On Transforming Relevance Scales

Lei Han[+], Kevin Roitero[*], Eddy Maddalena[△], Stefano Mizzaro[*], and Gianluca Demartini[+]
[+]University of Queensland, Australia. [*]University of Udine, Italy. [△]University of Southampton, UK.

ABSTRACT

Information Retrieval (IR) researchers have often used existing IR evaluation collections and transformed the relevance scale in which judgments have been collected, e.g., to use metrics that assume binary judgments like Mean Average Precision. Such scale transformations are often arbitrary (e.g., 0,1 mapped to 0 and 2,3 mapped to 1) and it is assumed that they have no impact on the results of IR evaluation. Moreover, the use of crowdsourcing to collect relevance judgments has become a standard methodology. When designing the crowdsourcing relevance judgment task, one of the decision to be made is the how granular the relevance scale used to collect judgments should be. Such decision has then repercussions on the metrics used to measure IR system effectiveness.

In this paper we look at the effect of scale transformations in a systematic way. We perform extensive experiments to study the transformation of judgments from fine-grained to coarse-grained. We use different relevance judgments expressed on different relevance scales and either expressed by expert annotators or collected by means of crowdsourcing. The objective is to understand the impact of relevance scale transformations on IR evaluation outcomes and to draw conclusions on how to best transform judgments into a different scale, when necessary.

KEYWORDS

Crowdsourcing, IR Evaluation, Assessor Agreement, Relevance Scales

ACM Reference Format:
Lei Han, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2019. On Transforming Relevance Scales. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3357384.3357988

1 WHY TRANSFORMING RELEVANCE SCALES?

One of the design decision to make when creating an Information Retrieval (IR) evaluation collection is which *relevance scale* to use for judgments. While, historically, binary relevance judgments have usually been collected, and metrics based on binary relevance like, e.g., precision, recall, and average precision, have been used, more recently multi-level relevance scales have become popular. There are several reasons for this. One reason is the availability of gain-based evaluation metrics like NDCG [10], which are defined also for non-binary relevance judgments.

Another reason can be found in the increasingly popular use of Crowdsourcing, which has become a standard methodology to create relevance labels for IR test collections. One of the main challenges of the use of crowdsourcing to collect relevance judgments is quality. To address this issue a number of approaches have been proposed including task design methods [14, 22] as well as multiple assignments of the same topic-document pair in order to aggregate judgments from different workers to improve the overall judgment quality [5, 25]. Some of the existing aggregation functions produce a value that is still in the original scale; some others like, e.g., arithmetic mean or median, can produce an aggregated value that does not belong to the original scale, thereby obtaining more fine-grained relevance values. Following this trend, recent research has specifically and deliberately looked at the effect of unbounded and fine-grained relevance scales, thereby producing relevance labels on scales that have 101 values [21] or are even unbounded [15, 24].

The existence of IR evaluation collections using different relevance scales does not allow for possible comparisons of evaluation results or merging of collections. Additionally, the evaluation of IR systems is limited to the relevance scale used by the specific test collection. This leads to a natural research question: *how to transform one scale into another?* Indeed, quite often in the past, IR researchers have transformed relevance judgments collected on a multi-level scale into a coarser-grained scale, e.g., from a 4-level (highly relevant, relevant, marginally relevant, and not relevant) scale to binary for the sake of using IR evaluation metrics that require binary judgments (e.g, Average Precision, Mean Reciprocal Rank, etc.) [8, 11, 12, 17, 20]. More in general, whenever an evaluation metric requires a coarse-grained scale, the relevance judgments collected on a fine-grained scale need to be transformed.

In these examples, researchers have arbitrarily defined certain scale *transformation thresholds* to map relevance judgments from a *source scale* (e.g., 4-levels labelled as 0, 1, 2, and 3) into a *target scale* (e.g., binary with levels labelled 0 and 1), for example, by transforming 0 and 1 judgments in the source scale as 0 in the target scale and 2 and 3 in the source scale as 1 in the target scale, and doing so for all topics in the collection. Another solution has been to transform 0 into 0 and 1, 2, 3 into 1. These approaches were named "rigid" and "relaxed", respectively, in early NTCIR editions (e.g., NTCIR-4 CLIR task) [12]. Thus, the problem is defined as finding the best thresholds in the source scale (which we call *cuts* in this paper) that allow us to then transform judgments from the source to the target scale.

We believe that a more principled approach is needed, and possible. The cuts should not be selected arbitrarily but rather on the basis of more informed decisions. For example, the same cut has been typically used uniformly across all topics, but different cuts for different topics (for example, based on the number of relevant documents available for that topic) may be more accurate. Even more,

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3357988

different cuts might be needed for different assessors. Indeed, there seems to be some evidence that the "one-cut-fits-all" approaches are inadequate since different topics show different relevance profiles [15, 24]. To the best of our knowledge, such a systematic study of scale transformations has never been performed so far.

Besides aiming at better understanding scale transformations, we also mention that another reason for this research is to save resources: in the above cited large scale crowdsourcing experiments [15, 21, 24], we estimate that gathering relevance judgments costed about $100 for each topic considering 129 systems and a pool depth of 10. If fine-grained to coarse-grained scale transformations were available and reliable, an experimenter could collect the fine-grained relevance labels only, and rely on scale transformations when coarse-grained labels are needed. Equivalently, one could run all the experiments anyway but scale transformations would allow to gather more reliable data with the same resources (i.e., money, time to design and run the experiments, data cleaning, etc.).

Finally, although in this paper we focus on relevance judgments, the same issue might be found in other scenarios like fake news detection (where, for example, six levels have been used [27] but it seems reasonable to use accuracy as an effectiveness measure, which requires binary values), sentiment classification where 5-star ratings are often transformed in three classes [1] or binarized in an ad-hoc manner (e.g., reviews with more than 3 stars are considered as positive and less than 3 stars as negative [19]), and many other classification-like problems. Our results might be useful also for those cases. While we do not address the last two issues in this paper, we rather present an in-depth analysis of the effects of relevance scale transformations on IR evaluation and present a comprehensive set of transformation approaches for crowdsourced relevance judgments systematically observing the effects on IR evaluation results. We also show that judgments collected for different topics are best transformed using different cuts in order to maximize assessor agreement and IR system ranking correlation. The main contribution of this paper is an understanding of the effects of transforming relevance judgments into a more coarse-grained scale than that used to originally collect the judgments and a set of guidelines on how to best select cuts to transform judgments into a different scale.

Thus, the research questions we focus on are the following:

(1) How to transform relevance scales with target scale data?
(2) How to transform scales without target scale data?
(3) Is there any difference when transforming expert or crowdsourced judgments?
(4) Should we transform scales differently for each topic or in the same way for an entire judgement collection?
(5) What is the effect of assuming unjudged documents as not-relevant on transforming scales?
(6) Which scale transformation method should we adopt to obtain IR evaluation results more similar to when using expert judgments?

2 RELATED WORK

Although different relevance scales have been used to create IR evaluation collections for some years, they have not been studied in much detail and with a systematic approach until recently. Most of the past IR research has adopted some kind "quick-and-dirty" scale transformation (usually from a few levels to two), without deep analyses of the appropriateness and implications of such transformations. In this section we provide an overview of different research around the use of relevance scales in the creation of IR evaluation collections.

