On Topic Difficulty in IR Evaluation: The Effect of Systems, Corpora, and System Components

Fabio Zampieri[†], Kevin Roitero[†], Shane Culpepper[‡], Oren Kurland[+], and Stefano Mizzaro[†]

[†]University of Udine (*Udine, Italy*), [‡]RMIT University (*Melbourne, Australia*), [+]Technion (*Haifa, Israel*)

We Thank SIGIR for Providing the Travel Grant

1. Contribution

- We address **topic difficulty** systematically
- We define **topic difficulty** as the average of AP over systems (AAP)
- We exploit **common topics** in different TREC datasets
- We exploit **common (sub-)corpora/documents** in the datasets
- We run a comprehensive set of **unofficial**, state-of-the-art, off-the-shelf systems on various TREC topics

2. Experimental Setting

Collections Used

Name	Track	Year	Topics	Official	Unofficial
T06	Ad Hoc	1997	50	74	158
T07	Ad Hoc	1998	50	103	158
T08	Ad Hoc	1999	50	129	158
R04	Robust	2004	249	110	158
C17	Common Core	2017	50	75	158

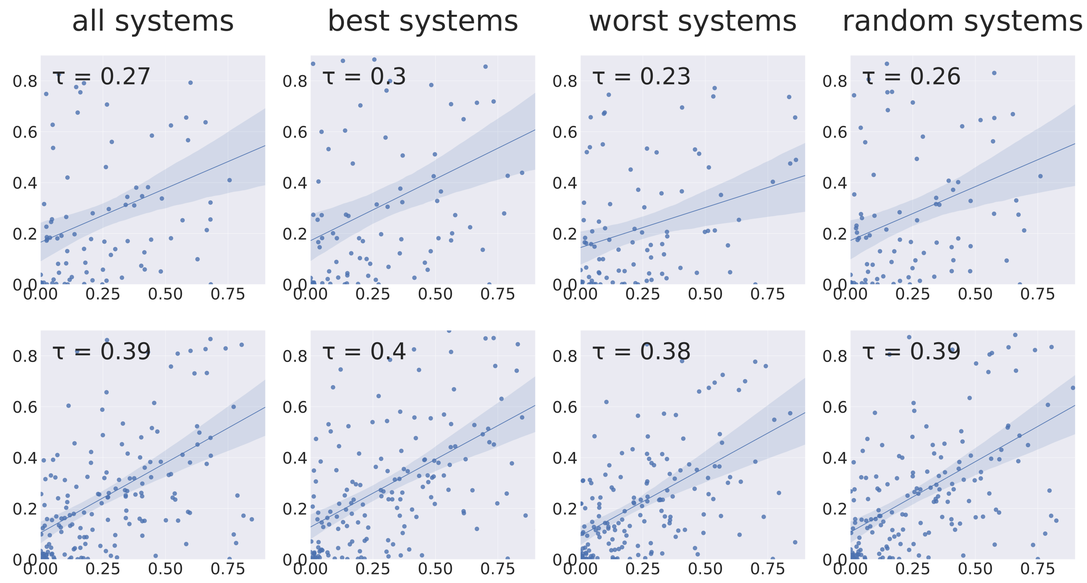
Corpora

Name	Corpus name	T06-8	R04	C17
FT	The *Financial Times*	x	x	
FR	*Federal Register*	x	x	
CR	*Congressional Record*	x		
FBIS	*FBI Service*	x	x	
NYT	The *New York Times*			x

Topic Overlap

	T06	T07	T08	C17	R04
T06	50	0	0	11	50
T07	0	50	0	17	50
T08	0	0	50	16	50
C17	11	17	16	50	50
R04	50	50	50	50	249

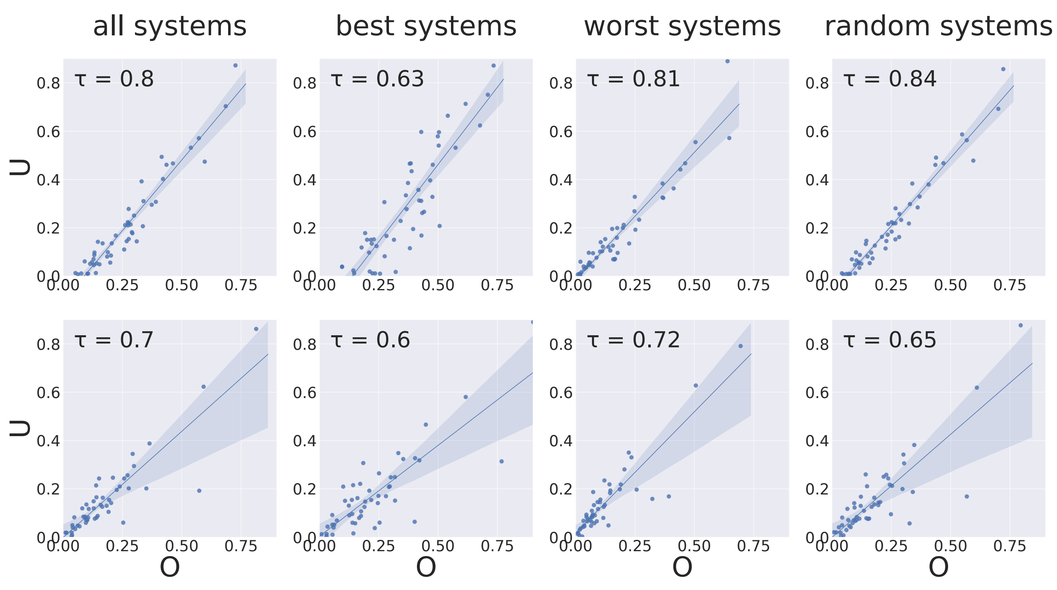
3. Corpora Effect



AAP values for: FT vs. FR (1st row), and FT vs. FBIS (2nd row). Each point is a topic.

