Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification

Tetsuya Sakai
Waseda University / Shinjuku-ku Okubo 3-4-1, Tokyo 169-8555, JAPAN
tetsuyasakai@acm.org

Abstract

Ordinal Classification (OC) is an important classification task where the classes are ordinal. For example, an OC task for sentiment analysis could have the following classes: *highly positive, positive, neutral, negative, highly negative*. Clearly, evaluation measures for an OC task should penalise misclassifications by considering the ordinal nature of the classes (e.g., *highly positive* misclassified as *positive* vs. misclassifed as *highly negative*). *Ordinal Quantification* (OQ) is a related task where the gold data is a *distribution* over ordinal classes, and the system is required to estimate this distribution. Evaluation measures for an OQ task should also take the ordinal nature of the classes into account. However, for both OC and OQ, there are only a small number of known evaluation measures that meet this basic requirement. In the present study, we utilise data from the SemEval and NTCIR communities to clarify the properties of nine evaluation measures in the context of OC tasks, and six measures in the context of OQ tasks.

1 Introduction

In NLP and many other experiment-oriented research disciplines, researchers rely heavily on evaluation measures. Whenever we observe an improvement in the score of our favourite measure, we either assume or hope that this implies that we have managed to moved our system a little towards what we ultimately want to achieve. Hence it is of utmost importance to examine whether evaluation measures are measuring what we want to measure, and to understand their properties.

This paper concerns evaluation measures for Ordinal Classification (OC) and Ordinal Quantification (OQ) tasks. In an OC task, the classes are ordinal, not nominal. For example, Task 4 (Sentiment Analysis in Twitter) Subtask C in SemEval-2016/2017 is defined as: given a set of tweets about a particular topic, estimate the sentiment conveyed by each tweet towards the topic on a five-point scale (*highly negative, negative, neutral, positive, highly positive*) (Nakov et al., 2016; Rosenthal et al., 2017). On the other hand, an OQ task involves a gold distribution of labels over ordinal classes and the system's estimated distribution. For example, Task 4 Subtask E of the SemEval-2016/2017 workshops is defined as: given a set of tweets about a particular topic, estimate the distribution of the tweets across the five ordinal classes already mentioned above (Nakov et al., 2016; Rosenthal et al., 2017). The Dialogue Breakdown Detection Challenge (Higashinaka et al., 2017) and the Dialogue Quality subtasks of the NTCIR-14 Short Text Conversation (Zeng et al., 2019) and the NTCIR-15 Dialogue Evaluation (Zeng et al., 2020) tasks are also OQ tasks. [1]

Clearly, evaluation measures for OC and OQ tasks should take the ordinal nature of the classes into account. For example, in OC, when a *highly positive* item is misclassified as *highly negative*, that should be penalised more heavily than when it is misclassified as *positive*. Surprisingly, however, there are only a small number of known evaluation measures that meet this requirement. In the present study, we use data from the SemEval and NTCIR communities to clarify the properties of nine evaluation measures in the context of OC tasks, and six measures in the context of OQ tasks. Some of these measures satisfy the aforementioned basic requirement for ordinal classes; others do not.

[1] In terms of data structure, we observe that the relationship between OC and OQ are similar to that between paired data and two-sample data in statistical significance testing. In OC, we examine which item is classified by the system into which class, and build a confusion matrix of gold and estimated classes. In contrast, in OQ, we compare the system's distribution of items with the gold distribution, but we do not concern ourselves with which item in one distribution corresponds to which item in the other.

*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing*, pages 2759–2769
August 1–6, 2021. ©2021 Association for Computational Linguistics

Section 2 discusses prior art. Section 3 provides formal definitions of the measures we examine, as this is of utmost importance for reproducibility. Section 4 describes the data we use to evaluate the measures. Sections 5 and 6 report on the results on the OC and OQ measures, respectively. Finally, Section 7 concludes this paper.

2 Prior Art

2.1 Evaluating Ordinal Classification

As we have mentioned in Section 1, Task 4 Subtask C of the SemEval-2016/2017 workshops is an OC task with five ordinal classes (Nakov et al., 2016; Rosenthal et al., 2017). While SemEval also features other OC tasks with fewer classes (e.g., Task 4 Subtask A from the same years, with three classes), we use the Subtask C data as having more classes should enable us to see more clearly the difference between measures that consider ordinal classes and those that do not.[2] Note that if there are only two classes, OC is reduced to nominal classification. Subtask C used two evaluation measures that consider the ordinal nature of the classes: *macroaveraged Mean Absolute Error* (MAE^M) and the standard *Mean Absolute Error* (MAE^μ) (Baccianella et al., 2009).

At ACL 2020, Amigó et al. (2020) proposed a measure specifically designed for OC, called *Closeness Evaluation Measure* (CEM^{ORD}), and discussed its axiomatic properties. Their meta-evaluation experiments primarily focused on comparing it with other measures in terms of how each measure agrees simultaneously with all of preselected "gold" measures. However, while their results showed that CEM^{ORD} is similar to all of these gold measures, the outcome may differ if we choose a different set of gold measures. Indeed, in the context of evaluating information retrieval evaluation measures, Sakai and Zeng (2019) demonstrated that a similar meta-evaluation approach called *unanimity* (Amigó et al., 2018) depends heavily on the choice of gold measures. Moreover, while Amigó et al. (2020) reported that CEM^{ORD} also performs well in terms of consistency of system rankings across different data (which they refer to as "robustness"), experimental details were not provided in their paper. Hence, to complement their work, the present study conducts extensive and reproducible experiments for OC measures. Our OC meta-evaluation experiments cover nine measures, including MAE^M, MAE^μ, and CEM^{ORD}.