2.1 Relevance Scales in Evaluation Initiatives

Current IR evaluation practice is based on the early work of Cleverdon in their Cranfield experiments where they defined a systematic process to evaluate and compare the effectiveness of IR systems [4]. Starting in 1992, TREC has been building standard IR evaluation collections following the Cranfield paradigm using a document collection, a set of search topics, and a collection of relevance judgments using a specific relevance scale [9]. Early TREC collections all used a binary relevance scale. Starting in the 2000s, TREC started to use multi-level relevance scale and evaluation metrics that support such judgments. Also in the early 2000s, other IR evaluation initiatives started to use multi-level relevance scales. For example, NTCIR-3 [7] collected relevance judgments on 4 levels and then used classic evaluation metrics that required binary judgments. To transform judgments into binary, they used all three possible cuts. In some later NTCIR initiatives [11, 12] 4 levels of relevance (i.e., H, R, P, and N) have been used and only 2 cuts have been considered (i.e., rigid HR->1 and relaxed HRP->1). In NTCIR-4 [20] organizers used only 3 levels and than binarized in both possible ways.

2.2 Relevance Scales in Evaluation Measures

From an overview of IR evaluation measures and of whether they assume a binary or non-binary scale [6], we can observe that early measures all assumed binary relevance judgements as pioneered by Cleverdon. Only around the 1990s the first measures assuming multi-level relevance have been proposed, although without much application in practice. Evaluation measures that use non-binary relevance judgements were proposed even before multi-level relevance scale collections were created. For example, *ndpm* [28] considers users expressing preferences for one document over another and thus possibly generating relevance classes of judged documents or even a full ranked list of documents by relevance. Only in the 2000s measures supporting multi-level relevance (e.g., NDCG [10] and RBP [18]) have become popular and used in practice.

2.3 Fine-grained Relevance Scales

More recently, systematic studies of the impact of different relevance scale may have on the collected relevance judgments and on the results of IR Evaluation have been carried on. For example, Maddalena et al. [15], Turpin et al. [24] looked at the possibility to collect judgments from a crowdsourcing platform using an unbounded relevance scale, based on the concept of Magnitude Estimation (ME) which allows to collect judgments in the $]0, +\infty[$ range. They have shown that in a crowdsourcing setting such a scale gives more flexibility to assessors to use their own preferred scale (e.g., a worker may judge relevance using only values in the $]0, 5]$ range while another may judge relevance in $]0, 10]$) also allowing workers to always go higher or lower in comparison to the judgments already given to previous documents. Later, Roitero et al. [21] compared the ME scale to a fine-grained but finite scale of 101 levels showing how such a scale carries both the benefits and flexibility of a scale like ME allowing crowd workers to pick their favourite way to judge relevance as well as having the advantages of a finite scale

Table 1: The five datasets we use in this work.

	2-levels	4-levels	101-levels
Expert	TR2 (TREC-8)	So4 (Sormunen)	
Crowd	S2	S4	S100

thus allowing for better cross-topic comparison of the judgements and of the IR system effectiveness evaluation.

3 EXPERIMENTAL SETTING

In this section we describe the datasets, the measures, and the transformation methods used in our experiments. We also briefly discuss the effect of unjudged documents.

3.1 Relevance Judgment Datasets

Aiming at investigating the effect of scale transformations, we identified a set of 18 search topics from TREC-8 [26] judged by NIST experts using a binary scale. Some of the documents retrieved for such topics were subsequently re-judged by Sormunen [23] on a 4-level ordinal relevance scale: `N`–not relevant (0); `M`–marginally relevant (1); `R`–relevant (2); `H`–highly relevant (3). Then, Roitero et al. [21] ran a crowdsourcing re-assessing exercise using a 101-level ordinal relevance scale. Such a crowdsourced reassessment produced a total of 4269 topic-document judgments, of which 90.9% have binary TREC relevance judgments available, and 18.9% have Sormunen 4-level ordinal judgments available. The differences in the overlap of judged documents among these collections are due to the different sampling strategies adopted: the 4-level collection [23] contains only a sample of the documents judged by TREC, constructed by skipping many documents already judged as not relevant by TREC assessors; the 101-level collection contains judgments for all documents retrieved in the first 10 ranking positions by at least one system; and in TREC-8 some systems did not contributed to the pool (so they might have unjudged documents in the first ranking positions). Additionally to these datasets, in this work we also use new crowdsourced re-assessments of the same documents used by Roitero et al. [21] using both a binary and a 4-level scale.

Thus, in summary (see Table 1), we use two expert-generated collections: one generated by NIST assessors for TREC-8 using a binary scale (in the following we refer to this data set with TR2, for TREC and binary) and one generated by assessors used for [23] using a 4-level scale (So4). We also use three crowdsourced collections: one generated using a 101-level scale (S100) where workers assign values by a slider bar in the same way as in [21], one using a 4-level scale (S4), i.e., the same one used in [23], where workers choose the values by radio buttons, and one using a binary scale (S2). In the crowdsourced collections we use in this work, each crowd worker had to judge 8 documents for a single topic in a HIT (Human Intelligence Task, the basic unit of work to be performed by a crowd worker) and each document has been independently judged by 10 different crowd workers to be able to aggregate their judgements and improve the dataset quality. The three crowdsourced datasets contain exactly the same 4269 topic-document pairs and only differ for workers who completed the HITs and for the relevance scale used to collect the judgments. The crowdsourced data collections have received ethics approval from the review board by the authors' institutions.

3.2 Measuring the Similarity of Relevance Judgment Sets

When transforming a judgment set into a different target scale, we are able to use the transformed judgments to evaluate IR system effectiveness and to compare the evaluation results with those obtained using the original judgment set. We can do this by means of Kendall's τ correlation between the IR system rankings generated using the two evaluation sets. Another method to compare the original relevance judgments with those transformed to the target scale is to rely on assessor agreement measures. Using all judgments in a crowdsourced IR collection, we measure the agreement among different assessors who contribute to the judgment set, which we define as *internal agreement*. When using this approach to compare original and transformed judgment sets, we can measure how much the judgments transformed in the target scale agree among themselves as compared to the internal agreement of the judgments in the original scale. When another dataset collected in the target scale is available, we can also define *external agreement* by measuring assessor agreement between the transformed judgments and the ones collected natively in the target scale, to check how the transformed judgments align with the (expert or crowd) judgments in the other available target-scale dataset. Note that by using internal agreement we identify the best cut that maximizes the agreement of the transformed judgments with themselves, while by using external agreement we select the best cut where the transformed judgment set is the closest to the one obtained in the target scale w.r.t. judgment quality. In the following section we report our experimental results comparing different transformation methods in terms of assessor agreement using Krippendorff's [13] α and in terms of IR system effectiveness evaluation using τ.

As shown in previous research [16, 21], there is large variance in assessor agreement values across different topics; thus, we measure per-topic agreement to perform per-topic cuts and transformations. This approach is unconventional for the classical IR evaluation setting, where a certain relevance cutoff is chosen a priori to be the same over all the topics in the collection. Note, however, that a per-topic approach does not create issues in performing IR evaluation. For example, NDCG values for different topics would still be comparable despite the fact that they originate from topics with different cuts since the relevance scale and gain values across topics are the same, and due to the fact that the computed gains are normalized. Different cuts over different topics may affect the number of relevant documents in each topic, which is anyway something that varies considerably across topics in IR evaluation collections [18]. Indeed, per-topic cuts may help reducing the variance in the number of relevant documents across topics.

3.3 Scale Transformation Methods

Given the datasets used in this work (see Table 1), we perform the following transformations of relevance judgment datasets into a target scale: So4 into binary, S4 into binary, S100 into binary, and S100 into 4 relevance levels. Note that we only perform transformations from fine-grained scales to coarse-grained scale as the opposite would require new information not available in the source dataset.