- **Topic difficulty** is quite **sensitive** to the document corpus.
- The **correlation** of AAP values computed over different sub-collections is **very low**: the highest correlation is between AAP values computed over FT and FBIS (2nd row), while other values do not exceed 0.3

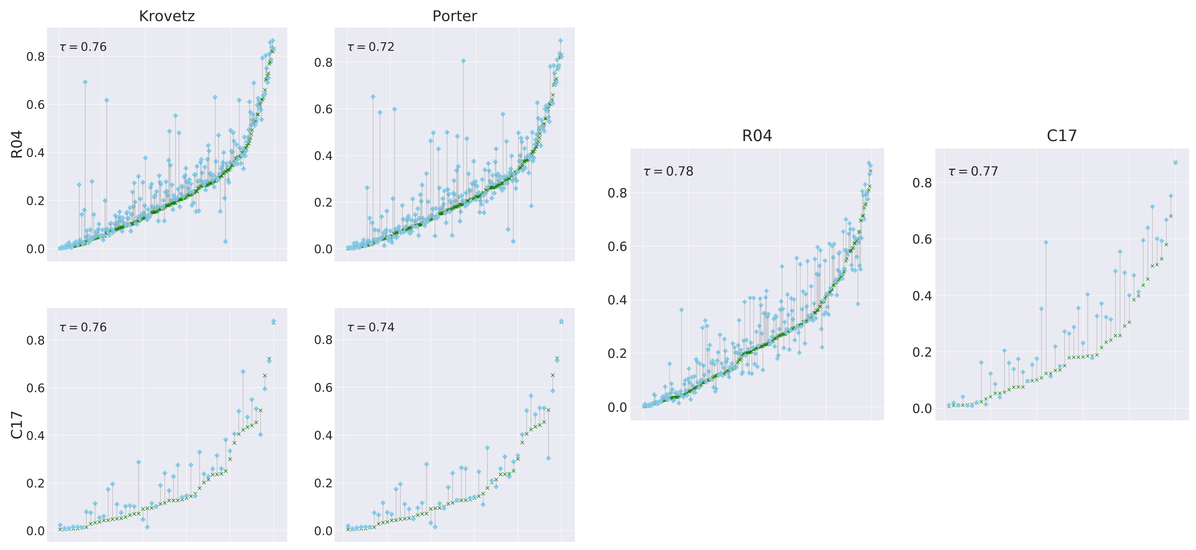
4. Systems Effect



AAP values for C17 (1st row) and R04 (2nd row), computed over different systems. Each point is a topic; x-axis: O = official systems; y-axis: U = unofficial systems.

- **High correlations** are observed in almost every case
- For a given corpus, **topic difficulty is quite stable** and does not change much across different sets of systems
- **Correlation** values **drop** when relying on the most effective systems only

5. System Components Effect



Differences in AAP computed over baselines and systems using stemmers (left), and using query expansion (right). Each point is a topic.

- For R04 (1st column), we see **frequent increases in AAP** and **infrequent decreases**
- However, for C17 (2nd column) **decreases in AAP are negligible** (the same is also true for TREC6-8, not shown)
- For a fixed subset of topics in a given collection, **topic difficulty can considerably change** if we add a stemming (left) or query expansion (right) to the set of systems

6. ANOVA – System Components

$$AP = system + topic + IRmodel + stemmer + query\ expansion + corpus + inter.$$

Factor	SS	DF	F	p-value	ω^2
corpus	15.7907	2	1133.24	< 1e-6	**0.0050**
topic	2528.42	248	1463.35	< 1e-6	**0.8157**
system	52.6792	168	45.007	< 1e-6	**0.0166**
ir_model	2.8554	22	18.6294	< 1e-6	0.0008
qe	2.0049	1	287.777	< 1e-6	0.0006
stemmer	0.3708	6	8.8723	< 1e-6	0.0001
corpus:system	5.9907	336	2.5591	< 1e-6	0.0011
corpus:qe	0.2012	2	14.4394	< 1e-6	6.045e-05

- The **topic effect** on AP scores is the **largest** (0.8157)
- The **system effect** on AP scores is significant but **small**
- Somewhat surprisingly, the **corpus interactions** have a **negligible effect** on AP scores

7. ANOVA – Effectiveness

$$AP = system + topic + corpus + corpus\text{-}system\ inter. + corpus\text{-}topic\ inter.$$

Factor	SS	DF	F	p-value	ω^2
corpus	1.5537	2	140.299	< 1e-6	**0.0003**
system	48.4639	168	52.0968	< 1e-6	**0.0103**
topic	3045.68	248	2217.86	< 1e-6	**0.6603**
corpus:topic	1120.13	496	407.84	< 1e-6	0.2423
corpus:system	6.4594	336	3.4718	< 1e-6	0.0009

- **Systems** have a **small effect** (0.0103) on AP scores, while
- **Topics** have the **greatest effect** (0.6603)
- The **interaction** effect **between corpus and topic** is also **large** but, perhaps surprisingly,
- Both the **relative effect of the corpus**, and the **interaction between corpus and system** is **negligible**

SIGIR 2019 21-25 July

Università degli Studi di Udine

RMIT University

Technion Israel Institute of Technology

SIGIR Special Interest Group on Information Retrieval