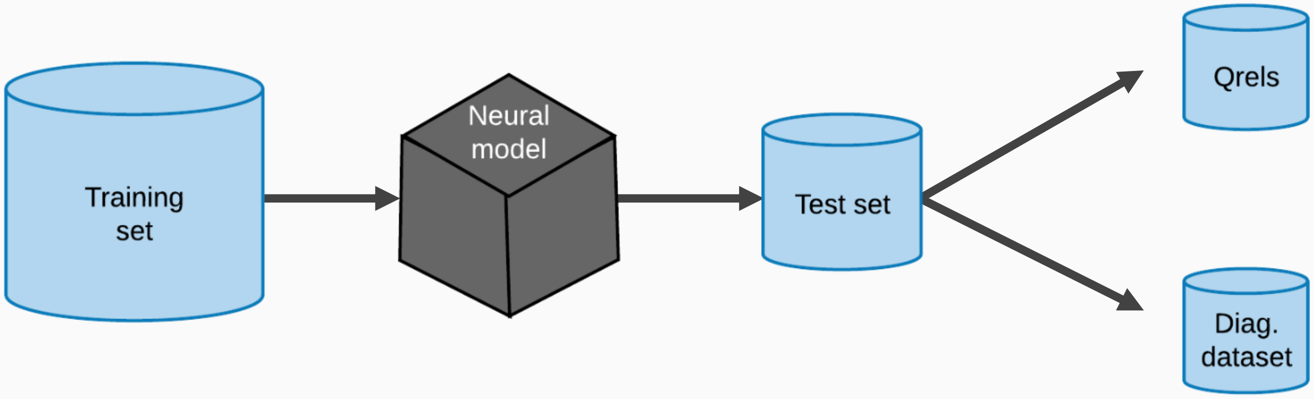
Assume	... and ...,
If	... $\mathbf{d}_1 = k \times \mathbf{d}_2$ ...,
Then	$S(\mathbf{d}_1, \mathbf{q}) \geq S(\mathbf{d}_2, \mathbf{q})$

Observation:
all conditions hold, add $(\mathbf{q}, \mathbf{d}_1, \mathbf{d}_{1,k})$ to our diagnostic dataset

How to **obtain** diagnostic datasets



How to **use** diagnostic datasets?



Retrieval Effectiveness

Explanation of Performance

Experiment

Setup

Results

Experimental **setup**

- WikiPassageQA (4K questions, 50K passages):
  - Given a **question** and all six-sentence passages making up a **Wikipedia document**, **rank the passages** such that the actual answers are ranked on top
  - For example on Granite: *How does the weathering affect granite?*

- 2 traditional baselines (Indri) and 4 neural IR models (MatchZoo)
- 5 diagnostic datasets (TFC1, TFC2, M-TDC, LNC2$^{\text{Test}}$, LNC2$^{\text{All}}$)

- Retrieval effectiveness (MAP, MRR, P@5)
- Axiomatic performance on diagnostic datasets (fraction of fulfilled diagnostic instances)

Experimental **results**

- **Neural models struggle** to outperform baselines on WikiPassageQA

		MAP
1	BM25	$0.52^{3,4}$
2	QL	$\mathbf{0.54}^{1,3,4}$
3	Duet	0.25
4	MatchPyramid	0.44^{3}
5	DRMM	$0.55^{1,2,3,4}$
6	aNMM	$\mathbf{0.57}^{1,2,3,4}$

Experimental **results**

- **Neural models struggle** to outperform baselines on WikiPassageQA
- **BM25** and **QL** fulfill **many diagnostic instances**, but not all due to (a.o) **document length differences**

		MAP	**Performance per axiom**				
			$\overline{\mathrm{TFC1}}$	$\overline{\mathrm{TFC2}}$	$\overline{\mathrm{M\text{-}TDC}}$	$\overline{\mathrm{LNC2}}^{Test}$	$\overline{\mathrm{LNC2}}^{All}$
1	BM25	0.52[3,4]	0.73	**0.98**	**1.00**	**0.80**	**0.80**
2	QL	**0.54**[1,3,4]	**0.87**	0.63	**0.94**	0.68	0.68
3	Duet	0.25					
4	MatchPyramid	0.44[3]					
5	DRMM	0.55[1,2,3,4]					
6	aNMM	**0.57**[1,2,3,4]					

Experimental **results**

- **Neural models struggle** to outperform baselines on WikiPassageQA
- **BM25** and **QL** fulfill **many diagnostic instances**, but not all due to (a.o) **document length differences**
- **Stronger neural models** fulfill **more diagnostic instances**