We distinguish two main classes of approaches to transform the relevance scale used by a collection of judgments. In the former only the judgment set to be transformed is available (in the source scale); in this case we use internal agreement (see Section 3.2). In the

latter scenario both the set to be transformed and a set of judgments created in the target scale (either by experts or crowdsourced) are available, and therefore we use external agreement (see Section 3.2).

3.3.1 Single Dataset Scale Transformation. To transform a crowdsourced judgment set into a target scale we need to select one of the possible cuts (e.g., to transform a 4-level judgment set into binary we have three possible choices). There are different possible approaches we can follow to decide on the best *cuts* for our crowdsourced judgment sets; for each of them we provide a long name, a short one (defined more in detail in Table 2), and a description:

HIT-centric, transform then aggregate (H_t+a^1). Given all documents judged by an individual crowd worker, we can first transform each individual judgments into the target scale (using one of the possible cuts) and then do the same for the other 9 judgments collected from the crowd for each document. We can then aggregate the 9 judgments for the same document (e.g., using majority vote or another aggregation function) thus obtaining two transformed judgments in the target scale: the individual worker judgment and the crowd aggregated judgment. This allows us to compute the Krippendorf's α agreement between an individual worker with respect to rest of the crowd. By computing this version of inter-annotator agreement for all possible cuts, we are able to identify the cut which maximizes the α value, in order to keep the highest judgment agreement.

HIT-centric, aggregate then transform (H_a+t^1). As variant of the previous approach, we first aggregate the 9 crowd judgments for the same document in the source scale and then transform both the individual worker judgment and the crowd aggregated judgment into the target scale to compute α agreement for a specific cut.

Topic-wide α (Tw_α^1). A third approach is to compute α on the entire worker-document matrix of judgments for a topic transformed in the target scale for each possible cut to find the one that maximizes α. Note that this method has a potential issue given by the fact that α is not a very reliable measure for sparse matrices [3].

3.3.2 Double Dataset Scale Transformation. In this context, we assume that both a judgment set in the source scale and one in the target scale are available. For example, besides a fine-grained crowdsourced judgment set, we could also have expert judgments collected in the target scale available (e.g., TR2 binary judgments) which we might leverage to better decide on the best cut to be used on the source scale dataset (e.g., S4). In this case, the possible scale transformation approaches are:

HIT-centric (H^2). All 8 documents judged by an individual crowd worker in a HIT are first transformed into the target scale using a certain cut. Then, the transformed judgments are compared to the second dataset (that was created using the target scale) to compute the α agreement value for a given worker and cut. We then make an average of α values over all workers contributing judgments for a certain topic and, thus, are able to identify the best cut using the highest α value.

Document-centric aggregate then transform (D_a+t^2). We first aggregate all relevance judgments collected for a document (e.g., 10 in our experimental design) in the source scale and then transform the aggregated document judgments into the target scale using a certain cut and we then measure the α agreement score for the specific cut. The best cut is identified using the highest α value.

Table 2: The 27 possible scale transformation methods either using a single crowd dataset in the source scale or using two datasets. We then apply these methods to perform three transformations: S4 into binary, S100 into binary and S100 into 4-level. Notation: Hit-centric (H), Document-centric (D), Topic-wide (Tw$_\alpha$), Aggregate (a), Transform (t), Single dataset (1), Double dataset (2) as superscript.

Single dataset (by internal agreement)		
H_t+a^1(S4→2)	H_t+a^1(S100→2)	H_t+a^1(S100→4)
H_a+t^1(S4→2)	H_a+t^1(S100→2)	H_a+t^1(S100→4)
Tw_α^1(S4→2)	Tw_α^1(S100→2)	Tw_α^1(S100→4)
Double dataset (by external agreement)		
H^2(S4→2, TR2)	H^2(S100→2, TR2)	H^2(S100→4, So4)
H^2(S4→2, S2)	H^2(S100→2, S2)	H^2(S100→4, S4)
D_a+t^2(S4→2, TR2)	D_a+t^2(S100→2, TR2)	D_a+t^2(S100→4, So4)
D_a+t^2(S4→2, S2)	D_a+t^2(S100→2, S2)	D_a+t^2(S100→4, S4)
D_t+a^2(S4→2, TR2)	D_t+a^2(S100→2, TR2)	D_t+a^2(S100→4, So4)
D_t+a^2(S4→2, S2)	D_t+a^2(S100→2, S2)	D_t+a^2(S100→4, S4)

Document-centric transform then aggregate (D_t+a^2). A variant of the previous approach consists in first transforming judgments into the target scale using a certain cut and then aggregating them (e.g., by majority vote) to compare them with judgments in the target scale (e.g., by experts) and measure α agreement for the specific cut. The best cut is identified using the highest α value.

We additionally distinguish two possible types of judgment datasets in the target scale: expert judgments (e.g., TR2 binary judgments) or crowd-generated aggregated judgments collected natively in the target scale (e.g, S2).

3.3.3 Possible Transformations. Given all the transformation methods presented so far and the datasets used in our experiments, we can generate 27 possible transformations (listed in Table 2 and described below) which we will experimentally analyze and compare in the remaining of this paper with the goal of drawing an understanding of how to best transform relevance judgments into a different relevance scale. Using the single dataset scale transformation method defined in Section 3.3.1 we can transform three crowd-generated judgment collections (i.e., S4 into binary, S100 into binary, and S100 into four levels) using the two HIT-centric methods or, alternatively, using topic-based α to select the best cut for each topic. This leads to nine possible transformations (see Table 2). By using the double dataset scale transformation method (see Section 3.3.2), using either of the two judgment sets (i.e., by experts like TR2 binary and So4 4-level judgments, or by the crowd such as S2 and S4) in the target scale we can transform judgments using the HIT-centric approach or the two possible document-centric approaches (i.e., performing first an aggregation and then a scale transformation or first a scale transformation and then a judgment aggregation). This leads to additional 18 possible transformations (double dataset by external agreement in Table 2). In addition, using the document-centric approach we are able to perform the transformation of 4-level expert judgment set (i.e., So4) into binary, which allows us to compare with binary expert TR2 judgments. Note that in this transformation, the aggregation step is not needed.

Table 3: Agreement α between transformed Sormunen judgments and TREC with and without the assumption that unjudged documents are not relevant.

Source scale	Target scale	Available documents	α values for transforming on left	middle	right
So4	TR2	805	0.595	0.136	−0.356
$So4_a$	TR2	3881	0.882	0.620	0.212
$So4_a$	$TR2_a$	4269	0.884	0.625	0.220

We also notice that when transforming a 4-level scale into a binary scale the number of possible cuts is 3: *left* (0 into 0 and 1, 2, 3 into 1), *middle* (0, 1 into 0 and 2, 3 into 1), and *right* (0, 1, 2 into 0 and 3 into 1). When transforming a 101-level scale into binary the number of cuts is 100, and when transforming a 101-level scale into four levels it is 161 700.

3.4 The Effect of Unjudged Documents

In both TR2 and So4 judgments, only a subset of documents have been judged by experts (see Section 3.1). Thus, we make the common assumption of unjudged documents to be not relevant. To verify this assumption, we add non-relevant labels to both TR2 and So4 judgments for unjudged documents that appear in our other experimental datasets (i.e., S2, S4, and S100). We thus obtain two additional datasets: $TR2_a$ and $So4_a$, where a denotes adding the additional non-relevant labels based on this assumption. Next, we perform the transformation of So4 4-level judgments into binary, and measure their agreement with TR2 judgments by means of α using the *doc-centric* approach in two ways: i) only over judgments which are available in both datasets and ii) over all documents with the assumptions that unjudged documents are not relevant.

