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

Fabio Zampieri
University of Udine
Udine, Italy
zampieri.fabio@spes.uniud.it

Kevin Roitero
University of Udine
Udine, Italy
roitero.kevin@spes.uniud.it

J. Shane Culpepper
RMIT University
Melbourne, Australia
shane.culpepper@rmit.edu.au

Oren Kurland
Technion
Haifa, Israel
kurland@ie.technion.ac.il

Stefano Mizzaro
University of Udine
Udine, Italy
mizzaro@uniud.it

ABSTRACT

In a test collection setting, topic difficulty can be defined as the average effectiveness of a set of systems for a topic. In this paper we study the effects on the topic difficulty of: (i) the set of retrieval systems; (ii) the underlying document corpus; and (iii) the system components. By generalizing methods recently proposed to study system component factor analysis, we perform a comprehensive analysis on topic difficulty and the relative effects of systems, corpora, and component interactions. Our findings show that corpora have the most significant effect on topic difficulty.

ACM Reference Format:
Fabio Zampieri, Kevin Roitero, J. Shane Culpepper, Oren Kurland, and Stefano Mizzaro. 2019. On Topic Difficulty in IR Evaluation: The Effect of Systems, Corpora, and System Components. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331279

1 INTRODUCTION

Topic difficulty, defined as the average effectiveness of a set of systems on a topic [9, 10], is a well-studied problem in the IR literature. It is loosely related to the problem of Query Performance Prediction (QPP), which aims to estimate the effectiveness of a system for a given query when no relevance judgments are available [2]. In classical QPP, however, the aim is to predict the performance of a specific system for a specific query; in this paper we study topic difficulty for a set of systems. This is a different problem that can be justified by the aim of understanding the "general" difficulty of a topic [7–10]. It also leads naturally to the research issue of finding representative sets of systems, i.e., sets for which difficulty would generalize to other sets. Our overall goal is to understand the effect of three factors (the set of systems, the document corpus, and the system components) on topic difficulty. To the best of our knowledge, this problem has only been investigated from a system effectiveness perspective. We achieve this goal by extending factor analysis methods recently proposed to study the effect of system components on effectiveness of systems [4–6]. We address four research questions:

RQ1. Given a collection, what is the effect of choosing a different set of systems on the difficulty of topics?
RQ2. Given a set of systems, what is the effect of the corpus of documents (or sub-corpora of a corpus) on topic difficulty?
RQ3. What is the effect of system components on topic difficulty?
RQ4. What is the relative effect of choosing different systems, corpora, and system components on topic difficulty?

2 RELATED WORK

A body of related work focuses on studying factors that affect system effectiveness, such as topic composition, collection, and system components. Sanderson et al. [11] investigated the effect of splitting a TREC collection into sub-collections based on system effectiveness, and identified several interesting sub-collection effects induced by the splits. Banks et al. [1] provided an overview of methods that can be applied to analyze the performance of IR systems on TREC collections and its relation to topics, collections and other factors. One common statistical tool used for this problem is the *Analysis of Variance* (ANOVA), which was recently used by Ferro and Silvello [5] to compare combinations of collections, metrics, and systems. They showed that stop lists, IR models, and component interactions have a significant but small effect on overall system effectiveness. The same approach was adopted by Ferro and Sanderson [4] and Ferro et al. [3], whose experiments show the existence of a significant sub-corpus effect relative to system effectiveness; however, the effect is smaller than both system and topic effects, with topic effect being the most significant. Similar experiments using the sub-corpora of a single collection showed that the system effect is smaller than the topic effect [4]. However, none of these studies specifically addresses the effect of factors on topic difficulty which we study here. Moreover, all of them compare sub-corpora of the same collection, which has some drawbacks. TREC corpora are built with a "working assumption" that they are somehow complete, and working on sub-corpora can sometimes negate this assumption. In this paper, we do not only analyze what happens on incomplete sub-corpora, but we are also able to compare across different corpora.

3 EXPERIMENTS

3.1 Experimental Setting

Datasets. Table 1 summarizes the datasets used for our experiments. We focus on five TREC (Text REtrieval Conference) collections. Our datasets are purposefully chosen to include overlapping sets of topics, systems, and corpora. The set of R04 topics includes TREC6 topics (301-350), TREC7 topics (351-400), TREC8 topics (401-450), half of the Robust03 topics (601-650), and 50 additional topics

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331279

Table 1: Datasets used in our experiments.