		MAP	Performance per axiom TFC1	TFC2	M-TDC	$\overline{\text{LNC2}}^{Test}$	$\overline{\text{LNC2}}^{All}$
1	BM25	$0.52^{3,4}$	0.73	**0.98**	**1.00**	**0.80**	**0.80**
2	QL	$\mathbf{0.54}^{1,3,4}$	**0.87**	0.63	**0.94**	0.68	0.68
3	Duet	0.25	0.69	0.56	0.48	0.19	0.47
4	MatchPyramid	0.44^{3}	0.79	0.58	0.63	0.00	0.19
5	DRMM	$0.55^{1,2,3,4}$	0.84	**0.60**	**0.76**	0.05	0.12
6	aNMM	$\mathbf{0.57}^{1,2,3,4}$	**0.85**	0.56	0.69	**0.38**	**0.47**

Experimental **results**

- **Neural models struggle** to outperform baselines on WikiPassageQA
- **BM25** and **QL** fulfill **many diagnostic instances**, but not all due to (a.o) **document length differences**
- **Stronger neural models** fulfill **more diagnostic instances**
  - **TFC1:** **DRMM** and **aNMM** best at matching query terms and aggregation

	MAP	Performance per axiom				
		TFC1	TFC2	M-TDC	$\mathrm{LNC2}^{Test}$	$\mathrm{LNC2}^{All}$
1 BM25	$0.52^{3,4}$	0.73	**0.98**	**1.00**	**0.80**	**0.80**
2 QL	$\mathbf{0.54}^{1,3,4}$	**0.87**	0.63	**0.94**	0.68	0.68
3 Duet	0.25	0.69	0.56	0.48	0.19	0.47
4 MatchPyramid	0.44^{3}	0.79	0.58	0.63	0.00	0.19
5 DRMM	$0.55^{1,2,3,4}$	0.84	**0.60**	**0.76**	0.05	0.12
6 aNMM	$\mathbf{0.57}^{1,2,3,4}$	**0.85**	0.56	0.69	**0.38**	**0.47**

Experimental **results**

- **Neural models struggle** to outperform baselines on WikiPassageQA
- **BM25** and **QL** fulfill **many diagnostic instances**, but not all due to (a.o) **document length differences**
- **Stronger neural models** fulfill **more diagnostic instances**
  - **TFC1:** **DRMM** and **aNMM** best at matching query terms and aggregation
  - **TFC2:** All neural models do **not strictly follow** this heuristic

		Performance per axiom				
	MAP	TFC1	TFC2	M-TDC	LNC2^{Test}	LNC2^{All}
1 BM25	$0.52^{3,4}$	0.73	**0.98**	**1.00**	**0.80**	**0.80**
2 QL	$\mathbf{0.54}^{1,3,4}$	**0.87**	0.63	**0.94**	0.68	0.68
3 Duet	0.25	0.69	0.56	0.48	0.19	0.47
4 MatchPyramid	0.44^{3}	0.79	0.58	0.63	0.00	0.19
5 DRMM	$0.55^{1,2,3,4}$	0.84	**0.60**	**0.76**	0.05	0.12
6 aNMM	$\mathbf{0.57}^{1,2,3,4}$	**0.85**	0.56	0.69	**0.38**	**0.47**

Experimental **results**

- **Neural models struggle** to outperform baselines on WikiPassageQA
- **BM25** and **QL** fulfill **many diagnostic instances**, but not all due to (a.o) **document length differences**
- **Stronger neural models** fulfill **more diagnostic instances**
  - **TFC1:** **DRMM** and **aNMM** best at matching query terms and aggregation
  - **TFC2:** All neural models do **not strictly follow** this heuristic
  - **M-TDC:** **DRMM** best at weighing query terms, neural models underperform on IDF heuristic

		MAP	Performance per axiom: TFC1	TFC2	M-TDC	LNC2^{Test}	LNC2^{All}
1	BM25	$0.52^{3,4}$	0.73	**0.98**	**1.00**	**0.80**	**0.80**
2	QL	$\mathbf{0.54}^{1,3,4}$	**0.87**	0.63	**0.94**	0.68	0.68
3	Duet	0.25	0.69	0.56	0.48	0.19	0.47
4	MatchPyramid	0.44^{3}	0.79	0.58	0.63	0.00	0.19
5	DRMM	$0.55^{1,2,3,4}$	0.84	**0.60**	**0.76**	0.05	0.12
6	aNMM	$\mathbf{0.57}^{1,2,3,4}$	**0.85**	0.56	0.69	**0.38**	**0.47**