Table 3 presents the number of judged documents in both So4 and TR2 datasets, with and without the assumption that unjudged documents imply non-relevance, and the results of the best cuts selected to transform So4 judgments into binary using TR2 as target scale dataset. The results show that by adding more non-relevant labels to So4 judgments, the agreement between the two expert judgments becomes higher. This shows (also due to the selection strategy followed to build So4, see Section 3.1) that unjudged documents in So4 are often labelled as non-relevant by experts in TR2 judgments, which is in line with the assumption that unjudged documents are assumed as non-relevant ones. Note also that, in all cases, the best cut selected using α values is the left one. As it was already observed by Sormunen [23], the relevance threshold is low in TREC judgments which is explained by the risk of possibly missing relevant documents [29]. Based on this observation, in the following we only focus on the results where we make this assumption. We report results both making the assumption and removing unjudged documents for S100 transformed into a 4-level scale using So4, as the results vary substantially in that case because of the many missing judgments in So4.

3.5 Evaluation of Scale Transformations

To understand which of the proposed scale transformation methods leads to better relevance judgment datasets in the target scale, we look at what results the generated judgements produce when used for the evaluation of IR system effectiveness. Specifically, we compare the evaluation results of the transformed judgments against expert judgments collected natively in the target scale. By using NDCG [10] as evaluation metric, we compute Kendall's τ correlation between the system ranking generated with the transformed judgments and the ranking generated with expert judgments assuming that the desired outcome is to obtain a transformed judgement set that leads to IR evaluation results as similar as possible to expert judgments. Thus, when comparing different alternative methods to transform a judgment set from a source scale to a target scale, we prefer the ones achieving higher τ values within the proposed evaluation approach.

4 RESULTS AND DISCUSSION

In this section, we present our findings from applying the scale transformation approaches presented above to the considered expert and crowdsourced relevance judgments datasets (So4, S2, S4, S100). Additionally to the 27 transformations of crowdsourced datasets (see Table 2), we report the transformation of the expert-generated dataset So4 into binary ($So4_a \rightarrow 2$, $TR2_a$).

4.1 So4→2 and S4→2

We begin with the simplest scale transformations, So4 and S4 to binary, where we have in total three possible choices to set cuts between any two neighbor relevance levels.

4.1.1 So4→2. Table 3 shows agreement values between expert judgments collected on a 4-level scale (So4) which are transformed into a binary scale with all possible cut choices. The results indicate that setting cuts just after 0 (left cut) outperforms the other two choices (middle and right). On a per-topic basis, the left cut works best in 17 out of the 18 considered topics, with only one exception where the difference between left- and middle-cut is not large. This is consistent with the definition of relevance used by experts in both TR2 and So4 judgments, as discussed at the end of Section 3.4.

4.1.2 S4→2. Next, we present the results involving judgments from crowd workers (i.e., S4). Figures 1c and 2c show the results of single dataset (i.e., S4) scale transformation with all three possible cuts (left, middle, and right), and scale transformation methods described in Section 3.3.1: HIT-centric a+t and t+a (Fig. 1c) and Topic-wide α (Fig. 2c). Note that HIT-centric a+t and t+a produce exactly the same result, just because we adopt majority vote (the same as median in the binary case) as our aggregation function for 2-level scales while take median values for 4-level scales. Actually, since we aggregate judgments from nine workers, the result of majority vote is always the same as the fifth value in a ranked list of 2-level judgments. On the other hand, the median value of a ranked list of nine 4-level judgments is the fifth value by definition. Therefore, no matter whether we use t+a or a+t in the scale transformation method, the result of aggregating nine judgments exactly equals to transforming the fifth value in a ranked list, which means that both methods produce the same transformation.

From the charts it is evident that while the agreement level measured by means of α varies from topic to topic, the best cut selected by each method is stable (i.e., middle cut). Figures 1ab and 2ab show the agreement between transformed judgments from source scale and target scale judgments with all possible cuts, by the double dataset methods defined in Section 3.3.2 where we use

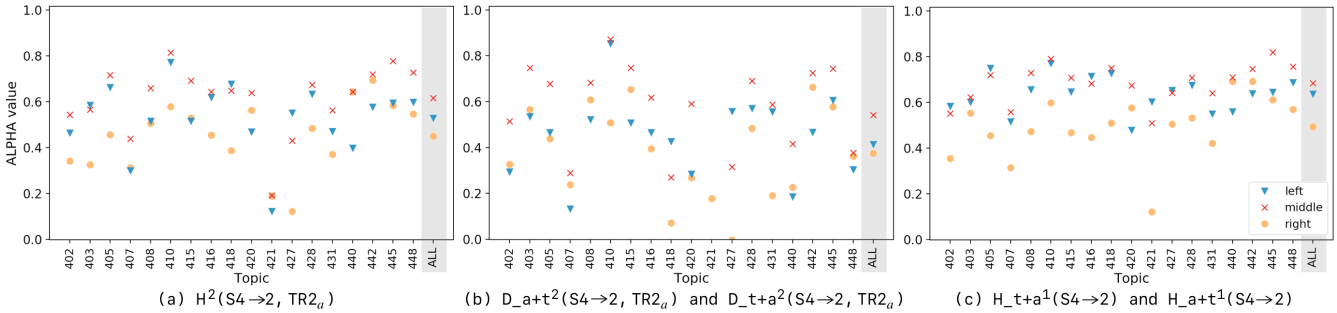


Figure 1: Transformation of S4 into 2 levels with TR2$_a$ (a) and (b) and HIT-centric single dataset transformation (c).

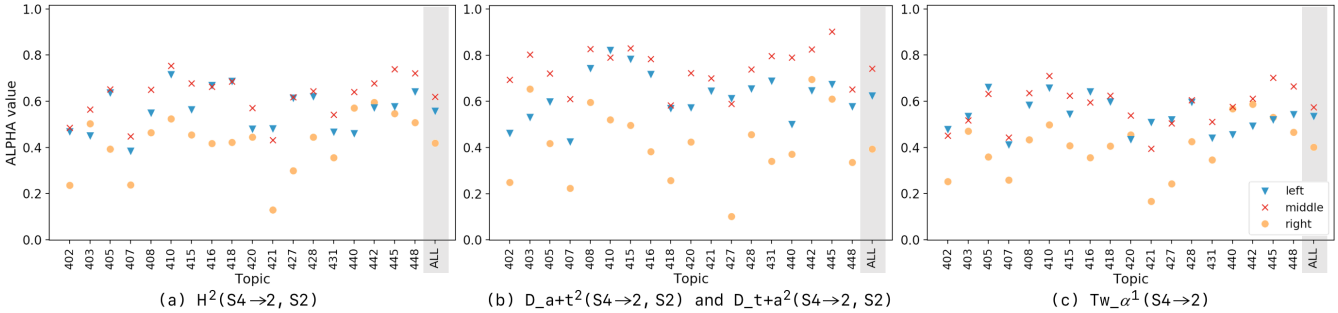


Figure 2: Transformation of S4 into 2 levels with S2 (a) and (b) and single dataset transformation with Tw_α^1 (c).

both TR2 (Fig. 1ab) and S2 (Fig. 2ab) as judgment dataset in the target scale. The results show differences between using expert (i.e., TR2) and crowd (i.e., S2) judgments to measure the agreement used to select the best cut. The best cuts of Topics 421 and 427, for example, are identified as right- and left-transformation when compared to expert judgments, while both are picked up as middle-transformation as the best when using crowdsourced labels. This is because crowd workers have given different judgments as compared to experts (e.g., only 34 out of all 342 documents are labelled as relevant by experts in Topic 421, while crowd workers in S2 have judged 197 documents in this topic as relevant).

For each of these nine results on transforming S4, as well as for the result on So4, we show (on the right-hand side of each chart) the average of α values at each cut across all topics, by which we are able to identify a single best cut for the entire judgment collection (i.e., S4 or So4).[1] We observe that when transforming crowd judgments (i.e., S4) into binary, middle cut works best on average, regardless of what method is chosen to make the transformation. This is different from binarizing expert judgments (i.e., So4), where the best cut is the left one (Tab. 3). Such result reveals how the relevance definition used by crowd workers differs from that of experts.

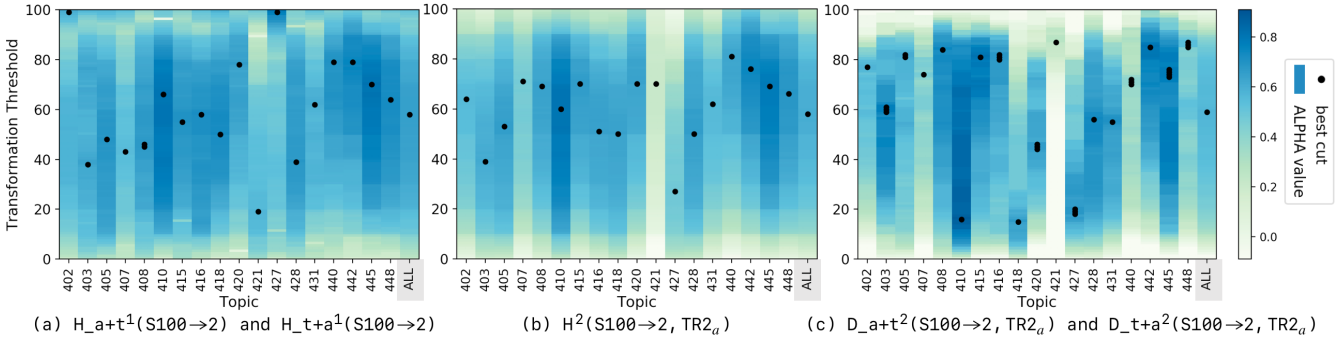
4.2 S100→2

In this section, we focus on transformation of S100 into a binary scale. In this scenario there are a total of 100 possible choices among which to pick the best cut. Figures 3 and 4 present the results of both single and double dataset scale transformations defined in Section 3.3 for all possible cut choices. Note that in these charts (and in the following ones) the y-axis is not the α value like in the previous section for S4 transformation into binary, but rather the cut in the 0..100 range; the α value for each cut is shown by the color density. The small black dot shows the best cut. In case of multiple cuts having the same best α, there are multiple black dots shown for some topics, such as, for example, Topic 403 in Figure 4c.

Unlike binarizing S4 scales, the selection of a method to transform S100 into 2 levels has an impact on picking up the best cut for different topics. In Topic 402, for example, the best cut is identified as 99 by HIT-centric methods in the single dataset transformation (see Figures 3a), while it is around 60 according to HIT-centric methods in the double dataset transformation (see Figures 3b and 4b) and below 50 in Doc-centric methods when comparing to S2 judgments (see Figure 4c). This shows how different scale transformation methods may lead to very different judgments in the target scale and potentially impact the evaluation of IR systems. There are two extreme cases in single dataset HIT-centric method: Topic 402 and 427 (in Figures 3a). The best cuts for both topics are selected as 99. This is because among all judgments in S100, no judgment at level 100 has been given by crowd workers to any document in these topics, and, therefore, by setting the cut at 99 (i.e., mapping 100 to 1 and others to 0) every judgment is transformed into 0, which, in this case, maximises α agreement. We note, however, that alternative scale transformation methods select more appropriate cuts to transform S100 judgments into binary.

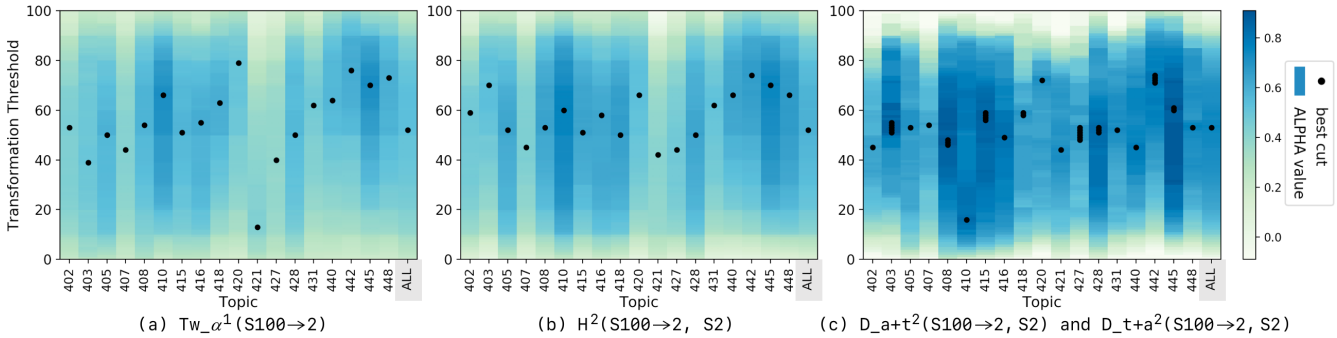
From Figure 5, we can observe that the cut that maximises α when applied on the entire collection rather than on a per-topic basis is 58. We can also notice that the lowest cuts (i.e., 0-10) lead to lower α values ($\alpha = 0.21$ when cut= 0) than the highest possible cuts (i.e., 90-100 with $\alpha = 0.51$ when cut= 99). This can be explained

[1]Note that an alternative approach to select a single cut for the entire collection is to re-run a transformation method over all judgments regardless of their topic. This approach, which we do not report about for space limitations, leads to consistent results to those reported in this paper.



(a) H_a+t^1(S100→2) and H_t+a^1(S100→2) (b) H^2(S100→2, TR2$_a$) (c) D_a+t^2(S100→2, TR2$_a$) and D_t+a^2(S100→2, TR2$_a$)

Figure 3: Transformation of S100 into 2 levels using a single dataset (a) and using TR2$_a$ (b) and (c).



(a) Tw_α^1(S100→2) (b) H^2(S100→2, S2) (c) D_a+t^2(S100→2, S2) and D_t+a^2(S100→2, S2)

Figure 4: Transformation of S100 into 2 levels using a single dataset (a) and using S2 (b) and (c).

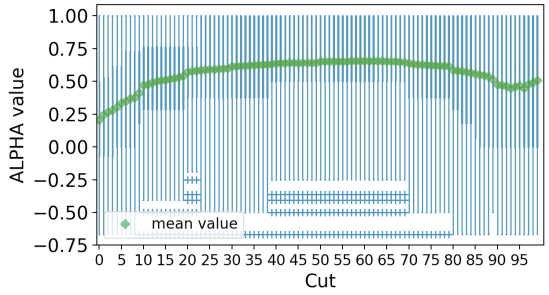


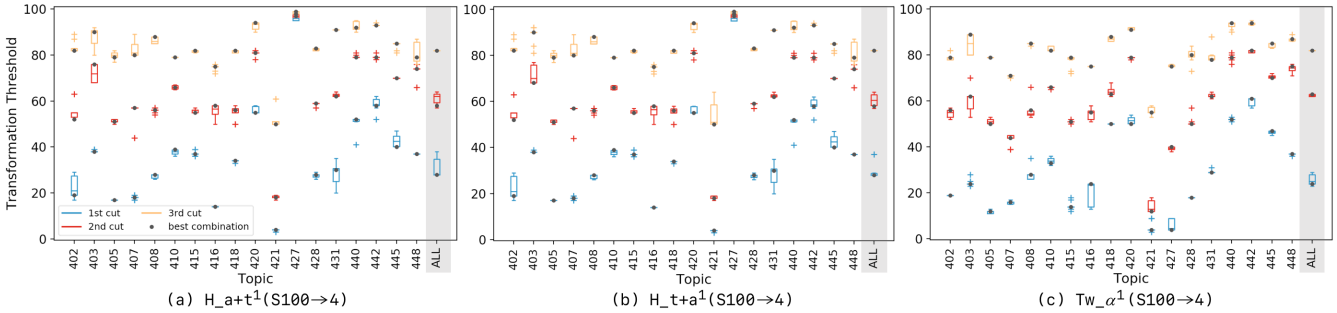
Figure 5: Inter-rater agreement over different cuts with H_t+a^1(S100→2) on the entire collection of S100.

by the unbalanced distribution of relevant/non-relevant documents in the collection. That is, by selecting a cut on the lower side of the scale we generate a binary judgment sets with many relevant documents. Contrary, by selecting a cut on the higher extreme of the scale we generate many not-relevant judgments, which is more aligned with the natural distribution of judgements (with few relevant and many not-relevant judgments).

4.3 S100→4

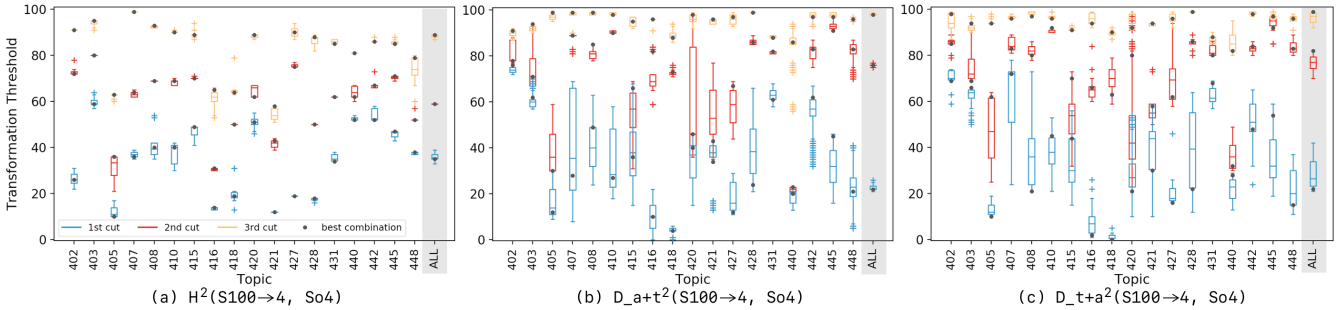
The most computationally challenging transformation is S100 into 4 levels as it implies selecting the best of 160K possible cuts.

4.3.1 Comparing Scale Transformation Methods. Figures 6, 7, 8, and 9 show for each transformation method a plot where for each topic we show the range of top 10 α values for each of the three points of a scale transformation cut. The black dots indicate the best cut for each topic. Looking at these results we can make the following observations. In HIT-centric single datasets methods, for topic 427 the methods select an extreme cut (which also happened for this topic in the previous results on transforming S100 into binary. This issue gets however fixed by using double dataset methods that allows to select a more natural cut by measuring agreement against a judgment set in the target scale. Because of So4 having a lot of unjudged documents, in this experiment we also compare the results of selecting the best cut using either So4 or So4$_a$ as target scale dataset (Fig. 7 and 8). We can observe that when using So4 without the assumption that unjudged are not relevant (and thus removing the unjudged from the agreement computation) we obtain a large interval of best cut points (shown by the large boxes in the plots of Fig. 7). This is because 81% of the documents have been removed from the calculation of α values. When we make the assumption that unjudged are not-relevant we have less variability in the selection of the best cut point (Fig. 8). When using So4$_a$ we tend to get higher cuts (i.e., to the right side of the scale) as compared to when we use S4 as target scale dataset. In detail, the top cut point selected by So4$_a$ methods tends to be 90 or higher and the difference from that selected by S4 methods is statistically significant (t-test $p < 0.001$). This shows how using either expert or crowdsourced datasets has a strong impact on the selection of the best cut and on the judgment scale transformation results. Finally, we also observe that double dataset doc-centric methods tend to display higher variability in the best cuts as compared to HIT-centric methods. This is because in HIT-centric methods, there is no aggregation function used in the scale transformation process. When



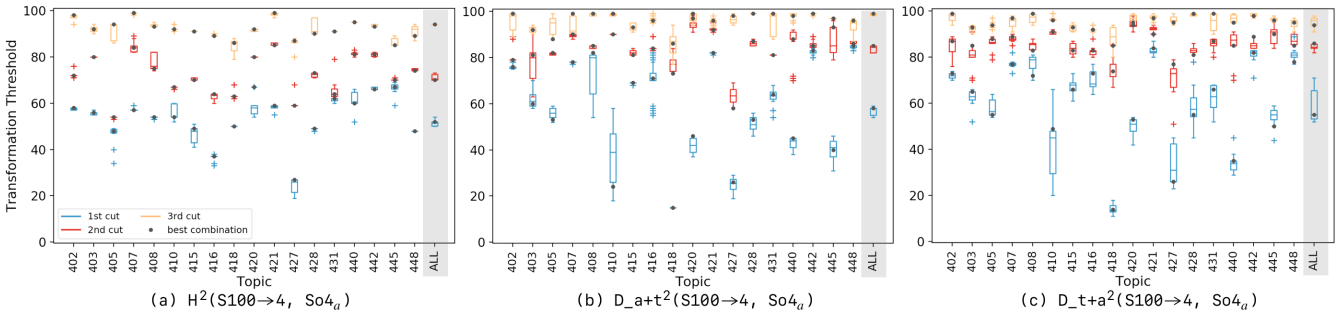
(a) H_a+t^1(S100→4) (b) H_t+a^1(S100→4) (c) Tw_α^1(S100→4)

Figure 6: Single dataset transformation of S100 into 4 levels.



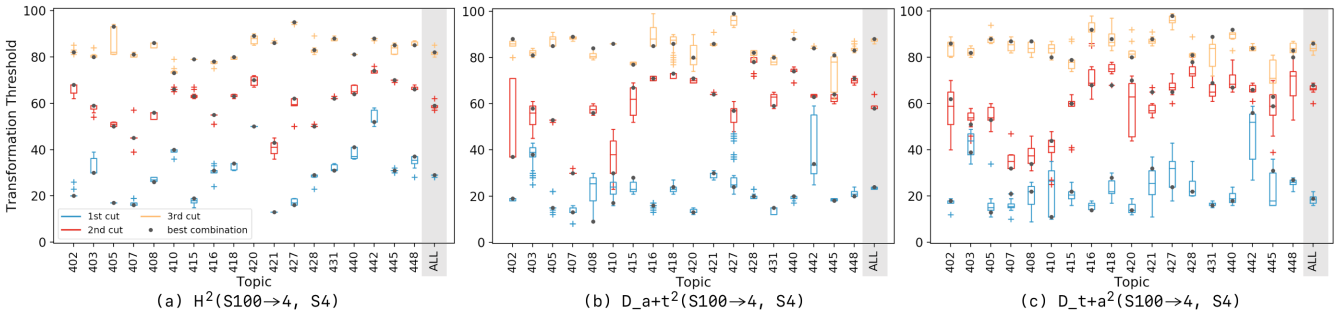
(a) H^2(S100→4, So4) (b) D_a+t^2(S100→4, So4) (c) D_t+a^2(S100→4, So4)

Figure 7: Transformation of S100 into 4 levels with So4.



(a) H^2(S100→4, So4$_a$) (b) D_a+t^2(S100→4, So4$_a$) (c) D_t+a^2(S100→4, So4$_a$)

Figure 8: Transformation of S100 into 4 levels with So4$_a$.



(a) H^2(S100→4, S4) (b) D_a+t^2(S100→4, S4) (c) D_t+a^2(S100→4, S4)

Figure 9: Transformation of S100 into 4 levels with S4.

performing judgment aggregation, with either D_a+t^2 or D_t+a^2, the distribution of the original relevance judgments shrinks into just one single value, which introduces *gaps* (i.e., not all scale values are used) in the distribution of the crowdsourced judgments. As a result, whichever cut is selected within the gap interval (i.e., the scale interval between two scale points used within aggregated relevance judgments), the transformed judgments remain the same, and consequently all these cuts receive the same α and can be equally considered the best. In general, a comparison between the a+t and t+a approaches does not show significant differences.

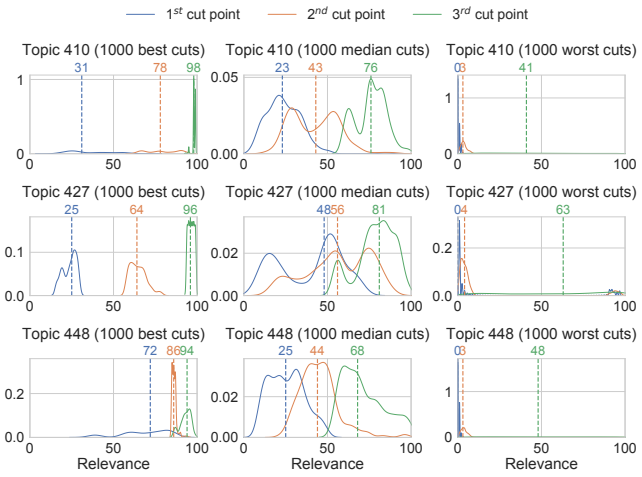


Figure 10: Distributions of the three cut points for the 1000 best (left column), median (center column), and worst (right column) cuts, over three topics in the collection.

4.3.2 Analysis of the cut distribution. To investigate how all possible cuts are distributed, we focus on D_a+t^2(S100→4, So4) which is the transformation method having the highest τ with per-topic cuts (0.820, see Tab. 4). We selected, out of the 160K possible cuts, 3 subgroups having respectively the 1,000 *best*, *median*, and *worst* cuts. Figure 10 shows that in such groups the cut variability is quite limited for the best cuts (left column) and the distance in terms of relevance is quite consistent among the three cut points. For median cuts (central column), the noise increases but the three cut points are still uniformly distributed and distinguishable. The cut points of the worst cuts (right columns) have values very close to scale boundaries. This happens in a consistent way across topics and different transformation methods.

We now look at relations between properties of cuts and intrinsic characteristics of topics. First, we group the *best*, *median*, and *worst* 1000 cuts, then we compute the mean of topics of the widths of the 1000 cuts ranges (intended as difference of relevance between the first and third cut point). We find that the mean cut width of *median* cuts correlate positively (Pearson 0.54, $p < 0.05$) with the ratio of relevant documents in the topic, whereas such a correlation is negative when considering the group of the 1000 worst cuts (Pearson -0.5, $p < 0.05$). This suggests that topics having a high ratio of relevant tend to have *worst* cuts within a small range, and *median* cuts within a wider range.

4.4 Evaluating Scale Transformation Methods

Table 4 shows the system ranking correlations computed when using relevance judgments obtained with the selected best cut and expert judgments. This allows us to compare the effect of different scale transformation methods on IR evaluation results, as described in Section 3.5. We can observe that in the S4 transformation to binary all methods select the middle cut as the best one, and thus the correlation values are very similar and only differ because of ties in the system rankings generated by the transformed judgments [2]. We thus cannot draw conclusions on which scale transformation method is best for this experiment as they all pick the same best cut. When comparing single-dataset and double-dataset methods

Table 4: Kendall's τ correlation between the IR system ranking generated using the transformed judgments and expert judgments for both selecting the best cut on a per-topic basis or for the entire collection. Bold indicates the best values per experiment only considering expert judgments in the same scale as the target (not reported for the first experiment as all methods lead to the same best cut). So4 values are grayed-out when the transformation is into binary.

	Per-topic-cut		Single-cut		Single-Cut
	TR2	So4	TR2	So4	
H_t+a^1(S4→2)	0.70	0.71	0.71	0.72	middle
H_a+t^1(S4→2)	0.70	0.71	0.73	0.74	middle
Tw_α^1(S4→2)	0.69	0.70	0.72	0.74	middle
H^2(S4→2, TR2)	0.72	0.74	0.72	0.73	middle
H^2(S4→2, S2)	0.72	0.73	0.71	0.72	middle
D_a+t^2(S4→2, TR2)	0.73	0.75	0.73	0.74	middle
D_a+t^2(S4→2, S2)	0.74	0.75	0.73	0.74	middle
D_t+a^2(S4→2, TR2)	0.73	0.75	0.71	0.72	middle
D_t+a^2(S4→2, S2)	0.73	0.73	0.71	0.72	middle
H_t+a^1(S100→2)	0.71	0.72	**0.77**	0.81	58
H_a+t^1(S100→2)	0.70	0.71	0.75	0.80	58
Tw_α^1(S100→2)	0.72	0.74	0.76	0.79	52
H^2(S100→2, TR2)	0.72	0.74	0.75	0.80	58
H^2(S100→2, S2)	0.76	0.79	0.76	0.79	52
D_a+t^2(S100→2, TR2)	0.76	0.78	0.76	0.80	59
D_a+t^2(S100→2, S2)	0.77	0.80	0.76	0.79	53
D_t+a^2(S100→2, TR2)	**0.77**	0.78	0.75	0.80	55
D_t+a^2(S100→2, S2)	0.77	0.81	0.75	0.81	56
H_t+a^1(S100→4)	0.75	0.76	0.76	0.77	(28, 58, 82)
H_a+t^1(S100→4)	0.74	0.74	0.74	0.76	(28, 58, 82)
Tw_α^1(S100→4)	0.74	0.76	0.75	0.77	(24, 63, 82)
H^2(S100→4, So4)	0.77	0.77	0.76	**0.79**	(52, 70, 94)
H^2(S100→4, S4)	0.75	0.76	0.76	0.77	(29, 59, 82)
D_a+t^2(S100→4, So4)	0.77	**0.82**	0.73	0.77	(58, 85, 99)
D_a+t^2(S100→4, S4)	0.74	0.76	0.73	0.76	(24, 58, 88)
D_t+a^2(S100→4, So4)	0.76	0.80	0.74	0.75	(55, 86, 94)
D_t+a^2(S100→4, S4)	0.73	0.76	0.74	0.76	(19, 68, 86)

we see that using a second dataset in the target scale consistently leads to more similar IR evaluation results. In the case of double-dataset methods, using expert generated judgments leads to higher correlation as compared to using crowd-generated judgments in the target scale. These results are also due to the fact that the second dataset that we are using to select the cuts is the same one we compare against when looking at IR system correlation. We can also observe that, when transforming S100 into 4 levels, document-centric methods lead to higher evaluation results especially when using expert judgments (i.e., So4) as target scale dataset. We also see that for document-centric methods selecting different cuts for each topic leads to more similar results as compared to selecting a single cut for the entire collection. However, on average, selecting the best cut independently for each topic is not necessarily always better than selecting a single best cut for the entire collection. Comparing the use of So4 and S4 in double dataset methods, we see that So4 leads to significantly higher cuts than S4. This confirms previous observations about the differences between experts and the crowd.