2.2 Evaluating Ordinal Quantification

As we have mentioned in Section 1, Task 4 Subtask E of the SemEval-2016/2017 workshops is an OQ task with five ordinal classes (Nakov et al., 2016; Rosenthal et al., 2017).[3] Subtask E used *Earth Mover's Distance* (EMD), remarking that this is "currently the only known measure for ordinal quantification" (Nakov et al., 2016; Rosenthal et al., 2017). Subsequently, however, Sakai (2018a) proposed a new suite of OQ measures based on *Order-aware Divergence* (OD),[4] and compared them with *Normalised Match Distance* (NMD), a normalised version of EMD. Sakai utilised data from the Third Dialogue Breakdown Detection Challenge (DBDC3) (Higashinaka et al., 2017), which features three ordinal classes, and showed that his *Root Symmetric Normalised OD* (RSNOD) measure behaves similarly to NMD. However, his experiments relied on the run submission files from his own team, as he did not have access to the entire set of DBDC3 submission files. On the other hand, the organisers of DBDC3 (Tsunomori et al., 2020) compared RSNOD, NMD, and the official measures of DBDC (namely, Mean Squared Error and Jensen-Shannon Divergence, which ignore the ordinal nature of the classes) using all the run submission files from DBDC3. They reported that RSNOD was the overall winner in terms of system ranking consistency and *discriminative power*, i.e., the ability of a measure to obtain many statistical significant differences (Sakai, 2006, 2007, 2014).

In addition to the aforementioned two Subtask E data sets from SemEval, the present study utilises three data sets from the Dialogue Quality (DQ) Subtasks of the recent NTCIR-15 Dialogue Evaluation (DialEval-1) Task (Zeng et al., 2020). Each DQ subtask is defined as: given a helpdesk-customer dialogue, estimate the probability distribution over the five-point Likert-scale Dialogue Quality ratings (See Section 4). Our OQ meta-evaluation experiments cover six measures, including NMD and RSNOD.

[2] SemEval-2018 Task 1 (Affect in Tweets) featured an OC task with four classes (Mohammad et al., 2018). However, the run submission files of this task are not publicly available.

[3] The *Valence Ordinal Classification* subtask of SemEval-2018 Task 1 (Affect in Tweets) is also an OQ task, with seven classes (Mohammad et al., 2018). However, the submission files of this task are not publicly available.

[4] See also Sakai (2017) for an earlier discussion on OD.

3 Evaluation Measure Definitions

3.1 Classification Measures

In the OC tasks of SemEval-2016/2017, a set of topics was given to the participating systems, where each topic is associated with N tweets. (N varies across topics.) Given a set C of ordinal classes represented by consecutive integers, each OC system yields a $|C| \times |C|$ confusion matrix for each topic. From this, we can calculate evaluation measures described below. Finally, the systems are evaluated in terms of *mean* scores over the topic set.

Let c_{ij} denote the number of items (e.g., tweets) whose true class is j, classified by the system into i $(i, j \in C)$ so that $N = \sum_j \sum_i c_{ij}$. Let $c_{\bullet j} = \sum_i c_{ij}$, $c_{i\bullet} = \sum_j c_{ij}$, and $C^+ = \{j \in C \mid c_{\bullet j} > 0\}$. That is, C^+ is the set of gold classes *that are not empty*. We compute MAE's as follows.

$$MAE^M = \frac{1}{|C^+|} \sum_{j \in C^+} \frac{\sum_{i \in C} |i-j| c_{ij}}{c_{\bullet j}} , \quad (1)$$

$$MAE^\mu = \frac{\sum_{j \in C} \sum_{i \in C} |i-j| c_{ij}}{N} . \quad (2)$$

Unlike the original formulation of MAE^M by Baccianella et al. (2009), ours explicitly handles cases where there are empty gold classes (i.e., j s.t. $c_{\bullet j} = 0$). Empty gold classes actually do exist in the SemEval data used in our experiments.

It is clear from the weights used above ($|i-j|$) that MAEs assume *equidistance*, although this is not guaranteed for ordinal classes. Hence Amigó et al. (2020) propose the following alternative:

$$CEM^{ORD} = \frac{\sum_{j \in C} \sum_{i \in C} prox_{ij} c_{ij}}{\sum_{j \in C} prox_{jj} c_{\bullet j}} , \quad (3)$$

where $prox_{ij} = -\log_2(\max\{0.5, K_{ij}\}/N)$, and

$$K_{ij} = \begin{cases} c_{\bullet i}/2 + \sum_{l=i+1}^{j} c_{\bullet l} & (i \leq j) \\ c_{\bullet i}/2 + \sum_{l=j}^{i-1} c_{\bullet l} & (i > j) \end{cases} . \quad (4)$$

Our formulation of $prox_{ij}$ with a max operator ensures that it is a finite value even if $K_{ij} = 0$.

We also consider *Weighted* κ (Cohen, 1968). We first compute the expected agreements when the system and gold labels are independent: $e_{ij} = c_{i\bullet} c_{\bullet j}/N$. Weighted κ is then defined as:

$$\kappa = 1 - \frac{\sum_{j \in C} \sum_{i \in C} w_{ij} c_{ij}}{\sum_{j \in C} \sum_{i \in C} w_{ij} e_{ij}} , \quad (5)$$

where w_{ij} is a predefined weight for penalising misclassification. In the present study, we follow the approach of MAEs (Eqs. 1-2) and consider *Linear* Weighted κ: $w_{ij} = |i-j|$. However, it should be noted here that κ is not useful if the OC task involves baseline systems such as the ones included in the aforementioned SemEval tasks: that is, a system that always returns Class 1, a system that always returns Class 2, and so on. It is easy to mathematically prove that κ returns a zero for all topics for all such baseline systems.

We also consider applying *Krippendorff's* α (Krippendorff, 2018) to OC tasks. The α is a measure of data label reliability, and can handle any types of classes by plugging in an appropriate distance function. Instead of the $|C| \times |C|$ confusion matrix, the α requires a $|C| \times N$ class-by-item matrix that contains label counts $n_i(u)$, which represents the number of labels which say that item u belongs to Class i. For an OC task, $n_i(u) = 2$ if both the gold and system labels for u is i; $n_i(u) = 1$ if either the gold or system label (but not both) for u is i; $n_i(u) = 0$ if neither label says u belongs to i. Thus, this matrix ignores which labels are from the gold data and which are from the system.

For comparing *two* complete sets of labels (one from the gold data and the other from the system), the definition of Krippendorff's α is relatively simple. Let $n_i = \sum_u n_i(u)$; this is the total number of labels that Class i received from the two sets of labels. The *observed coincidence* for Classes i and j $(i, j \in C, i \neq j)$ is given by $O_{ij} = \sum_u n_i(u) n_j(u)$, while the *expected coincidence* is given by $E_{ij} = n_i n_j/(2N-1)$. The α is defined as:

$$\alpha = 1 - \frac{\sum_i \sum_{j>i} O_{ij} \delta_{ij}^2}{\sum_i \sum_{j>i} E_{ij} \delta_{ij}^2} , \quad (6)$$

where, for ordinal data,

$$\delta_{ij}^2 = (\sum_{k=i}^{j} n_k - \frac{n_i + n_j}{2})^2 , \quad (7)$$

and for interval data, $\delta_{ij}^2 = |i-j|^2$ (Krippendorff, 2018). We shall refer to these two versions of α as α-ORD and α-INT, respectively. Unlike κ, the α's can evaluate the aforementioned baseline systems without any problems.

The three measures defined below ignore the ordinal nature of the classes. That is, they are *axiomatically incorrect* as OC evaluation measures.

First, let us consider two different definitions of "Macro F1" found in the literature (Opitz and Burst, 2019): to avoid confusion, we give them different names in this paper. For each $j \in C^+$, let $Prec_j = c_{jj}/c_{j\bullet}$ if $c_{j\bullet} > 0$, and $Prec_j = 0$ if $c_{j\bullet} = 0$ (i.e., the system never chooses Class j). Let $Rec_j = c_{jj}/c_{\bullet j}$. Also, for any positive values p and r, let $f1(p,r) = 2pr/(p+r)$ if $p+r > 0$, and let $f1(p,r) = 0$ if $p = r = 0$. Then:

$$F1^M = \frac{1}{|C^+|} \sum_{j \in C^+} f1(Prec_j, Rec_j) \,. \qquad (8)$$

Now, let $Prec^M = \sum_{j \in C^+} Prec_j/|C^+|$, $Rec^M = \sum_{j \in C^+} Rec_j/|C^+|$, and

$$HMPR = f1(Prec^M, Rec^M) \,. \qquad (9)$$

HMPR stands for *Harmonic mean of Macroaveraged Precision and macroaveraged Recall*. Opitz and Burst (2019) recommend what we call $F1^M$ over what we call HMPR. Again, note that our formulations use C^+ to clarify that empty gold classes are ignored.

Finally, we also consider Accuracy:[5]

$$Accuracy = \frac{\sum_{j \in C} c_{jj}}{N} \,. \qquad (10)$$

From Eqs. 2 and 10, it is clear that MAE$^\mu$ and Accuracy ignore *class imbalance* (Baccianella et al., 2009), unlike the other measures.

3.2 Quantification Measures

In an OQ task, a comparison of an estimated distribution and the gold distribution over $|C|$ ordinal classes yields one effectiveness score, as described below. The systems are then evaluated by *mean* scores over the test *instances*, e.g., topics (Nakov et al., 2016; Rosenthal et al., 2017) or dialogues (Zeng et al., 2019, 2020). Let p_i denote the estimated probability for Class i, so that $\sum_{i \in C} p_i = 1$. Similarly, let p^*_i denote the true probability. We also denote the entire probability distributions by p and p^*, respectively.

Let $cp_i = \sum_{k \le i} p_k$, and $cp^*_i = \sum_{k \le i} p^*_k$. *Normalised Match Distance* (NMD) used in the NTCIR Dialogue Quality Subtasks (Zeng et al., 2019, 2020) is given by (Sakai, 2018a):

$$NMD(p, p^*) = \frac{\sum_{i \in C} |cp_i - cp^*_i|}{|C| - 1} \,. \qquad (11)$$

This is simply a normalised version of EMD used in the OQ tasks of SemEval (See Section 2.2) (Nakov et al., 2016; Rosenthal et al., 2017).

We also consider two measures that can handle OQ tasks from Sakai (2018a). First, a *Distance-Weighted sum of squares* for Class i is defined as:

$$DW_i = \sum_{j \in C} |i - j|(p_j - p^*_j)^2 \,. \qquad (12)$$

Note that the above assumes equidistance. Let $C^* = \{i \in C | p^*_i > 0\}$. That is, C^* is the set of classes with a positive gold probability. *Order-aware Divergence* is defined as:

$$OD(p \parallel p^*) = \frac{1}{|C^*|} \sum_{i \in C^*} DW_i \,, \qquad (13)$$

with its symmetric version $SOD(p, p^*) = (OD(p \parallel p^*) + OD(p^* \parallel p))/2$. *Root (Symmetric) Normalised Order-aware Divergence* is defined as:

$$RNOD(p \parallel p^*) = \sqrt{\frac{OD(p \parallel p^*)}{|C| - 1}} \,, \qquad (14)$$

$$RSNOD(p, p^*) = \sqrt{\frac{SOD(p, p^*)}{|C| - 1}} \,. \qquad (15)$$

The other three measures defined below ignore the ordinal nature of the classes (Sakai, 2018a); they are *axiomatically incorrect* as OQ measures. *Normalised Variational Distance* (NVD) is essentially the Mean Absolute Error (MAE):

$$NVD(p, p^*) = \frac{1}{2} \sum_{i \in C} |p_i - p^*_i| \,. \qquad (16)$$

Root Normalised Sum of Squares (RNSS) is essentially the Root Mean Squared Error (RMSE):

$$RNSS(p, p^*) = \sqrt{\frac{\sum_{i \in C} (p_i - p^*_i)^2}{2}} \,. \qquad (17)$$

The advantages of RMSE over MAE is discussed in Chai and Draxler (2014).

The *Kullback-Leibler divergence* (KLD) for system and gold probability distributions over classes is given by:

$$KLD(p \parallel p^*) = \sum_{i \in C \text{ s.t. } p_i > 0} p_i \log_2 \frac{p_i}{p^*_i} \,. \qquad (18)$$

As this is undefined if $p^*_i = 0$, we use the more convenient *Jensen-Shannon divergence* (JSD) instead, which is symmetric (Lin, 1991):

$$JSD(p, p^*) = \frac{KLD(p \parallel p^M) + KLD(p^* \parallel p^M)}{2} \,, \qquad (19)$$

where $p^M_i = (p_i + p^*_i)/2$.

[5] Since the system and the gold data have the same total number of items to classify (i.e., N), accuracy is the same as *microaveraged* F1/recall/precision.

Short name in this paper	Evaluation venue	Task/subtask	Task Type	language	#ordinal classes	test data sample size	#runs used
Sem16T4C	SemEval-2016	Task 4 Subtask C	OC	E	5	100	12
Sem17T4C	SemEval-2017	Task 4 Subtask C	OC	E	5	125	20
Sem16T4E	SemEval-2016	Task 4 Subtask E	OQ	E	5	100	12
Sem17T4E	SemEval-2017	Task 4 Subtask E	OQ	E	5	125	14
DQ-{A, E, S}	NTCIR-15 (2020)	DialEval-1 DQ	OQ	C+E	5	300	22 (13+9)

Table 1: Task data used in our OC and OQ meta-evaluation experiments (C: Chinese, E: English).

(I) Sem16T4C	α-INT	HMPR	F1^M	MAE^M	κ	CEM^{ORD}	Accuracy	MAE$^\mu$
α-ORD	1.000	0.818	0.879	0.818	0.879	0.606	-0.030	-0.152
α-INT	-	0.818	0.879	0.818	0.879	0.606	-0.030	-0.152
HMPR	-	-	0.818	0.879	0.939	0.606	-0.030	-0.152
F1^M	-	-	-	0.818	0.879	0.727	0.091	-0.030
MAE^M	-	-	-	-	0.879	0.727	0.091	-0.030
κ	-	-	-	-	-	0.667	0.030	-0.091
CEM^{ORD}	-	-	-	-	-	-	0.364	0.242
Accuracy	-	-	-	-	-	-	-	0.879
(II) Sem17T4C	α-INT	HMPR	F1^M	MAE^M	κ	CEM^{ORD}	Accuracy	MAE$^\mu$
α-ORD	0.989	0.821	0.821	0.789	0.758	0.695	0.453	0.337
α-INT	-	0.811	0.811	0.800	0.768	0.684	0.442	0.326
HMPR	-	-	0.895	0.926	0.832	0.789	0.526	0.432
F1^M	-	-	-	0.863	0.768	0.789	0.526	0.453
MAE^M	-	-	-	-	0.842	0.842	0.579	0.484
κ	-	-	-	-	-	0.747	0.463	0.368
CEM^{ORD}	-	-	-	-	-	-	0.716	0.600
Accuracy	-	-	-	-	-	-	-	0.863

Table 2: System ranking similarity in terms of Kendall's τ for each OC task. Correlation strengths are visualised in colour ($\tau \geq 0.8$, $0.6 \leq \tau < 0.8$, and $\tau < 0.6$) to clarify the trends.

4 Task Data

Table 1 provides an overview of the SemEval and NTCIR task data that we leveraged for our OC and OQ meta-evaluation experiments. From SemEval-2016/2017 Task 4 (Sentiment Analysis in Twitter) (Nakov et al., 2016; Rosenthal et al., 2017), we chose Subtask C as our OC tasks, and Subtask E as our OQ tasks for the reason given in Section 2.1.[6] Moreover, for the OQ meta-evaluation experiments, we also utilise the DQ (Dialogue Quality) subtask data from NTCIR-15 DialEval-1 (Zeng et al., 2020). As these subtasks require participating systems to estimate three different dialogue quality score distributions, namely, *A-score* (task accomplishment), *E-score* (dialogue effectiveness), and *S-score* (customer satisfaction), we shall refer to the subtasks as DQ-A, DQ-E, and DQ-S hereafter. We utilise both Chinese and English DQ runs for our OQ meta-evaluation (22 runs in total), as the NTCIR task evaluates all runs using gold distributions that are based on the Chinese portion of the parallel dialogue corpus (Zeng et al., 2020). As the three NTCIR data sets are larger than the two SemEval data sets both in terms of sample size and the number of systems, we shall focus on the OQ meta-evaluation results with the NTCIR data; the results with Sem16T4E and Sem17T4E can be found in the Appendix.

5 Meta-evaluation with Ordinal Classification Tasks

5.1 System Ranking Similarity

Table 2 shows, for each OC task, the Kendall's τ rank correlation values (Sakai, 2014) between two system rankings for every pair of measures. We can observe that: (A) the α's, the two "Macro F1" measures (F1^M and HMPR), MAE^M and κ produce similar rankings; (B) MAE$^\mu$ and Accuracy (i.e., the two measures that ignore class imbalance) produce similar rankings, which are *drastically* different from those of Group A; and (C) CEM^{ORD} produces a ranking that is substantially different from the above two groups, although the ranking is closer to those of Group A. The huge gap between Groups A and B strongly suggests that MAE$^\mu$ and Accuracy are not useful even as secondary measures for evaluating OC systems.

It should be noted that the SemEval 2016/2017 Task 4 Subtask C actually reported MAE$^\mu$ scores in addition to the primary MAE^M scores, and the

[6] We do not use the Arabic data from 2017 as only two runs were submitted to Subtasks C and E (Rosenthal et al., 2017).

Measure	Mean τ	Measure	Mean τ
Sem16T4C			
(a) Full split (50 vs. 50) $V_{E2} = 0.00211$		(b) 10 vs. 10 $V_{E2} = 0.00730$	
κ	0.976♯	α-ORD	0.872♠
α-ORD	0.962♭	κ	0.868♠
α-INT	0.935♣	α-INT	0.863♠
MAE$^{\mu}$	0.929♣	HMPR	0.799♣
HMPR	0.904♢	MAE^{M}	0.780♡
MAE^{M}	0.901♢	F1^{M}	0.758‡
F1^{M}	0.884‡	MAE$^{\mu}$	0.753‡
CEM^{ORD}	0.806	Accuracy	0.625†
Accuracy	0.799	CEM^{ORD}	0.595
Sem17T4C			
(c) Full split (62 vs. 63) $V_{E2} = 0.00114$		(d) 10 vs. 10 $V_{E2} = 0.00503$	
α-ORD	0.910♠	HMPR	0.768♠
F1^{M}	0.908♠	α-INT	0.761♣
α-INT	0.907♠	F1^{M}	0.760♣
HMPR	0.901♣	κ	0.751♡
CEM^{ORD}	0.871‡	α-ORD	0.742♡
MAE$^{\mu}$	0.869‡	CEM^{ORD}	0.729♢
κ	0.866‡	MAE$^{\mu}$	0.700†
MAE^{M}	0.850†	MAE^{M}	0.697†
Accuracy	0.818	Accuracy	0.663

Table 3: System ranking consistency for the OC tasks. ♯/♭/♠/♣/♡/♢/ ‡ /† means "statistically significantly outperforms the worst 8/7/6/5/4/3/2/1 measure(s)," respectively. V_{E2} is the residual variance computed from each 1000×9 trial-by-measure matrix of τ scores, which can be used for computing effect sizes. For example, from Part (a), the effect size for the difference between α-ORD and CEM^{ORD} can be computed as $(0.962 - 0.806)/\sqrt{0.00211} = 3.40$.

system rankings according to these two measures were completely different even in the official results. For example, in the 2016 results (Table 12 in Nakov et al. (2016)), while the baseline run that always returns *neutral* is ranked at 10 among the 12 runs according to MAE^{M}, the same run is ranked at the top according to MAE$^{\mu}$. Similarly, in the 2017 results (Table 10 in Rosenthal et al. (2017)), a run ranked at 10 (tied with another run) among the 20 runs according to MAE^{M} is ranked at the top according to MAE$^{\mu}$. Our results shown in Table 2 generalise these known discrepancies between the rankings.