Experimental **results**

- **Neural models struggle** to outperform baselines on WikiPassageQA
- **BM25** and **QL** fulfill **many diagnostic instances**, but not all due to (a.o) **document length differences**
- **Stronger neural models** fulfill **more diagnostic instances**
  - **TFC1:** **DRMM** and **aNMM** best at matching query terms and aggregation
  - **TFC2:** All neural models do **not strictly follow** this heuristic
  - **M-TDC:** **DRMM** best at weighing query terms, neural models underperform on IDF heuristic
  - **LNC2:** All neural models **struggle, but can learn** to not over-penalize long documents

			Performance per axiom				
		MAP	TFC1	TFC2	M-TDC	$\mathrm{LNC2}^{Test}$	$\mathrm{LNC2}^{All}$
1	BM25	$0.52^{3,4}$	0.73	**0.98**	**1.00**	**0.80**	**0.80**
2	QL	$\mathbf{0.54}^{1,3,4}$	**0.87**	0.63	**0.94**	0.68	0.68
3	Duet	0.25	0.69	0.56	0.48	0.19	0.47
4	MatchPyramid	0.44^{3}	0.79	0.58	0.63	0.00	0.19
5	DRMM	$0.55^{1,2,3,4}$	0.84	**0.60**	**0.76**	0.05	0.12
6	aNMM	$\mathbf{0.57}^{1,2,3,4}$	**0.85**	0.56	0.69	**0.38**	**0.47**

Experimental **results**

- **Neural models struggle** to outperform baselines on WikiPassageQA
- **BM25** and **QL** fulfill **many diagnostic instances**, but not all due to (a.o) **document length differences**
- **Stronger neural models** fulfill **more diagnostic instances**
- **QL outperforms aNMM** on all axioms but is less effective: we only included **four axioms**

			Performance per axiom				
		MAP	TFC1	TFC2	M-TDC	LNC2^{Test}	LNC2^{All}
1	BM25	$0.52^{3,4}$	0.73	**0.98**	**1.00**	**0.80**	**0.80**
2	QL	$\mathbf{0.54}^{1,3,4}$	**0.87**	0.63	**0.94**	0.68	0.68
3	Duet	0.25	0.69	0.56	0.48	0.19	0.47
4	MatchPyramid	0.44^{3}	0.79	0.58	0.63	0.00	0.19
5	DRMM	$0.55^{1,2,3,4}$	0.84	**0.60**	**0.76**	0.05	0.12
6	aNMM	$\mathbf{0.57}^{1,2,3,4}$	**0.85**	0.56	0.69	**0.38**	**0.47**

Experimental **results**

- **Neural models struggle** to outperform baselines on WikiPassageQA
- **BM25** and **QL** fulfill **many diagnostic instances**, but not all due to (a.o) **document length differences**
- **Stronger neural models** fulfill **more diagnostic instances**
- **QL outperforms aNMM** on all axioms but is less effective: we only included **four axioms**
- **"Correlation"** between MAP and average axiomatic performance = **0.44** (N=6)

			Performance per axiom				
		MAP	TFC1	TFC2	M-TDC	LNC2^{Test}	LNC2^{All}
1	BM25	$0.52^{3,4}$	0.73	**0.98**	**1.00**	**0.80**	**0.80**
2	QL	$\mathbf{0.54}^{1,3,4}$	**0.87**	0.63	**0.94**	0.68	0.68
3	Duet	0.25	0.69	0.56	0.48	0.19	0.47
4	MatchPyramid	0.44^{3}	0.79	0.58	0.63	0.00	0.19
5	DRMM	$0.55^{1,2,3,4}$	0.84	**0.60**	**0.76**	0.05	0.12
6	aNMM	$\mathbf{0.57}^{1,2,3,4}$	**0.85**	0.56	0.69	**0.38**	**0.47**

Conclusions

Summary

What's next?

Conclusions

- Summary
  - Diagnostic datasets offer us a tool to **diagnose retrieval models** that is rooted in axiomatic thinking
  - The approach is **model-agnostic:** we can diagnose any IR model based on its **output**
  - There is no shortage of resources for diagnostic datasets as they **do not require relevance labels**

- Future work
  - **More** axioms, datasets (tasks), models (toolkits), ...
  - How can we “**fix**” neural models based on the axiomatic insights?

An Axiomatic Approach to Diagnosing Neural IR Models

Daniël Rennings, Felipe Moraes, Claudia Hauff
Delft University of Technology, the Netherlands

drennings@deloitte.nl