Acronym	Track	Year	Topics	Official	Unofficial
TREC6	Ad Hoc	1997	50	74	158
TREC7	Ad Hoc	1998	50	103	158
TREC8	Ad Hoc	1999	50	129	158
R04	Robust	2004	249	110	158
C17	Common Core	2017	50	75	158

Table 2: The number of common topics between collections.

	R04	C17	TREC6	TREC7	TREC8
C17	50	50			
TREC6	50	11	50		
TREC7	50	17	0	50	
TREC8	50	16	0	0	50

Table 3: Corpora of documents used in the datasets.

Acronym	Corpus name	TREC6-8	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

that were specifically introduced in R04. C17 has 50 topics, which were also originally included in the R04 set of topics; C17 has a few topics that overlap with TREC6-8 (see Table 2). Table 3 shows the document corpora used in each collection: R04 and TREC6-8 share, apart from C17, the same corpora; C17 is based only on NYT.

For each of the TREC collections we use the officially-submitted runs. We also supplement available runs using several open source search engines in order to produce system configurations that are directly comparable across collections: Terrier, Atire, and Indri (www.terrier.org, www.atire.org, www.lemurproject.org). The 158 system variants are generated by systematically alternating and combining the ranker, stemmer, and stopword configurations, but fixing configurations to be identical across all test collections. Henceforth we distinguish between official systems/runs (O) from TREC, and unofficial system configurations (U) generated by us. Both O and U systems produce ranked lists of 1000 documents.

Metrics. We use *Average Precision* (AP) as an effectiveness measure. Given a system s_i and a topic t_j, we denote the corresponding score which is a real number between 0 and 1 as $AP(s_i, t_j)$. By averaging the AP values over each topic, we obtain the *Average AP* (AAP), a measure of topic difficulty [9, 10]: $AAP(t_j) = \frac{1}{m} \sum_{i=1}^{m} AP(s_i, t_j)$. A high AAP value indicates that the topic is easy, and a low AAP indicates that the topic is difficult for a specific collection and set of system runs. We use Kendall's τ as the primary correlation coefficient in this work, as it is well-suited to compute partial correlations in fully-ranked data [1].

3.2 Results

RQ1: System Effects. We first illustrate and discuss how topic difficulty changes when we select a different set of systems. In Figure 1, scatter plots of AAP values for R04 and C17 topics are shown; the other collections, not shown due to space limits, exhibit similar trends. Columns correspond to subsets of systems, each containing 30 elements (with the exception of the first column, which represents the set of all systems), while rows correspond

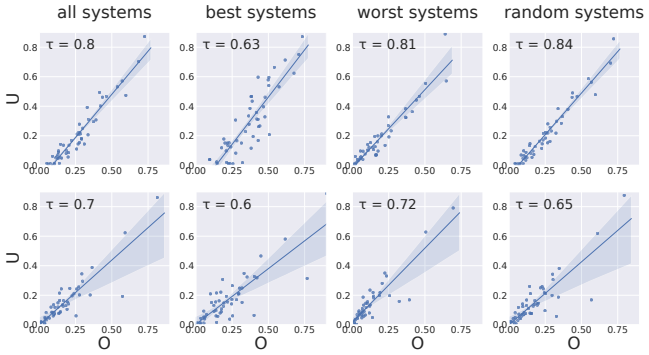

Figure 1: Scatterplots of AAP values for C17 (first row) and R04 (second row), computed over different sets of systems (y-axis: U = Unofficial; x-axis: O = Official).