5 CONCLUSIONS AND FUTURE WORK

Selecting the right relevance scale to be used within the creation of an IR evaluation collection is a key decision to make. When reusing existing collections, it may be necessary to transform judgments that have been originally collected in a fine-grained scale into a different relevance scale. In this paper we presented an extensive study of relevance scale transformation methods over different datasets. To the best of our knowledge, this is the first study of this kind. We looked both at classic transformations previously adopted in the IR evaluation literature (i.e., the binarization of 4-level judgments) up to extreme cases in which we transformed a fine-grained (i.e., 101 relevance levels) scale into 4 levels thus considering 160K possible ways to transform it. Our results indicate that the method we select to transform judgements have strong implications on the results of IR evaluation experiments. We observed that:

- Transforming the scale of a judgment collection is best done *on a per-topic basis* rather than selecting the same cut for the entire collection as all proposed methods tend to select quite different cuts for different topics in our collection. Selecting cuts per-topic or one single cut for the entire collection appears, however, not to have a large impact on the IR evaluation results as compared to those obtained with expert judgments.
- Transforming the scale of an expert-judged collection and a crowd-judged collection should not necessarily be done in the same way. In our experiments, when binarizing S4 and So4 we found that the best cuts for these two dataset are different (i.e., middle and left cut, respectively, for S4 $\longrightarrow$ 2 and So4 $\longrightarrow$ 2).
- The classic assumption that unjudged documents are considered not-relevant may have strong implication on the way we transform judgments from one scale to another. This results is not surprising as such an assumption is known to be invalid [29].
- When comparing the IR evaluation outcomes obtained with transformed judgments with those obtained with native expert judgments, we observed that document-centric methods lead to more similar results and should thus be preferred.

In the future, we plan to differentiate the results from identifying the best cut using the closest α of transformed judgment set by internal agreement to that of original judgment set, instead of picking up the best cut by the highest α. We will also investigate the use of system ranking correlation measures (e.g., τ) as an alternative to assessor agreement for the selection of the best cut, perform in-depth comparative analysis of per-topic versus single-cut approaches, and design supervised methods to select cuts for new datasets.

Acknowledgements. This work is partially supported by the ARC Discovery Project (Grant No. DP190102141) and has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 732328.

REFERENCES

[1] Ahmed Abbasi, Ammar Hassan, and Milan Dhar. 2014. Benchmarking Twitter Sentiment Analysis Tools.. In *LREC*, Vol. 14. 26–31.
[2] Ben Carterette. 2009. On Rank Correlation and the Distance Between Rankings. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. ACM, New York, NY, USA, 436–443. https://doi.org/10.1145/1571941.1572017
[3] Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2017. Let's Agree to Disagree: Fixing Agreement Measures for Crowdsourcing. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*.
[4] Cyril W Cleverdon. 1962. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. (1962).
[5] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics* (1979), 20–28.
[6] Gianluca Demartini and Stefano Mizzaro. 2006. A classification of IR effectiveness metrics. In *European Conference on Information Retrieval*. Springer, 488–491.
[7] Koji Eguchi, Keizo Oyama, Emi Ishida, Noriko Kando, and Kazuko Kuriyama. 2002. Overview of the Web Retrieval Task at the Third NTCIR Workshop.. In *NTCIR*.
[8] Norbert Fuhr, Saadia Malik, and Mounia Lalmas. 2004. Overview of the INitiative for the Evaluation of XML Retrieval (INEX) 2003. In *INEX 2003 Workshop Proceedings*. 1–11.
[9] Donna K Harman. 1993. *The first text retrieval conference (TREC-1)*. Vol. 500. US Department of Commerce, National Institute of Standards and Technology.
[10] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.
[11] Kazuaki Kishida, Kuang-hua Chen, Sukhoon Lee, Kazuko Kuriyama, Noriko Kando, Hsin-Hsi Chen, and Sung Hyon Myaeng. 2007. Overview of CLIR Task at the Sixth NTCIR Workshop.. In *NTCIR*.
[12] Kazuaki Kishida, Kuang-hua Chen, Sukhoon Lee, Kazuko Kuriyama, Noriko Kando, Hsin-Hsi Chen, Sung Hyon Myaeng, and Koji Eguchi. 2004. Overview of CLIR task at the fourth NTCIR workshop. In *Working Notes of the Fourth NTCIR Workshop Meeting, June 2-4*. 1–59.
[13] Klaus Krippendorff. 2007. Computing Krippendorff's alpha reliability. *Departmental papers (ASC)* (2007), 43.
[14] Eddy Maddalena, Marco Basaldella, Dario De Nart, Dante Degl'Innocenti, Stefano Mizzaro, and Gianluca Demartini. 2016. Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.
[15] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. 2017. On crowdsourcing relevance magnitudes for information retrieval evaluation. *ACM Transactions on Information Systems (TOIS)* 35, 3 (2017), 19.
[16] Eddy Maddalena, Kevin Roitero, Gianluca Demartini, and Stefano Mizzaro. 2017. Considering Assessor Agreement in IR Evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '17)*. ACM, New York, NY, USA, 75–82. https://doi.org/10.1145/3121050.3121060
[17] Saadia Malik, Mounia Lalmas, and Norbert Fuhr. 2005. Overview of INEX 2004. In *Advances in XML Information Retrieval*, Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Zoltán Szlávik (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–15.
[18] Alistair Moffat and Justin Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1, Article 2 (Dec. 2008), 27 pages. https://doi.org/10.1145/1416950.1416952
[19] Rodrigo Moraes, João Francisco Valiati, and Wilson P Gavião Neto. 2013. Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications* 40, 2 (2013), 621–633.
[20] Keizo Oyama. 2005. Overview of the NTCIR-5 WEB navigational retrieval subtask 2 (Navi-2). In *Proc. 5th NTCIR Workshop Meeting on Evaluation of Information Access Technologies, 2005*.
[21] Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. 2018. On fine-grained relevance scales. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 675–684.
[22] Falk Scholer, Diane Kelly, Wan-Ching Wu, Hanseul S Lee, and William Webber. 2013. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 623–632.
[23] Eero Sormunen. 2002. Liberal Relevance Criteria of TREC -: Counting on Negligible Documents?. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)*. ACM, New York, NY, USA, 324–330. https://doi.org/10.1145/564376.564433
[24] Andrew Turpin, Falk Scholer, Stefano Mizzaro, and Eddy Maddalena. 2015. The Benefits of Magnitude Estimation Relevance Assessments for Information Retrieval Evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 565–574. https://doi.org/10.1145/2766462.2767760
[25] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. 2014. Community-based Bayesian Aggregation Models for Crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. ACM, 155–164. https://doi.org/10.1145/2566486.2567989
[26] Ellen M. Voorhees and Donna K. Harman. 1999. Overview of the Eighth Text REtrieval Conference (TREC-8). In *The Eighth Text REtrieval Conference (TREC-8)*. 1–24.
[27] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceeding of 55th ACL*. 422–426.
[28] YY Yao. 1995. Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information Science* 46, 2 (1995), 133–145.
[29] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 307–314.