5.2 System Ranking Consistency

For each measure, we evaluate its system ranking consistency (or "robustness" (Amigó et al., 2020)) across two topic sets as follows (Sakai, 2021): (1) randomly split the topic set in half, produce two system rankings based on the mean scores over each topic subset, and compute a Kendall's τ score for the two rankings; (2) repeat the above 1,000 times and compute the mean τ; (3) conduct a randomised paired Tukey HSD test at $\alpha = 0.05$ with 5,000 trials on the mean τ scores to discuss statistical significance.[7]

Table 3 (a) and (c) show the consistency results with the OC tasks. For example, Part (a) shows that when the 100 topics of Sem16T4C were randomly split in half 1,000 times, κ statistically significantly outperformed all other measures, as indicated by a "♯." Table 3 (b) and (d) show variants of these experiments where only 10 topics are used in each topic subset, to discuss the robustness of measures to small sample sizes. If we take the averages of (a) and (c), the top three measures are the two α's and κ, while the worst two measures are CEM^{ORD} and Accuracy; we obtain the same result if we take the averages of (b) and (d). Thus, although Amigó et al. (2020) reported that CEM^{ORD} performed well in terms of "robustness," this is not confirmed in our experiments.

Recall that κ has a practical inconvenience: it cannot distinguish between baseline runs that always return the same class. While SemEval16T4C contains one such run (which always returns *neutral*), SemEval17T4C contains as many as five such runs (each always returning one of the five ordinal classes). This is probably why κ performs well in Table 3(a) and (b) but not in (c) and (d).

5.3 Discriminative Power

In the information retrieval research community, *discriminative power* (Sakai, 2006, 2007, 2014) is a widely-used method for comparing evaluation measures (e.g., Anelli et al. (2019); Ashkan and Metzler (2019); Chuklin et al. (2013); Clarke et al. (2020); Golbus et al. (2013); Lu et al. (2016); Kanoulas and Aslam (2009); Leelanupab et al. (2012); Robertson et al. (2010); Valcarce et al. (2020)). Given a set of systems, a p-value for the difference in means is obtained for every system pair (preferrably with a multiple comparison procedure (Sakai, 2018b)); highly discriminative measures are those than can obtain many small p-values. While highly discriminative measures are not necessarily *correct*, we do want measures to be sufficiently discriminative so that we can draw some useful conclusions from experiments. Again, we use randomised paired

[7] The Tukey HSD (Honestly Significant Differences) test is a multiple comparison procedure: that is, it is like the t-test, but can compare the means of more than two systems while ensuring that the familywise Type I error rate is α. The randomised version of this test is free from assumptions such as normality and random sampling from a population (Sakai, 2018b).

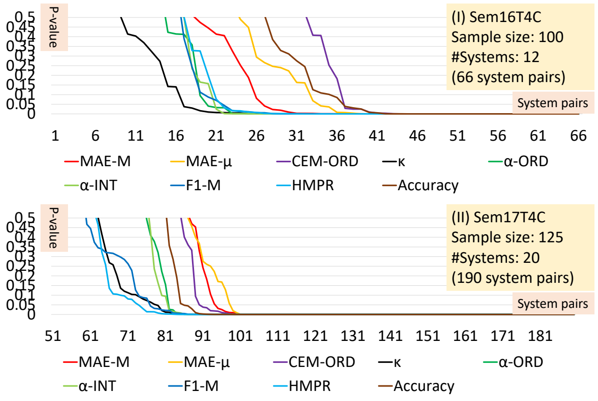

Figure 1: Discriminative power with randomised Tukey HSD tests ($B = 5,000$ trials) for each OC task.

	(I)	(II)	(III)	(IV)	(V)	(VI)
κ	✓		✓	A	**good**	**good**
α-ORD α-INT	✓	✓	✓	A	**good**	fair
CEM^{ORD}	✓	✓	✓	A/B	poor	poor
MAE^{M}	✓	✓	✓	A	fair	poor
F1^{M}		✓	✓	A	fair	fair
HMPR		✓	✓	A	fair	fair
MAE$^{\mu}$	✓	✓		B	fair	poor
Accuracy		✓		B	poor	poor

Table 4: Summary of the properties of OC measures. (I) handles ordinal classes; (II) handles systems that always return the same class; (III) handles class imbalance; (IV) system ranking similarity; (V) system ranking consistency; (VI) discriminative power.

Tukey HSD tests with 5,000 trials for obtaining the p-values.

Figure 1 shows the discriminative power curves for the OC tasks. Curves that are closer to the origin (i.e., those with small p-values for many system pairs) are considered good. We can observe that (i) CEM^{ORD}, Accuracy, MAE^{M}, and MAE$^{\mu}$ are the least discriminative measures in both tasks. (ii) Among the other measures that perform better, κ performs consistently well. Again, the fact that κ distinguishes itself from others in the SemEval16T4C results probably reflects the fact that the data set contains only one run that always returns the same class, which cannot be handled properly by κ.

5.4 Recommendations for OC Tasks

Table 4 summarises the properties of the nine measures we examined in the context of OC tasks. Column (IV) shows that, for example, the Group A measures produce similar rankings. Based on this table, we recommend (Linear Weighted) κ as the primary measure for OC tasks if the tasks do not involve multiple baseline runs that always return the same class. Such runs are unrealistic, so this limitation may not be a major problem. On the other hand, if the tasks do involve such baseline runs (as in SemEval), we recommend α-ORD as the primary measure. In either case, it would be good to use both κ and α-ORD to examine OC systems from multiple angles. According to our consistency and discriminative power experiments, using α-INT instead of α-ORD (i.e., assuming equidistance) does not seem beneficial for OC tasks.