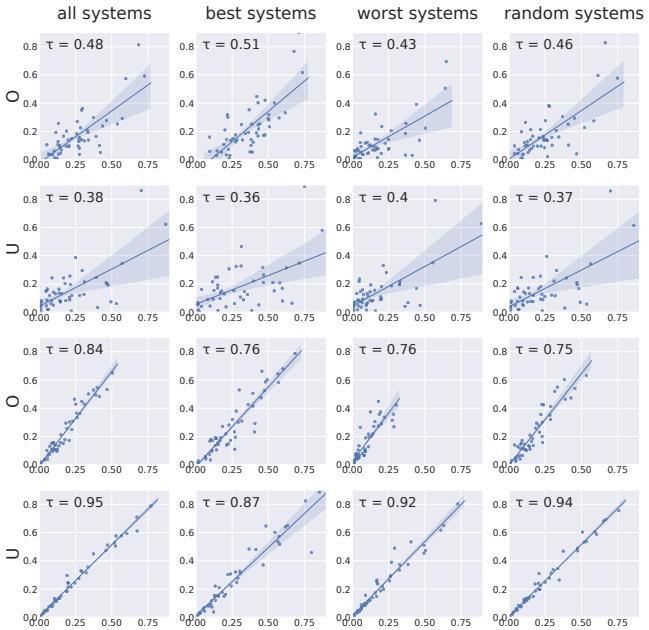

Figure 2: Scatterplots of AAP values computed over R04 vs. C17 (first two rows), and R04 vs. TREC6 (3rd and 4th rows), using either the official (O) runs (1st and 3rd row) or the unofficial (U) ones.

to collections. For each plot, a point is defined by the AAP value computed over the set of official systems (on the x axis) and the AAP value computed over the set of unofficial systems (on the y axis). High correlations are observed in almost every case. Selecting a particular group of systems does not seem to affect the correlation, even though a significant overall drop can be seen when τ values are computed using only the best systems (i.e., the 30 best official and the 30 best unofficial). Therefore, for a given corpus, topic difficulty seems quite stable and does not appear to change much across different sets of systems, although they heavily differ in terms of implementation and components. The correlation values drop, however, when relying only on the most effective systems.

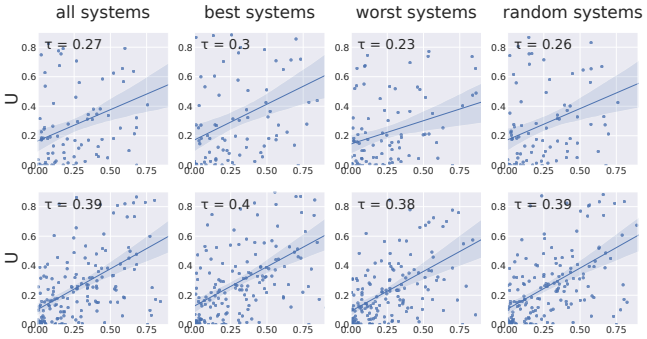

Figure 3: Scatterplots of AAP values computed over R04 sub-collections: FT vs FR (1st row) and FT vs. FBIS (2nd row).

RQ2: Corpora Effects. We now turn to the effect of document corpora on topic difficulty. In Figure 2, we see that the τ correlation between AAP values of R04 and C17 is 0.48 for official systems (1st row, 1st column), and 0.38 for unofficial ones (2nd row, 1st column). It is somewhat higher for official systems, although they differ across collections whereas the unofficial configurations are identical. Similar results are observed when selecting a particular subset of systems (columns 2-4). In contrast, the correlations between R04 and TREC6 are very high: 0.84 when computed over official systems (3rd row, 1st column), and 0.95 when computed over unofficial systems (4th row, 1st column). Also in this case, selecting a subset of systems does not seem to affect correlations. We obtained the same results for TREC7-8 (not shown here).

As R04 and C17 include different document corpora (see Table 3), these results suggest that topic difficulty is indeed quite sensitive to the document corpus. When comparing these results to previous work [3, 4], we observe two differences: only sub-corpora were used, not different corpora as we do here, and system effectiveness was studied, not topic difficulty as we do here.

Figure 3 provides also evidence to sub-corpora effects over R04. It shows how topic difficulty changes across the sub-corpora of R04 (shown in Table 3). Here again, 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.

To summarize: (i) we find very low correlations when changing significantly the corpus (R04 vs. C17), thereby generalizing the finding about low correlations on different sub-corpora also to the case of different complete corpora; (ii) in one case (R04 vs. C17), we find the strange result that computing AAP using the same unofficial system set leads to lower correlation than when using the official—and different—system set; but this is not confirmed on other datasets; finally (iii) if the changes to the corpus are small (R04 vs. TREC6) then correlations are high.

RQ3: System Component Effects. We now turn to our third research question, which focuses on the impact of system components on topic difficulty; in particular, we consider stemming and query expansion. Since these are quite dramatic changes to the systems, we expect quite significant changes to AAP values, and probably low correlations. Figure 4 shows, for each topic in the R04 and C17 collections, the difference of AAP values computed over the baselines (i.e., systems without stemmer and query expansion) and

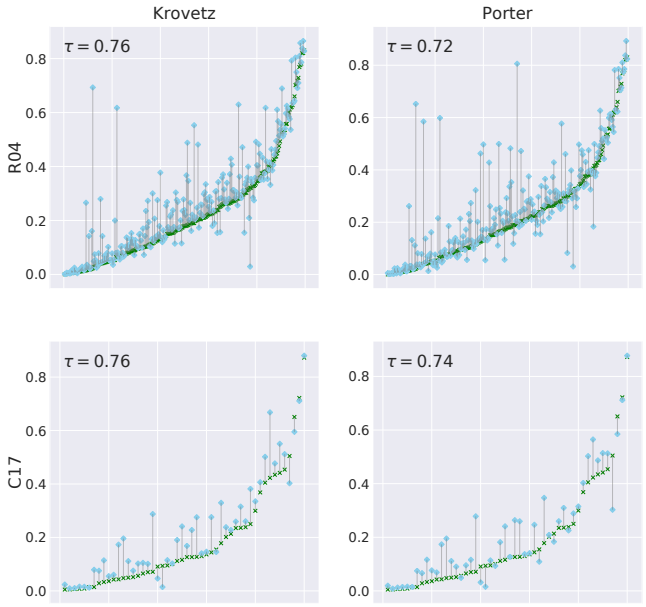

Figure 4: Differences between AAP values computed over baselines (i.e., systems without stemmer and query expansion) and those computed over systems using stemmers.

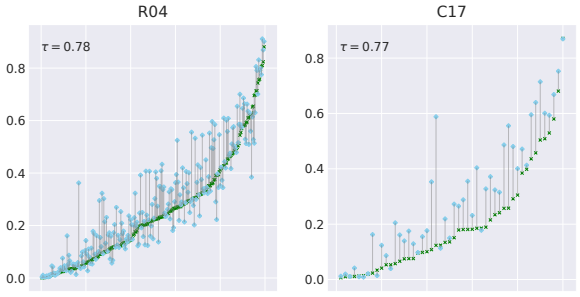

Figure 5: Differences in AAP computed over baselines and over systems using query expansion.

when using two common stemmers (Krovetz and Porter). Due to space limitations, we do not show the results for the all stemmer and collection combinations. For many of the topics, stemming leads to little or no significant improvement in terms of AAP. In a few cases, however, there are significant increases and decreases in AAP, which occur for the same topics across different stemmers. The highest differences in AAP was observed for the R04 topics (see the 1st row), which appear to be quite sensitive to the stemmer used.

Figure 5 shows the AAP differences between the baselines and systems using query expansion for R04 and C17. 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).

The results show that system components can have variable effects on topic difficulty. In particular, we see that, for a fixed subset of topics in a given collection, topic difficulty can considerably change if we add a stemming or query expansion to the set of

Table 4: ANOVA table for the model described by Eq. 1.

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. However, the τ correlations, shown in Figures 4 and 5, are quite high: somehow unexpectedly, relative topic difficulty remains quite stable despite the changes to the systems (stemming or query expansion) are quite significant.

RQ4: Comparing relative effects with ANOVA. In an attempt to provide a more principled and, at the same time, concise analysis, we investigate the effects of systems, corpora, and system components using ANOVA as part of our final research question. In particular, we define two ANOVA models (see Equations (1) and (2)), which are described below. Tables 4 and 5 show the outcome of each ANOVA test. For each factor, we report the Sum of Squares (SS), the Degrees of Freedom (DF), the F statistics, the p-value, and the effect-size (ω^2) which quantifies the proportional variance of each factor [4–6]. The first model decomposes the effectiveness (measured by AP) into system, topic, and corpus effects:

$$AP(i,j) = \mu + s_i + t_j + c_z + c_z s_i + c_z t_j + \epsilon_{ij} \quad (1)$$

where terms identify $AP(i,j)$ of i-th system and j-th topic, grand mean (μ), z-th corpus (c_z), corpus-system ($c_z s_i$) and corpus-topic ($c_z t_j$) interactions, and model error (ϵ_{ij}). Table 4 shows the results of the ANOVA analysis for Eq. (1). All effects are statistically significant. Systems have a small effect (0.0103), 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. The second model focuses on system components:

$$AP(i,j) = \mu + s_i + t_j + mo_q + st_k + qe_y + c_z + c_z s_i + c_z t_j + \epsilon_{ij} \quad (2)$$

where terms identify IR model (mo_q), stemmer (st_k), query expansion (qe_y), corpus-system ($c_z s_i$) and corpus-topic ($c_z t_j$) interactions. The results of the ANOVA test for Eq. (2) are shown in Table 5. All effects are statistically significant, and the topic effect is the largest (0.8157); the system effect is significant but small. Again, somewhat surprisingly, the corpus interactions have a negligible effect on AP scores. All other effects are not significant. In summary, the ANOVA analyses show that AP scores are affected mostly by topics and systems, with the greatest effects being attributable to the topic effect; furthermore, system components, corpus, and the interaction between corpus and systems have very little effect on AP. Nevertheless, the impact of topics on AP clearly varies based on the corpus.

4 CONCLUSIONS AND FUTURE WORK

This is the first study that specifically addresses topic difficulty in a systematic way: we use different corpora, not just sub-corpora; we run the same set of systems across different datasets; and we rely on datasets featuring common topics. To do so, we exploit the topic overlap between C17 and R04 with previous collections, and we supplement our analysis using a comprehensive set of unofficial but reproducible systems.

Table 5: ANOVA table for the model described by Eq. 2.

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

We find that topic difficulty is affected by the document corpora of collections: there is a significant corpus-effect on topic difficulty in all of the collections tested. Also, there is a significant system-effect, although not so large. Finally, we see a smaller effect of system components on topic difficulty, with the exception of a few limited cases. Although the standard ANOVA analysis shows a strong variance across topics and system effects that are higher than the corpus effects, we alsof find that topic difficulty is reasonably stable across system sets and system components, thus confirming that it is a reasonable and measurable concept. We found only two exceptions with low correlations: the comparison across the different corpora of R04 and C17 and the comparison across R04 sub-corpora (Figures 2 and 3). Although the latter might be due to the incomplete nature of sub-corpora, the former confirms that topic difficulty is mostly affected by the underlying document collection.

In the future we plan to extend the analysis to more collections, to fine-tune the parameters of the unofficial systems to each dataset, and to study more system and topic components.

Acknowledgements. This work was partially supported by the Israel Science Foundation (grant no. 1136/17), the Australian Research Council's *Discovery Projects* Scheme (DP170102231), a Google Faculty Award, and an Amazon Research Award.

REFERENCES

[1] David Banks, Paul Over, and Nien-Fan Zhang. 1999. Blind men and elephants: Six approaches to TREC data. *Information Retrieval* 1, 1 (1999), 7–34.
[2] David Carmel and Elad Yom-Tov. 2010. Estimating the query difficulty for information retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 2, 1 (2010), 1–89.
[3] Nicola Ferro, Yubin Kim, and Mark Sanderson. 2019. Using Collection Shards to Study Retrieval Performance Effect Sizes. *ACM TOIS* 5, 44 (2019), 59.
[4] Nicola Ferro and Mark Sanderson. 2017. Sub-corpora impact on system effectiveness. In *Proceedings of the 40th ACM SIGIR*. ACM, 901–904.
[5] Nicola Ferro and Gianmaria Silvello. 2016. A general linear mixed models approach to study system component effects. In *39th ACM SIGIR*. 25–34.
[6] Nicola Ferro and Gianmaria Silvello. 2018. Toward an anatomy of IR system component performances. *JASIST* 69, 2 (2018), 187–200.
[7] Donna Harman and Chris Buckley. 2009. Overview of the reliable information access workshop. *Information Retrieval* 12, 6 (2009), 615–641.
[8] Stefano Mizzaro, Josiane Mothe, Kevin Roitero, and Md Zia Ullah. 2018. Query Performance Prediction and Effectiveness Evaluation Without Relevance Judgments: Two Sides of the Same Coin. In *The 41st ACM SIGIR (SIGIR '18)*. 1233–1236.
[9] Stefano Mizzaro and Stephen Robertson. 2007. Hits Hits TREC: Exploring IR Evaluation Results with Network Analysis. In *Proceedings 30th SIGIR*. 479–486.
[10] Kevin Roitero, Eddy Maddalena, and Stefano Mizzaro. [n. d.]. Do Easy Topics Predict Effectiveness Better Than Difficult Topics?. In *ECIR2017*. 605–611.
[11] Mark Sanderson, Andrew Turpin, Ying Zhang, and Falk Scholer. 2012. Differences in effectiveness across sub-collections. In *Proc. of the 21st ACM CIKM*. 1965–1969.