The Many Benefits of Annotator Rationales for Relevance Judgments

Tyler McDonnell[†], Mucahid Kutlu[⋆], Tamer Elsayed[⋆], and Matthew Lease[‡]
[†]Dept. of Computer Science, University of Texas at Austin, USA
[⋆]Dept. of Computer Science and Engineering, Qatar University, Qatar
[‡]School of Information, University of Texas at Austin, USA
[†]tmcdonnell@utexas.edu, [⋆]{mucahidkutlu, telsayed}@qu.edu.qa, [‡]ml@utexas.edu

Abstract

When collecting subjective human ratings of items, it can be difficult to measure and enforce data quality due to task subjectivity and lack of insight into how judges arrive at each rating decision. To address this, we propose requiring judges to provide a specific type of *rationale* underlying each rating decision. We evaluate this approach in the domain of Information Retrieval, where human judges rate the relevance of Webpages. Cost-benefit analysis over 10,000 judgments collected on Mechanical Turk suggests a win-win: *experienced* crowd workers provide rationales with no increase in task completion time while providing further benefits, including more reliable judgments and greater transparency[1].

1 Introduction

Ensuring data quality remains a significant challenge in crowdsourcing [Kittur *et al.*, 2013], especially with paid microtask platforms such as Mechanical Turk (MTurk) in which inexpert, remote, unknown annotators are provided only rudimentary communication channels and training. The annotation process is opaque, with only the final labels being observable. The key idea of rationales [Zaidan *et al.*, 2007] is to ask human annotators to provide justifications for their labeling decisions in a particular, constrained form. As with Zaidan *et al.* [2007], we emphasize that the idea of rationales generalizes beyond the particular annotation task or form of rationale used. However, while rationales were originally conceived merely to support a specific machine learning goal (with trusted annotators assumed), we hypothesize that rationales offer far broader applicability and potential benefits.

We ground our investigation of annotator rationales in the specific Information Retrieval (IR) task of *relevance assessment*, which calls on human judges to rate the relevance of *documents* (e.g., Webpages) to search queries. We ask *assessors* to provide a rationale for each judgment by copy-and-pasting a short document excerpt (2-3 sentences) supporting their judgment. Table 2 shows examples. To collect relevance judgments, we created three task designs. Our Standard Task collects relevance judgments without rationales and slightly outperforms prior work [Hosseini *et al.*, 2012] without any use of Honey-Pot questions or platform-specific worker filtering mechanisms. Our Rationale Task achieves further improvement by asking judges to provide rationales; the submitted rationales themselves are completely ignored. Finally, our Two-Stage Task asks one judge to complete the Rationale Task, then a second *reviewer* to verify or fix that judgment. With the same number of workers and task cost, the Two-Stage Task yields further improvement in quality. We also present a heuristic algorithm for exploiting similarity in rationales selected by different workers (Section 5).

We believe that rationales stand to promote greater transparency and trust in crowdsourcing. Our analysis conducted over 10,000 MTurk judgments[2] shows the practical effectiveness of our approach. We also believe that rationales offer a myriad of further benefits (Section 2).

2 Motivations for Annotator Rationales

Enhancing transparency. As discussed earlier, annotator rationales offer a simple, concise, and light-weight form of communication to explain a given answer and demonstrate that it represents a thoughtful decision. When a worker disagrees with "expert" opinion or accepted gold for objective tasks, a rationale can help establish the validity of an alternative answer or reveal errors in the gold standard. For subjective tasks in which answer quality can be difficult to directly evaluate or verify, rationales provide a focused context to interpret a given answer and assess whether it is plausible.

Enhancing quality. Collecting rationales may also help to encourage more thoughtful decision making and discourage any temptation to cheat. When one need only provide a label, it is rather easy to click and be done without giving the task much thought. However, when one is forced to provide a rationale for one's decisions, greater care and reflection is needed. *We hypothesize that creating a plausible rationale for a randomly-selected answer would be at least as effortful as simply undertaking the task in good faith*. Moreover, because

[1]The full vesion of this work appears in [McDonnell *et al.*, 2016].

[2]http://github.com/tylermcdonnell/WhyIsThatRelevant

rationales can be checked relatively easily (even for subjective tasks), *we hypothesize this will reduce the temptation to cheat* (due to greater perceived risk of getting caught).

Enabling crowd verification. Rationales also create a new opportunity for utilizing iterative task design in the spirit of Find-Fix-Verify [Bernstein *et al.*, 2010]. While labels alone do not provide sufficient information for such iterative refinement, rationales could enable one worker's label and/or rationale to be further revised or refined by a subsequent worker (Section 4). Moreover, because rationales make it easier to verify worker answers, there is increased opportunity for delegating such verification tasks to the crowd.

Improving aggregation. As in Zaidan *et al.* [2007], collecting rationales generally enables *dual-supervision* of a learner over rationales and labels. In the context of crowdsourcing, while there has been work on label aggregation [Sheshadri and Lease, 2013], we are not familiar with any work proposing dual-supervision for aggregation. In this paper, we present a heuristic algorithm to filter judgments based on rationale overlap prior to aggregation (Section 5).

3 Related Work

Effective task design. Alonso [2009] recommends collecting optional, free-form, task-level feedback from workers. In contrast, we assume rationales are required, constrained, and example-specific. Because rationales are strictly-defined, it is possible to provide clear instructions about what is expected (e.g., in our work, a document extract of specified length). Moreover, because rationales are document extracts, they enable dual-supervision, as in Zaidan *et al.* [2007]'s work, and can provide additional domain-specific value (e.g., in our task, implicitly marking relevant document passages).

Relevance judging and agreement. To create a useful gold-standard to train and evaluate IR systems, relevance judges are typically instructed to assess a simplified form of *topical relevance* which ignores various factors, such as redundancy in search results, the searcher's prior knowledge about the topic, etc. [Voorhees, 2001]. For 25 years, NIST TREC (`trec.nist.gov`) has organized shared task evaluations and collected and shared relevance judgment datasets to support IR evaluation [Voorhees *et al.*, 2005].

4 Task Design

We describe 3 task designs: a Standard Task (no rationales), a Rationale Task, and a Two-Stage Rationale Task.

Standard task. We selected a balanced, quaternary scale with the following named categories: {*Definitely Not Relevant, Probably Not Relevant, Probably Relevant, Definitely Relevant*}. Another important design decision was to avoid reliance on any platform-specific worker filtering, geographic restrictions, or honey-pot verification questions. We set task payment at \$0.05 (roughly \$6.00/hr) for all task types.

Rationale task. Our rationale task extends the standard task to also request a rationale from the document of roughly 2-3 sentences in length to support the worker's decision.

Two-Stage rationale task. We deployed a sequential task to collect a relevance judgment and rationale from a single judge, with four *reviewers* then asked to confirm or modify it.

Algorithm 1 Threshold Filtering

```
1: procedure FILTER-BY-THRESHOLD(J_d)
2:     T ← SELECT-THRESHOLD(J_d)
3:     selected ← ∅
4:     for each (j_1, j_2) ∈ COMBINATIONS(J_d, 2) do
5:         if SIMILARITY(j_1, j_2) ≥ T then
6:             selected ← selected ∪ j_1 ∪ j_2
7:     return selected
F
8: procedure SELECT-THRESHOLD(J_d)
9:     T ← 0
10:    for each (j_1, j_2) ∈ COMBINATIONS(J_d, 2) do
11:        T ← max(T, SIMILARITY(j_1, j_2))
12:    return ROUND-DOWN(T, 10)
```

5 Filtering Judgments by Rationale Overlap

Assuming our task design motivates workers to quickly find clear rationales for their judgments, maximizing per-task compensation, we hypothesize that judges will tend to converge on similar document extracts as rationales: one of the first plausible rationales found in a document. We exploit such correlation between *overlap* in rationales and judging accuracy by filtering out judgments whose rationales exhibit poor overlap with other annotators. (Algorithm 1) computes the similarity between each pair of rationales provided for a document, computes a similarity threshold T, and selects all judgments whose rationales have a similarity score with at least one other rationale that is $\geq T$.

6 Evaluation

6.1 Experimental Setup

We collect *ad hoc* Web search relevance judgments for the ClueWeb09 dataset (`lemurproject.org/clueweb09`). Search topics and judgments are drawn from the 2009 TREC Web Track [Clarke *et al.*, 2010]. We select 700 documents to judge from different topics covering 43 of the 50 topics in the Web Track. We collect 5 crowd responses per Webpage (700x5=3500 judgments) for each task design: Standard, Rationale, and Two-Stage and evaluate against both the included ternary gold standard and a binarized version in which we collapse *relevant* and *highly relevant* distinctions.

6.2 Individual and Consensus Accuracy

In addition to measuring simple accuracy to evaluate the quality of crowd judgments vs. TREC gold, we also adopt Cohen's Kappa κ_C [Carletta, 1996; Artstein and Poesio, 2008; Bailey *et al.*, 2008], which accounts for chance in measuring agreement between two raters. Cohen's Weighted κ incorporates weights for treating disagreements differently, so we can assign "partial credit" for almost-correct answers.

Table 1 shows binary and ternary quality of crowd judgments, as measured by both simple accuracy and Cohen's κ_C, reported for individual judgments and consensus induced from aggregating 5 judgments. Our Standard Task is intended to serve as a strong baseline vs. prior work, and its binary accuracy of 86% actually outperforms the 80-82% binary accuracy achieved by Hosseini *et al.* [2012]'s careful task de-

				Binary		Ternary	
Row	**Task**	**Filter**	**Judgments**	**Accuracy**	**Cohen's κ_C**	**Accuracy**	**Cohen's κ_C**
1	Standard	-	Single Judge	0.65	0.36	0.47	0.34
2	Rationale	-	Single Judge	0.80	0.51	0.64	0.50
3	Two-Stage	-	Judge + Reviewer	0.85	0.58	0.75	0.60
4	Standard	-	5 Judges (EM)	0.86	0.59	0.75	0.46
5	Rationale	-	5 Judges (EM)	0.92	0.80	0.84	0.80
6	Rationale	THRESHOLD	5 Judges (EM)	0.96	0.85	0.91	0.84
7	Two-Stage	-	Judge + 4 Reviewers (EM)	0.96	0.85	0.91	0.85

Table 1: Quality of judgments obtained vs. TREC gold using different task designs (Standard, Rationale, and Two-Stage) and individual vs. aggregate judging, measuring simple accuracy vs. Cohen's Weighted Kappa κ_C for binary vs. ternary relevance.

sign. Moreover, unlike them, we do not rely on any HoneyPot questions or platform-specific worker filtering.

6.3 Filtering Judgments via Rationale Overlap

Using THRESHOLD FILTERING method (Algorithm 1), we observe accuracy gains across the board. Consensus results using THRESHOLD filtering (Row 6) vs. no filtering (Row 5) show binary judging of 96% accuracy & $\kappa_C = 0.85$ vs. 92% accuracy & $\kappa_C = 0.80$ and ternary judging of 91% accuracy & $\kappa_C = 0.84$ vs. 84% accuracy & $\kappa_C = 0.80$. This indicates that accurate assessors do select similar document extracts as rationales, indicating a correlation between *overlap* in annotator rationales and judging accuracy.

6.4 Cost-Benefit Analysis of Rationales

While Table 1 shows simple accuracy for the binary relevance of Standard vs. Rationale tasks using either 1 judgment (individual judging) or 5 judgments (aggregate consensus), Figure 1 shows the full range of how accuracy varies across the full range of [1:5] judgments. We randomly sample n judgments (x-axis) and apply MV consensus (EM results were similar), averaging over 20 random trials for each judgment count. Binary accuracy of Standard judging exhibits fairly consistent gains as judgments increase, achieving 86% accuracy with 5 judgments. In contrast, Rationale Judging approaches 90% accuracy with only three judgments.

Figure 2 plots the average time the subset of *experienced* workers (who completed 20 or more total tasks) spent completing each of their first 20 tasks, clearly showing a decrease in task time with more experience. Intuitively, both Standard and Rationale tasks involve overhead for reading instructions and task familiarization. For Standard, we see task time rapidly fall off after this early phase, whereas Rationale task time drops more slowly. However, task time critically converges in both cases for *experienced* workers. We thus hypothesize that both tasks effectively require the same mental processes: reviewing text in order to formulate a relevance decision; the Rationale task simply makes this explicit.

6.5 Two-Stage Task Results

Our Two-Stage Task (Section 4) collects a judgment and a rationale from a single assessor, then asks 4 subsequent *reviewers* to either confirm or modify the initial judgment.

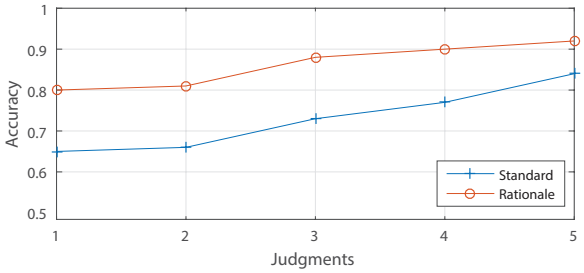

Figure 1: Judging accuracy vs. number of judgments, with MV for aggregation in the case of multiple judgments.

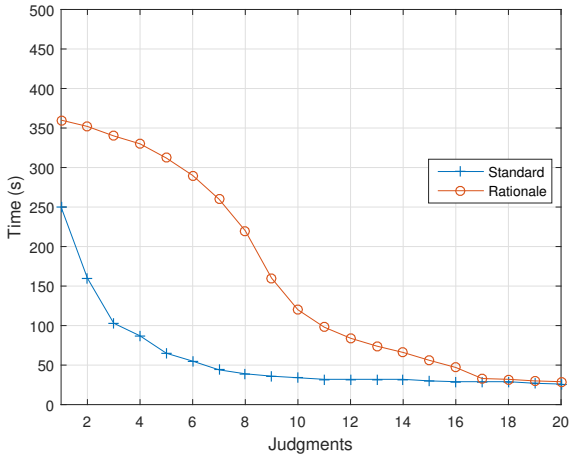

Figure 2: Average completion time vs. completed task count on Standard vs. Rationale tasks for experienced workers.

We found that second-stage reviewers never introduced new judgment errors in the second phase, but fixed an error made by the initial judge 82% of the time. Additionally, Table 1 shows that Two-Stage with 2 judgments (Row 3) achieves much higher accuracy (85% binary, 75% ternary) vs. either Rationale (Row 2) or Standard (Row 1). Most sig-

Query (Alice)	dogs for adoption
Narrative (Alice)	I want to find information on adopting a dog. This includes names and locations of rescue organizations or vehicles (e.g. classifieds) as well as documents with info on qualifications, fees (if any), what to expect, resources, etc. Organizations may be rescue organizations, pounds, shelters, etc. but not breeders or pet shops, unless the pet shop runs adoption fairs.

	Document 1	**Document 2**	**Document 3**
W1 Judgment (Tom)	Probably Not Relevant	Definitely Relevant	Definitely Not Relevant
W1 Rationale (Tom)	*Rooterville Sanctuary. For adoption: pets, pig, pigs, piggy, piggies, pork.*	*View our rescue dogs - visit our organization or contact us directly to see what is available.*	*The dogs listed here all require a new home. These dogs all deserve that second chance and you may be that special person to give it to them. View Rescue Dogs adoption fees. Contact us for more info.*
W2 Judgment	Probably Not Relevant	Probably Relevant	Definitely Relevant
W2 Reasoning	I agree that this organization is probably not likely to be one where Alice will find the animal she is looking for, since they seem to focus on pigs, though they mention dogs	It is a site that lists dog rescue organizations, which is what Alice is searching for. But it is an Australian website. I suspect Alice was looking for an organization in the US.	Tom provided a lot of information that shows why this page should be useful for Alice.
Gold Standard	Probably Not Relevant	Probably Relevant	Definitely Relevant

Table 2: Examples of the Two-Stage Task with worker responses for three different documents.

nificantly, Table 1 shows that Two-Stage with 2 judgments matches Standard's performance with 5 judges (Row 4) with 3 fewer judgments and higher ternary κ_C: 0.60 vs. 0.46.

Qualitative analysis

Table 2 presents a subset of judgments on three documents judged for the same search topic. The Table shows the judgment and rationale provided by the initial annotator, as well as the subsequent reviewer's judgment and reasoning.

Document 1. The judge rated the document to be *Probably Not Relevant*, providing a rationale which suggested that the sanctuary appeared to specialize in pigs, not dogs. The reviewer affirmed this judgment, citing Tom's rationale.

Document 2. The judge indicated *Definitely Relevant* because the website explicitly advertises dog adoptions. However, the reviewer tweaked the judgment to *Probably Relevant*, understanding Tom's justification but noting that the rescue organization is based in Australia and that, "I suspect Alice was looking... in the US." Such transparency of thought is invaluable since there is nothing explicit in the Narrative supporting the reviewer's supposition.

Document 3. The judge selected *Definitely Not Relevant*, but gave a rationale suggesting the website was quite relevant. The reviewer caught this, mentioning Tom's rationale, and suggested the submitted judgment was an accident.

Each example highlights the utility of rationales as a source of transparency and verifiability not possible with traditional relevance judging. In each case, the judge's rationale enabled the reviewer to weigh the judge's reasoning against their own.

7 Conclusion

We believe that forming a rationale is critical to forming a coherent judgment, whether or not task instructions explicitly require it. Our results show that requiring annotators to provide rationales incurs almost no additional time for *experienced* annotators (who complete 20 or more tasks), suggesting that annotators might be already doing so implicitly. By choosing to capture this critical reasoning process, a variety of benefits can be realized to improve transparency of work and quality of data from crowdsourcing, especially for subjective tasks in which multiple answers may be valid.

In future work, we plan to further investigate sequential task iteration beyond two-stages, dynamic collection of judgments based on rationale overlap, dual-supervision of aggregation with rationales, and the validity of using crowdsourcing labels for conducting repeatable, reliable, and rigorous A/B system testing evaluations [Blanco *et al.*, 2011].

Acknowledgments

We thank the many talented crowd contributors who provided the data for our study. This work was made possible by NPRP grant NPRP 7-1313-1-245 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

[Alonso, 2009] Omar Alonso. Guidelines for designing crowdsourcing-based relevance experiments, 2009. CiteSeerX

DOI 10.1.1.149.6649.

[Artstein and Poesio, 2008] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[Bailey *et al.*, 2008] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P de Vries, and Emine Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *SIGIR*, pages 667–674. ACM, 2008.

[Bernstein *et al.*, 2010] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. Soylent: a word processor with a crowd inside. In *UIST*, pages 313–322. ACM, 2010.

[Blanco *et al.*, 2011] Roi Blanco, Harry Halpin, Daniel M Herzig, Peter Mika, Jeffrey Pound, Henry S Thompson, and Thanh Tran Duc. Repeatable and reliable search system evaluation using crowdsourcing. In *SIGIR*, pages 923–932. ACM, 2011.

[Carletta, 1996] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.

[Clarke *et al.*, 2010] Charles L Clarke, Nick Craswell, and Ian Soboroff. Overview of the TREC 2009 Web Track. In *Proceedings of NIST TREC*, 2010.

[Hosseini *et al.*, 2012] Mehdi Hosseini, