Unanimity-Aware Gain for Highly Subjective Assessments

Tetsuya Sakai
Waseda University
tetsuyasakai@acm.org

ABSTRACT

IR tasks have diversified: human assessments of items such as social media posts can be highly subjective, in which case it becomes necessary to hire many assessors per item to reflect their diverse views. For example, the value of a tweet for a given purpose may be judged by (say) ten assessors, and their ratings could be summed up to define its gain value for computing a graded-relevance evaluation measure. In the present study, we propose a simple variant of this approach, which takes into account the fact that some items receive unanimous ratings while others are more controversial. We generate simulated ratings based on a real social-media-based IR task data to examine the effect of our unanimity-aware approach on the system ranking and on statistical significance. Our results show that incorporating unanimity can affect statistical significance test results even when its impact on the gain value is kept to a minimum. Moreover, since our simulated ratings do not consider the correlation present in the assessors' actual ratings, our experiments probably underestimate the effect of introducing unanimity into evaluation. Hence, if researchers accept that unanimous votes should be valued more highly than controversial ones, then our proposed approach may be worth incorporating.

CCS CONCEPTS

•**Information systems → Retrieval effectiveness;**

KEYWORDS

effect sizes; evaluation measures; inter-assessor agreement; *p*-values; social media; statistical significance

1 INTRODUCTION

In traditional test-collection-based IR experiments, we often rely on our experience which says that system rankings would remain stable even if the set of document relevance assessments are replaced by another [13]. However, IR tasks have diversified: human assessments of items such as social media posts can be highly subjective, in which case it becomes necessary to hire many assessors per item to reflect their diverse views. For example, the value of a tweet for a given purpose may be judged by (say) ten assessors, and their ratings could be summed up to define its gain value for computing a graded-relevance evaluation measure (e.g. [8, 11]). In the present study, we propose a simple variant of this approach, which takes into account the fact that some items receive unanimous ratings while others are more controversial. We generate simulated ratings based on a real social-media-based IR task data to examine the effect of our unanimity-aware approach on the system ranking and on statistical significance. Our results show that incorporating unanimity can affect statistical significance test results even when its impact on the gain value is kept to a minimum. Moreover, since our simulated ratings do not consider the correlation present in the assessors' actual ratings, our experiments probably underestimate the effect of introducing unanimity into evaluation. Hence, if researchers accept that unanimous votes should be valued more highly than controversial ones, then our proposed approach may be worth incorporating.

EVIA 2017, co-located with NTCIR-13, Tokyo, Japan.

2 RELATED WORK

2.1 Document Relevance Assessments

Due to lack of space, we refer the reader to Sakai [7] for a short overview of studies related to inter-assessor agreement. Below, we briefly discuss two studies that helps us to explain the novelty of our approach to utilising multiple relevance assessments.

Megorskaya *et al.* [4] studied the benefit of communication between multiple assessors in the context of gamified relevance assessment for web search evaluation. The premise in their work is that every document needs to finally receive a single relevance level, as a result of a consensus between the assessors or an overruling by a "referee" etc. This is in contrast to our work, where we are interested in assessment tasks where there may be no such thing as *the* correct assessment, and therefore it is important to preserve different subjective views in the data and in evaluation.

Turpin *et al.* [12] propose to use *magnitude estimation* in document relevance assessments in order to obtain ratio-scale judgments instead of the traditional ordinal- or interval-scale ones, and to interpret the ratio-scale judgments directly as the gain values for computing normalised discounted cumulative gain (nDCG) and expected reciprocal rank (ERR). This is achieved by instructing the assessor to give an arbitrary score to his first document and subsequently to give a "relative" score to each of the remaining documents, where "relative" means "in comparison to the preceding document." While their approach and ours both produce continuous relevance assessments, unanimity across judges was not within the scope of their study.

2.2 Social Media Assessments

Wang *et al.* [14] examined the effect of assesser differences in the context of the TREC Tweet Timeline Generation task, by devising two sets of tweet equivalence classes constructed by different assessors. Their conclusion is similar to that of Voorhees [13] who examined the effect of *document relevance* assessor differences: despite the substantial differences in the two sets of clusters, system rankings and the absolute evaluation measure scores based on these two sets were very similar. Sakai *et al.* [8] used graded-relevance measures to evaluate a community QA answer ranking task; each answer was assessed by four assessors and its gain value for computing the measures was determined as the sum of assessors' grades.

Table 1: Examples of $RawG_i$, WG_i, UG_i when $D_{max} = 3$.

Item i	Ratings ($N = 5$)	$RawG_i$	D_i	WG_i	$UG_i(p = 0.2)$	$UG_i(p = 0.1)$
Item 1	2 2 2 2 2	10	0	10.0	13	11.5
Item 2	1 1 2 3 3	10	2	3.3	11	10.5
Item 3	0 2 2 3 3	10	3	0.0	10	10
Item 4	1 1 1 1 1	5	0	5.0	8	6.5
Item 5	0 0 0 0 3	3	3	0.0	3	3
Item 6	0 0 0 0 2	2	2	0.7	3	2.5
Item 7	0 0 0 0 1	1	1	0.7	3	2

More recently, Shang *et al.* [11] reported on the NTCIR-12 Short Text Conversation task which is basically a tweet retrieval task: in their Japanese subtask, the sum of scores from ten assessors were used to define the gain value of each tweet. Note that these studies do not take into account whether the assessors are unanimous or not; the sum is all that matters.

2.3 Li and Yoshikawa

The recent work of Li and Yoshikawa [2] is similar in spirit to ours, and deserves a detailed explanation. They consider the problem of assessing the similarity between two documents using many assessors, and propose to incorporate what they call "confusability" into measures such as Pearson's correlation. Specifically, when computing a correlation value, they propose to weight each labelled item i by $1 - c_i$, where c_i is a normalised measure of confusability based on the difference (D_i) between the highest and the lowest ratings for item i, and so on[1]. Li and Yoshikawa remark that the same idea can be applied to other measures such as nDCG, although they do not provide any details: here, let us try to faithfully apply their idea to ranked retrieval evaluation based on a group of assessors. Let N be the number of assessors per item, and suppose that each assessor assigns to each item a rating on a scale from $0, 1, \ldots, D_{max}$. A straightforward way to define the final relevance level or the actual gain value for each item would be to just sum up the ratings [8, 11]: then we would have relevance levels from 0 to ND_{max}. For any item i with N independent assessments, let $RawG_i$ denote the gain value thus obtained. The above approach of Li and Yoshikawa suggests that we modify each gain value as follows:

$$WG_i = (1 - c_i)RawG_i = (1 - D_i/D_{max})RawG_i \ . \quad (1)$$

Let us consider Items 1-3 shown in Table 1 with $N = 5, D_{max} = 3$. Clearly, $RawG_1 = RawG_2 = RawG_3 = 10$, and according to Eq. 1, $WG_1 = 10$, $WG_2 = 3.3$, $WG_3 = 0$. Thus Items 2 and 3 are considered worse than (say) Item 4 in Table 1, since $WG_4 = RawG_4 = 5$. Clearly, a more careful consideration is in order.

2.4 Maddalena et al.

More recently, at ICTIR 2017, Maddalena *et al.* [3] proposed an evaluation approach whose motivation is almost identical as ours: they also claim that the distribution of the scores from different assessors should be utilised for IR evaluation. More specifically, they propose to replace a gain value of a document with an *interval* of gain values or even with a *distribution* of gain values, so that the final evaluation measures are also intervals or distributions. They call their measures *agreement-aware* measures.

In contrast to their novel approaches, our proposal simply utilises the original score distribution across assessors to adjust the gain value of each document so that a traditional evaluation measure can be computed. It remains to be seen how the interval and distribution measures of Maddalena *et al.* can effectively be utilised in IR evaluation venues such as CLEF, NTCIR and TREC.

3 PROPOSED METHOD

Our proposal is very simple and highly intuitive. Given a constant p ($0 \leq p \leq 1$), let us define the *unanimity-aware* gain as follows:

$$UG_i = RawG_i + pN(D_{max} - D_i) \quad (2)$$

if $RawG_i > 0$; otherwise $UG_i = RawG_i = 0$. Here, $D_{max} - D_i$ is a simple measure of unanimity where D_i is, as before, the difference between the maximum and the minimum among the N ratings[2]. Whereas, p controls the impact of unanimity on the gain. Thus, while we mainly want to reflect $RawG_i$ in our evaluation, we apply an "upgrade" according to the degree of unanimity. When the ratings of an item are perfectly unanimous (i.e., $D_i = 0$), we are giving it an extra pND_{max}; that is, we shall pretend that pN extra assessors gave it the highest possible rating.

For Items 1-3 shown in Table 1, $p = 0.2$ implies that $UG_1 = 13, UG_2 = 11, UG_3 = 10$; this is clearly more intuitive than the WG_i values. On the other hand, consider Items 5-7 in Table 1: note that if $p = 0.2$, $UG_5 = UG_6 = UG_7 = 3$. If this is not desirable, $p = 0.1$ may be used instead as shown in the same table; however, we shall discuss how to set an appropriate p elsewhere with real assessments in our future work. Hereafter, we only consider a modest impact by letting $p = 0.2$, as the focus of the present study is to demonstrate that our approach has a practical impact on experimental results even with a small p.

Our approach suggests a slight departure from traditional IR evaluation at the implementation level as well. In traditional IR, we usually prepare discrete relevance levels (e.g. relevant, highly relevant, etc.) to define the gold standard: we know the number of relevance levels in advance, and we map each relevance level to a *gain value* at the time of measure calculation. In contrast, our approach suggests that we retain the individual ratings in the test collection, from which gain values can be computed on the fly; there is no longer the notion of a predefined set of relevance levels. It is easy to see that the highest possible value of UG_i is $(1 + p)ND_{max}$. Fortunately, there is a readily available IR evaluation tool that accommodates not only relevance-level-based computation but also direct gain-value-based computation, as we shall discuss below.

4 EXPERIMENTS

Let us demonstrate the effect of introducing unanimity-aware gain to an IR task where the ratings of the items can be highly subjective. To this end, we chose to use the recent NTCIR-12 Short Text Conversation (STC) Chinese subtask data [11] for the following reasons: (1) STC requires the system to return a "reasonable" tweet as a response to a human tweet, and the assessments are expected

[1] Li and Yoshikawa [2] also considered using *standard deviation* and *entropy* to quantify c_i, but the present study focusses on the simplest case that relies on D_i as we believe that evaluation methods should be as simple as possible.

[2] Variants are possible of course: for example, we could obtain maximum and minimum values after removing outlier ratings.

to be highly subjective; (2) STC was the largest task of NTCIR-12, with 44 runs from 16 teams for the Chinese subtask[3]. The STC Chinese test collection contains 100 topics (i.e., input Weibo tweets) with relevance assessments ("qrels") containing the following relevance levels: $L0$ (judged nonrelevant); $L1$ (relevant); and $L2$ (highly relevant).

From the official qrels, we created 15 simulated variants with $D_{max} \in \{2, 4, 8\}$ and $N \in \{5, 10, 20, 40, 80\}$, as follows. For each judged tweet of each topic, $L0$ is replaced with $(0, 0, \ldots)$; whereas, both $L1$ and $L2$ are replaced with N simulated ratings obtained by random sampling from a uniform distribution over $[0, D_{max}]$. We then compute the unanimity-aware gains using Eq. 2, and evaluate up to top 10 Weibo tweets from each run. Note that randomly sampling N times implies that as N gets large, we are more likely to obtain both 0 and D_{max} among the N observations and therefore D_i is more likely to be D_{max}, i.e., UG_i is more likely to reduce to $RawG_i$ (Eq. 2). In contrast, *real* ratings of different assessors are probably correlated with one another, which should generally make D_i smaller than our simulated ratings. Hence, this experiment probably *underestimates* the impact of introducing unanimity-aware gain into evaluation.

We use the three official measures from STC: nG@1 (normalised gain at rank 1)[4], P+ (see below), and nERR (normalised expected reciprocal rank) [11]. While the official STC Chinese subtask used the `NTCIREVAL`[5] toolkit by giving a gain value of 1 to each $L1$-relevant tweet and 3 to each $L2$-relevant one, we utilised an alternative functionality of the same tool, which enables us to feed gain values of relevant items directly to it without considering the number of relevance levels. This feature was already available in `NTCIREVAL` for the purpose of accommodating the *global gain* proposed by Sakai and Song [9], which is an idea for obtaining a real-valued gain for each relevant web page for search result diversification evaluation. In their work, global gain was computed from intent-aware probabilities and per-intent graded relevance assessments.

P+, an official measure from STC but nevertheless less well-known than nDCG and nERR, deserves a brief explanation here. Just like nERR, it is a measure suitable for navigational intents. Just as Average Precision (AP) employs (binary) precision to measure the utility of the top r documents for a user group who abandon the ranked list at r, P+ employs the *blended ratio*, which combines precision and cumulative gain, for the same purpose. Furthermore, just as AP assumes that the distribution of users abandoning the ranked list is uniform across *all* relevant documents (even if some of them are not retrieved) [5], P+ assumes that the distribution is uniform over all relevant documents *ranked at or above* r_p, the *preferred rank* [11]. Given a ranked list, the preferred rank is the rank of the most relevant document that is closest to the top. In our case, the preferred rank is the rank of the document in the `res` file that has the highest UG_i value *and* is closest to the top. Both AP and P+ represent the *expected* utility over their user abandonment distributions.

[3] The Japanese subtask actually collected $N = 10$ individual ratings for each tweet, but had only 25 runs from seven teams. We plan to use this data set as well in a follow up study.
[4] Note that neither Discounting nor Cumulation of "nDCG" does not apply at rank 1.
[5] http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html

Table 2: Comparing the system rankings of $RawG_i$ vs. UG_i with Kendall's τ with 95% confidence intervals.

D_{max}	$N = 5$	$N = 10$	$N = 20$	$N = 40$	$N = 80$
(a) Mean nG@1					
2	.987 [.968, 1.007]	.989 [.970, 1.007]	1 [.995, 1.005]	1 [.995, 1.005]	1 [.995, 1.005]
4	.992 [.976, 1.005]	.983 [.962, 1.004]	.996 [.985, 1.006]	1 [.995, 1.005]	1 [.995, 1.005]
8	.985 [.968, 1.003]	.985 [.963, 1.008]	.998 [.994, 1.005]	.992 [.978, 1.008]	1 [.995, 1.005]
(b) Mean P+					
2	.985 [.964, 1.006]	.994 [.982, 1.005]	.998 [.991, 1.005]	1 [1, 1]	1 [1, 1]
4	.983 [.965, 1.002]	.992 [.977, 1.005]	.998 [.994, 1.004]	1 [.997, 1.003]	1 [.997, 1.003]
8	.989 [.972, 1.007]	.992 [.981, 1.005]	.994 [.982, 1.005]	.996 [.986, 1.006]	1 [1, 1]
(c) Mean nERR					
2	.996 [.985, 1.005]	.994 [.982, 1.005]	1 [1, 1]	1 [1, 1]	1 [1, 1]
4	.996 [.986, 1.006]	.994 [.982, 1.005]	.998 [.990, 1.005]	1 [1, 1]	1 [1, 1]
8	.989 [.977, 1.005]	.992 [.978, 1.005]	.998 [.991, 1.005]	.998 [.990, 1.005]	1 [1, 1]

5 RESULTS AND DISCUSSIONS

Table 2 compares the system rankings based on $RawG_i$ vs. UG_i in terms of Kendall's τ with 95% confidence intervals, for nG@1, P+ and nERR averaged over the 100 official Chinese STC topics. It can be observed that all of the upper confidence limits are above one, meaning that the systems rankings based on $RawG_i$ and UG_i are *statistically* equivalent. However, except where the 95% CIs are "[1, 1]," the two rankings are *not* identical, even with $p = 0.2$. Recall also that we should expect to see lower rank correlations if we use real assessors' ratings with correlations among them.

Probably a more practical concern than the change in the overall system ranking is: does the proposed method affect statistical significance test results? If a researcher is interested in comparing every system pair, then conducting a pairwise test such as the paired t-test repeatedly (without correcting α) is not the correct approach: one elegant solution would be to use the randomised Tukey HSD (Honestly Significantly Difference) test [1], which is free from distributional assumptions and ensures that the *familywise error rate* (i.e., the probability of incorrectly obtaining a statistically significant difference for at least one system pair) is α. We use the `Discpower`[6] toolkit to conduct the randomised Tukey HSD test from each topic-by-run score matrix, with $B = 5,000$ trials for each test. The STC Chinese subtask had 16 participating teams, and one run from each team (specifically, best run in terms of the official Mean nG@1 score) is considered in this analysis, giving us $16 * 15/2 = 120$ comparisons. Do $RawG_i$ and UG_i give us similar p-values and similar research conclusions?

Table 3 summarises the discrepancy between the significance test results with $RawG_i$ and those with UG_i: these are the comparisons where the difference is statistically significant at $\alpha = 0.05$ according to one while not significant according to the other. p-values, absolute score differences ($|d_{XY}|$), and *effect sizes* (ES_{HSD}) are also shown; ES_{HSD} is computed by dividing $|d_{XY}|$ by the residual standard deviation of each experimental condition [6][7]. It can be observed that the effect of introducing unanimity-aware gain

[6] http://research.nii.ac.jp/ntcir/tools/discpower-en.html
[7] This form of effect size measures the difference between two systems in standard deviation units; unlike the p-value, is not a function of the sample size.

Table 3: Discrepancies at $\alpha = 0.05$: p-values (those below α shown in bold), absolute differences, and effect sizes.

	D_{max}	N	system pair	(I) $RawG_i$			(II) UG_i		
				p-value	$\lvert d_{XY} \rvert$	ES_{HSD}	p-value	$\lvert d_{XY} \rvert$	ES_{HSD}
(a) nG@1	2	5	MSRSC-C-R1 vs. Grad1-C-R1	.064	.1373	1.808	**.034**	.1361	2.045
			ICL00-C-R1 vs. Grad1-C-R1	.062	.1378	1.814	**.029**	.1377	2.069
			cyut-C-R1 vs. HITSZ-C-R1	.064	.1375	1.811	**.033**	.1363	2.048
		10	ICL00-C-R1 vs. PolyU-C-R1	**.043**	.1564	1.723	.057	.1452	1.756
		20	Nders-C-R1 vs. picl-C-R2	.051	.1622	1.629	**.045**	.1632	1.650
	4	5	cyut-C-R1 vs. HITSZ-C-R1	.073	.1390	1.752	**.038**	.1418	1.953
	8	5	MSRSC-C-R1 vs. Grad1-C-R1	**.048**	.1481	1.797	.057	.1413	1.821
			ICL00-C-R1 vs. Grad1-C-R1	.052	.1469	1.782	**.049**	.1434	1.848
			cyut-C-R1 vs. HITSZ-C-R1	.054	.1466	1.779	**.024**	.1515	1.952
		10	ICL00-C-R1 vs. PolyU-C-R1	**.035**	.1666	1.668	.055	.1568	1.653
(b) P+	2	5	MSRSC-C-R1 vs. Poly-U-C-R1	**.037**	.1272	2.340	.051	.1176	2.431
			Grad1-C-R1 vs. HITSZ-C-R1	**.037**	.1273	2.342	.059	.1150	2.377
		10	ICL00-C-R1 vs. Grad1-C-R1	**.047**	.1311	2.251	.070	.1207	2.242
		20	ICL00-C-R1 vs. Grad1-C-R1	**.049**	.1330	2.193	.053	.1319	2.188
	4	5	MSRSC-C-R1 vs. Grad1-C-R1	**.048**	.1240	2.332	.068	.1146	2.343
			PolyU-C-R1 vs. HITSZ-C-R1	.076	.1186	2.230	**.047**	.1197	2.447
	8	5	PolyU-C-R1 vs. HITSZ-C-R1	.082	.1181	2.196	**.045**	.1210	2.400
		10	ICL00-C-R1 vs. Grad1-C-R1	**.032**	.1363	2.298	.086	.1223	2.159
			Nders-C-R1 vs. PolyU-C-R1	**.035**	.1357	2.288	.057	.1272	2.245
(c) nERR	2	5	BUPTTeam-C-R4 vs. ITNLP-C-R3	**.045**	.1347	2.208	.052	.1268	2.283
			MSRSC-C-R1 vs. Grad1-C-R1	.060	.1312	2.151	**.046**	.1280	2.375
	4	5	OKSAT-C-R1 vs. PolyU-C-R1	**.046**	.1380	2.157	.055	.1319	2.191
	8	5	MSRSC-C-R1 vs. Grad1-C-R1	**.037**	.1436	2.151	.052	.1369	2.139
		10	ICL00-C-R1 vs. Grad1-C-R1	**.042**	.1531	2.005	.052	.1480	2.008

cannot be overlooked, even with $p = 0.2$. For example, when $D_{max} = 2, N = 5$, there are three discrepancies between nG@1 based on $RawG_i$ and that based on UG_i among the 120 comparisons. Whereas, as was anticipated in Section 4, it can be observed that the impact of introducing unanimity is not observed for $N = 40, 80$. Again, with real ratings that tend to resemble one another and make D_i smaller that these random ratings do, we will probably observe a more substantial impact of introducing unanimity-aware gain into evaluation.

6 CONCLUSIONS AND FUTURE WORK

We proposed a simple and intuitive approach to incorporating the assessors' subjective yet unanimous decisions into gain-value-based retrieval evaluation, and demonstrated that this will affect experimental outcomes. Our results show that incorporating unanimity can affect statistical significance test results even when its impact on the gain value is kept to a minimum. Moreover, since our simulated ratings do not consider the correlation present in the assessors' actual ratings, our experiments probably underestimate the effect of introducing unanimity-aware gain into evaluation. Hence, if researchers accept that unanimous votes should be valued more highly than controversial ones, then our proposed approach may be worth incorporating. We also demonstrated how the proposed approach of directly feeding gain values to an existing evaluation tool can be accomplished, while bypassing the notion of discrete relevance levels.

Following the present study, the proposed unanimity-aware gain approach was applied to the recent NTCIR-13 Short Text Conversation (STC-2) Chinese subtask [10], with $p = 0.2$. There, according to the randomised Tukey HSD test, three extra statistically significantly different system pairs were obtained by using unanimity-aware nG@1 instead of the traditional nG@1; one extra statistically significantly different system pair was obtained by using unanimity-aware P+ instead of the traditional P+. Thus the sets of statistically significantly different system pairs according to the unanimity-aware approach were *supersets* of the corresponding sets based on the traditional gain values. However, there were only $N = 3$ assessors per topic.

In future work, we would like to apply our approach to diverse social-media-related tasks with many assessors (i.e., a large N), where, unlike our simulated ratings, correlations among the assessors are present. With real ratings, we expect to observe a larger impact of introducing unanimity-aware gain on the system ranking and statistical significance than we did in our simulations.

REFERENCES

[1] Ben Carterette. 2012. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM TOIS* 30, 1 (2012).

[2] Jiyi Li and Masatoshi Yoshikawa. 2016. Evaluation with Confusable Ground Truth. In *Proceedings of AIRS 2016 (LNCS 9994)*. 363–369.

[3] Eddy Maddalena, Kevin Roitero, Gianluca Demartini, and Stefano Mizzaro. 2017. Considering Assessor Agreement in IR Evaluation. In *Proceedings of ACM ICTIR 2017*. 75–82.

[4] Olga Megorskaya, Vladimir Kukushkin, and Pavel Serdyukov. 2015. On the Relation between Asessor's Agreement and Accuracy in Gamified Relevance Assessment. In *Proceedings of ACM SIGIR 2015*. 605–614.

[5] Stephen E. Robertson. 2008. A New Interpretation of Average Precision. In *Proceedings of ACM SIGIR 2008*. 689–690.

[6] Tetsuya Sakai. 2014. Statistical Reform in Information Retrieval? *SIGIR Forum* 48, 1 (2014), 3–12.

[7] Tetsuya Sakai. 2017. The Effect of Inter-Assessor Disagreement on IR System Evaluation: A Case Study with Lancers and Students. In *Proceedings of EVIA 2017*.

[8] Tetsuya Sakai, Daisuke Ishikawa, Noriko Kando, Yohei Seki, Kazuko Kuriyama, and Chin-Yew Lin. 2011. Using Graded-Relevance Metrics for Evaluating Community QA Answer Selection. In *Proceedings of ACM WSDM 2011*. 187–196.

[9] Tetsuya Sakai and Ruihua Song. 2011. Evaluating Diversified Search Results Using Per-Intent Graded Relevance. In *Proceedings of ACM SIGIR 2011*. 1043–1052.

[10] Lifeng Shang, Tetsuya Sakai, Hang Li, Ryuichiro Higashinaka, Yusuke Miyao, Yuki Arase, and Masako Nomoto. 2017. Overview of the NTCIR-13 Short Text Conversation Task. In *Proceedings of NTCIR-13*.

[11] Lifeng Shang, Tetsuya Sakai, Zhengdong Lu, Hang Li, Ryuichiro Higashinaka, and Yusuke Miyao. 2016. Overview of the NTCIR-12 Short Text Conversation Task. In *Proceedings of NTCIR-12*. 473–484.

[12] Andrew Turpin, Falk Scholer, Stefano Mizzaro, and Eddy Maddalena. 2015. The Benefits of Magnitude Estimation Relevance Assessments for Information Retrieval Evaluation. In *Proceedings of ACM SIGIR 2015*. 565–574.

[13] Ellen M. Voorhees. 1998. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *Proceedings of ACM SIGIR 1998*. 315–323.

[14] Yulu Wang, Garrick Sherman, Jimmy Lin, and Miles Efron. 2015. Assessor Differences and User Preferences in Tweet Timeline Generation. In *Proceedings of ACM SIGIR 2015*. 615–624.