(I) DQ-A	RSNOD	RNOD	NVD	JSD	NMD
RNSS	0.835	0.913	0.939	0.905	0.636
RSNOD	-	0.870	0.861	0.827	0.766
RNOD	-	-	0.939	0.939	0.723
NVD	-	-	-	0.931	0.680
JSD	-	-	-	-	0.714
(II) DQ-E	**RSNOD**	**RNOD**	**NVD**	**JSD**	**NMD**
RNSS	0.931	0.922	0.913	0.913	0.688
RSNOD	-	0.957	0.948	0.948	0.758
RNOD	-	-	0.957	0.991	0.749
NVD	-	-	-	0.948	0.758
JSD	-	-	-	-	0.758
(III) DQ-S	**RSNOD**	**RNOD**	**NVD**	**JSD**	**NMD**
RNSS	0.861	0.974	0.957	0.922	0.558
RSNOD	-	0.887	0.887	0.853	0.662
RNOD	-	-	0.983	0.948	0.584
NVD	-	-	-	0.965	0.584
JSD	-	-	-	-	0.619

Table 5: System ranking similarity in terms of Kendall's τ for each OQ task (NTCIR). Correlation strengths are visualised in colour ($\tau \geq 0.9$, $0.8 \leq \tau < 0.9$, and $\tau < 0.8$) to clarify the trends.

6 Meta-evaluation with Ordinal Quantification Tasks

6.1 System Ranking Similarity

Table 5 shows, for each OQ task from NTCIR, the Kendall's τ between two system rankings for every pair of measures. It is clear from the "NMD" column that NMD is an outlier among the six measures. In other words, among the only axiomatically correct measures for OQ tasks, RNOD and RSNOD are the ones that produce rankings that are similar to those produced by well-known measures such as JSD and NVD (i.e., normalised MAE; see Eq. 16). Also, in Table 5(I) and (III), it can be observed that the ranking by RSNOD lies somewhere between that by NMD (let us call it "Group X") and those by the other measures ("Group Y"). However, this is not true in Table 5(II), nor with our SemEval results (See Appendix Table 8).

DQ-A			
(a) Full split (150 vs. 150) $V_{E2} = 0.00130$		(b) 10 vs. 10 $V_{E2} = 0.00871$	
RNOD	0.909♣	JSD	0.558♣
RNSS	0.885‡	RNOD	0.507♢
NVD	0.882‡	NVD	0.497♢
JSD	0.879‡	RNSS	0.456‡
RSNOD	0.820†	NMD	0.424†
NMD	0.717	RSNOD	0.404
DQ-E			
(c) Full split (150 vs. 150) $V_{E2} = 0.000519$		(d) 10 vs. 10 $V_{E2} = 0.00403$	
NMD	0.865♣	JSD	0.624♣
JSD	0.842♡	RNOD	0.610♡
RNSS	0.835♢	RNSS	0.594‡
NVD	0.819‡	NVD	0.592‡
RNOD	0.813	RSNOD	0.563†
RSNOD	0.811	NMD	0.502
DQ-S			
(e) Full split (150 vs. 150) $V_{E2} = 0.00105$		(f) 10 vs. 10 $V_{E2} = 0.00656$	
RNSS	0.906♡	JSD	0.580♣
JSD	0.901♢	RNOD	0.547♡
RNOD	0.897♢	NVD	0.523‡
NVD	0.870‡	RNSS	0.514‡
RSNOD	0.861†	RSNOD	0.448†
NMD	0.745	NMD	0.421

Table 6: System ranking consistency for the OQ tasks (NTCIR). ♣/♡/♢/ ‡ /† means "statistically significantly outperforms the worst 5/4/3/2/1 measure(s)," respectively. V_{E2} is the residual variance computed from each 1000×6 trial-by-measure matrix of τ scores, which can be used for computing effect sizes. For example, Part (a), the effect size for the difference between RNOD and NMD can be computed as $(0.909 - 0.717)/\sqrt{0.00130} = 5.33$ (i.e., over five standard deviations apart).

6.2 System Ranking Consistency

Table 6 shows the system ranking consistency results with the OQ tasks from NTCIR. These experiments were conducted as described in Section 5.2. If we take the averages of (a), (c), and (e) (i.e., experiments where the 300 dialogues are split in half), the worst measure is NMD, followed by RSNOD. Moreover, the results are the same if we take the averages of (b), (d), and (f) (i.e., experiments where two disjoint sets of 10 dialogues are used), we obtain the same result. Hence, among the axiomatically correct measures for OQ tasks, RNOD appears to be the best in terms of system ranking consistency, and that introducing symmetry (Compare Eqs. 14 and 15) may not be a good idea from a statistical stability point of view. Note that, for comparing a system distribution with a gold distribution, symmetry is not a requirement.

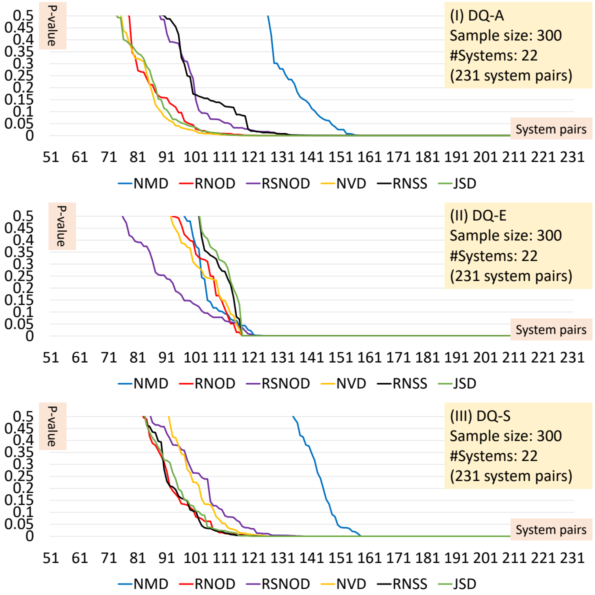

Figure 2: Discriminative power with randomised Tukey HSD tests ($B = 5,000$ trials) for each OQ task from NTCIR.

6.3 Discriminative Power

Figure 2 shows the discriminative power curves for the OQ tasks from NTCIR. We can observe that: (i) NMD performs extremely poorly in (I) and (III), which is consistent with the full-split consistency results in Table 6(a) and (e); (ii) RNOD outperforms RSNOD in (I) and (III). Although RSNOD appears to perform well in (II), if we consider the 5% significance level (i.e., 0.05 on the y-axis), the number of statistically significantly different pairs (out of 231) is 117 for RNOD, 116 for RSNOD, NMD, and NVD, and 115 for RNSS and JSD. That is, RNOD performs well in (II) also. These results also suggest that introducing symmetry to RNOD (i.e., using RSNOD instead) is not beneficial.

6.4 Recommendations for OQ Tasks

Table 7 summarises the properties of the six measures we examined in the context of OQ tasks. Column (III) indicates that NMD is an outlier in terms of system ranking. Based on this table, we recommend RNOD as the primary measure of OQ tasks, as evaluating OQ systems do not require the measures to be symmetric. As a secondary measure, we recommend NMD (i.e., a form of Earth Mover's Distance) to examine the OQ systems from a different angle, although its statistical stability (in terms of system ranking consistency and discriminative power) seems relatively unpredictable. Although the NTCIR Dialogue Quality subtasks (Zeng et al.,

	(I)	(II)	(III)	(IV)	(V)
NMD	✓	✓	X	poor	poor
RSNOD	✓	✓	Y	poor	fair
RNOD	✓		Y	**good**	fair
NVD		✓	Y	**good**	fair
RNSS		✓	Y	**good**	fair
JSD		✓	Y	**good**	fair

Table 7: Summary of the properties of OQ measures. (I) handles ordinal classes; (II) symmetric; (III) system ranking similarity; (IV) system ranking consistency; (V) discriminative power.

2019, 2020) have used NMD and RSNOD as the official measures, it may be beneficial for them to replace RSNOD with RNOD.

7 Conclusions

We conducted extensive evaluations of nine measures in the context of OC tasks and six measures in the context of OQ tasks, using data from SemEval and NTCIR. As we have discussed in Sections 5.4 and 6.4, our recommendations are as follows.

OC tasks Use (Linear Weighted) κ as the primary measure if the task does not involve multiple runs that always return the same class (e.g., one that always returns Class 1, another that always returns Class 2, etc.). Otherwise, use α-ORD (i.e., Krippendorff's α for ordinal classes) as the primary measure. In either case, use both measures.

OQ tasks Use RNOD as the primary measure, and NMD as a secondary measure.

All of our evaluation measure score matrices are available from `https://waseda.box.com/ACL2021PACKOCOQ`, to help researchers reproduce our work.

Among the above recommended measures, recall that Linear Weighted κ and RNOD assume equidistance (i.e., they rely on $w_{ij} = |i - j|$), while α-ORD and NMD do not. Hence, if researchers want to avoid relying on the equidistance assumption (i.e., satisfy the *ordinal invariance* property (Amigó et al., 2020)), α-ORD can be used for OC tasks and NMD can be used for OQ tasks. However, we do not see relying on equidistance as a *practical* problem. For example, note that the Linear Weighted κ is just an instance of the Weighted κ family: if necessary, the weight w_{ij} can be set for each pair of Classes i and j according to practical needs. Similarly, $w_{ij} = |i - j|$ (Eq. 12) for RNOD (and other equidistance-based measures) may be replaced with a different weighting scheme (e.g., something similar to the $prox_{ij}$ weights of CEM^{ORD}) if need be.

Our final and general remark is that it is of utmost importance for researchers to understand the properties of evaluation measures and ensure that they are appropriate for a given task. Our future work includes evaluating and understanding evaluation measures for tasks other than OC and OQ.

Acknowledgement

This work was partially supported by JSPS KAKENHI Grant Number 17H01830.

References

Enrique Amigó, Julio Gonzalo, Stefano Mizzaro, and Jorge Carrillo de Albornoz. 2020. An effectiveness metric for ordinal classification: Formal properties and experimental results. In *Proceedings of ACL 2020*.

Enrique Amigó, Damiano Spina, and Jorge Carrillo de Albornoz. 2018. An axiomatic analysis of diversity evaluation metrics: Introducting the rank-biased utility metric. In *Proceedings of ACM SIGIR 2018*, pages 625–634.

Vito Walter Anelli, Tommaso Di Noia, Eugenio Di Sciascio, Claudio Pomo, and Azzurra Ragone. 2019. On the discriminative power of hyper-parameters in cross-validation and how to choose them. In *Proceedings of ACM RecSys 2019*, pages 447–451.

Azin Ashkan and Donald Metzler. 2019. Revisiting online personal search metrics with the user in mind. In *Proceedings ACM SIGIR 2019*, pages 625–634.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Evaluation measures for ordinal regression. In *Proceedings of ISDA 2009*, pages 283–287.

T. Chai and R.R. Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7:1247–1250.

Aleksandr Chuklin, Pavel Serdyuov, and Maarten de Rijke. 2013. Click model-based information retrieval metrics. In *Proceedings of ACM SIGIR 2013*, pages 493–502.

Charles L.A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. 2020. Offline evaluation without gain. In *Proceedings of ICTIR 2020*, pages 185–192.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Peter B. Golbus, Javed A. Aslam, and Carles L.A. Clarke. 2013. Increasing evaluation sensitivity to diversity. *Information Retrieval*, 16:530–555.

Ryuichiro Higashinaka, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Nobuhiro Kaji. 2017. Overview of Dialogue Breakdown Detection Challenge 3. In *Proceedings of Dialog System Technology Challenge 6 (DSTC6) Workshop*.

Evangelos Kanoulas and Javed A. Aslam. 2009. Empirical justification of the gain and discountfunction for nDCG. In *Proceedings of ACM CIKM 2009*, pages 611–620.

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology (Fourth Edition)*. SAGE Publications.

Teerapong Leelanupab, Guido Zuccon, and Joemon M. Jose. 2012. A comprehensive analysis of parameter settings for novelty-biased cumulative gain. In *Proceedings of ACM CIKM 2012*, pages 1950–1954.

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Xiaolu Lu, Alistair Moffat, and J. Shane Culpepper. 2016. The effect of pooling and evaluation depth on IR metrics. *Information Retrieval Journal*, 19(4):416–445.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Fabrizio Sebastiani. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California. Association for Computational Linguistics.

Juri Opitz and Sebastian Burst. 2019. Macro F1 and macro F1.

Stephen E. Robertson, Evangelos Kanoulas, and Emine Yilmaz. 2010. Extending average precision to graded relevance judgements. In *Proceedings of ACM SIGIR 2010*, pages 603–610.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada. Association for Computational Linguistics.

Tetsuya Sakai. 2006. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of ACM SIGIR 2006*, pages 525–532.

Tetsuya Sakai. 2007. Alternatives to bpref. In *Proceedings of ACM SIGIR 2007*, pages 71–78.

Tetsuya Sakai. 2014. Metrics, statistics, tests. In *PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173)*, pages 116–163. Springer.

Tetsuya Sakai. 2017. Towards automatic evaluation of multi-turn dialogues: A task design that leverages inherently subjective annotations. In *Proceedings of EVIA 2017*, pages 24–30.

Tetsuya Sakai. 2018a. Comparing two binned probability distributions for information access evaluation. In *Proceedings of ACM SIGIR 2018*, pages 1073–1076.

Tetsuya Sakai. 2018b. *Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power*. Springer.

Tetsuya Sakai. 2021. On the instability of diminishing return IR measures. In *Proceedings of ECIR 2021 Part I (LNCS 12656)*, pages 572–586.

Tetsuya Sakai and Zhaohao Zeng. 2019. Which diversity evaluation measures are "good"? In *Proceedings of ACM SIGIR 2019*, pages 595–604.

Yuiko Tsunomori, Ryuichiro Higashinaka, Tetsuro Takahashi, and Michimasa Inaba. 2020. Selection of evaluation metrics for dialogue breakdown detection in dialogue breakdown detection challenge 3 (in Japanese). *Transactions of the Japanese Society for Artificial Intelligence*, 35(1).

Daniel Valcarce, Alejandro, Bellogín, Javier Parapar, and Pablo Castells. 2020. Assessing ranking metrics in top-N recommendation. *Information Retrieval Journal*, 23:411–448.

Zhaohao Zeng, Sosuke Kato, and Tetsuya Sakai. 2019. Overview of the NTCIR-14 short text conversation task: Dialogue quality and nugget detection subtasks. In *Proceedings of NTCIR-14*, pages 289–315.

Zhaohao Zeng, Sosuke Kato, Tetsuya Sakai, and Inho Kang. 2020. Overview of the NTCIR-15 dialogue evaluation task (DialEval-1). In *Proceedings of NTCIR-15*, pages 13–34.

Appendix

For completeness, this appendix reports on the OQ experiments based on SemEval16T4E and SemEval17T4E, which we omitted in the main body of the paper. However, we view the OQ results based on the three NTCIR data sets as more reliable than these additional results, as the SemEval score matrices are much smaller than those from NTCIR (See Table 1).

Table 8 shows the system ranking similarity results with SemEval16T4E and SemEval17T4E; this table complements Table 5 in the paper.

Table 9 shows the system ranking consistency results with SemEval16T4E and SemEval17T4E; this table complements Table 6 in the paper.

Figure 3 shows the discriminative power curves for SemEval16T4E and SemEval17T4E; this figure complements Figure 2 in the paper.

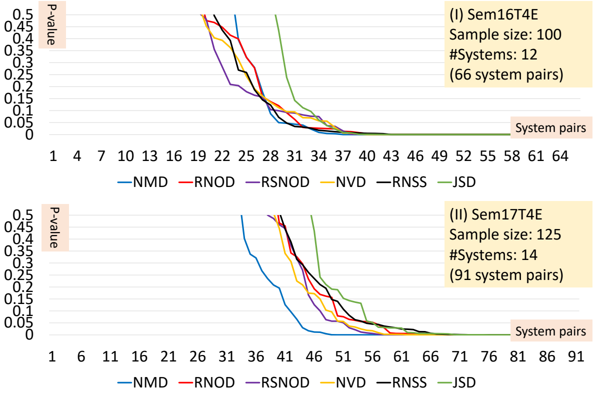

Figure 3: Discriminative power with randomised Tukey HSD tests ($B = 5,000$ trials) for each OQ task (SemEval).

(I) Sem16T4E	RSNOD	RNOD	NVD	JSD	NMD
RNSS	0.848	0.909	0.818	0.788	0.818
RSNOD	-	0.939	0.909	0.879	0.848
RNOD	-	-	0.909	0.879	0.909
NVD	-	-	-	0.970	0.939
JSD	-	-	-	-	0.909
(II) Seml17T4E	RSNOD	RNOD	NVD	JSD	NMD
RNSS	0.912	0.912	0.890	0.868	0.780
RSNOD	-	1.000	0.978	0.956	0.868
RNOD	-	-	0.978	0.956	0.868
NVD	-	-	-	0.978	0.890
JSD	-	-	-	-	0.912

Table 8: System ranking similarity in terms of Kendall's τ for each OQ task (SemEval). Correlation strengths are visualised in colour ($\tau \geq 0.9$, $0.8 \leq \tau < 0.9$, and $\tau < 0.8$) to clarify the trends.

Measure	Mean τ	Measure	Mean τ
(I) Sem16T4E			
Full split (50 vs. 50) $V_{E2} = 0.00175$		10 vs. 10 $V_{E2} = 0.00368$	
JSD	0.934♣	JSD	0.771♣
RNOD	0.847♡	RNOD	0.708♢
NVD	0.831♢	NVD	0.705♢
RNSS	0.815‡	RNSS	0.690‡
NMD	0.788†	NMD	0.674
RSNOD	0.767	RSNOD	0.673
(II) Sem17T4E			
Full split (62 vs. 63) $V_{E2} = 0.00107$		10 vs. 10 $V_{E2} = 0.00342$	
NMD	0.905♣	NMD	0.705♣
NVD	0.878♡	JSD	0.672♡
JSD	0.867♢	NVD	0.601♢
RSNOD	0.859‡	RNOD	0.588†
RNOD	0.826†	RSNOD	0.583†
RNSS	0.765	RNSS	0.557

Table 9: System ranking consistency for the OQ tasks (SemEval). ♣/♡/♢/ ‡ /† 5 4 3 2 1 means "statistically significantly outperforms the worst 5/4/3/2/1 measure(s)," respectively. V_{E2} is the residual variance computed from each 1000×6 split-by-measure matrix of τ scores, which can be used for computing effect sizes. For example, from (I) Left, the effect size for the difference between JSD and RNOD can be computed as $(0.934-0.847)/\sqrt{0.00175} = 2.08$ (i.e., about two standard deviations apart).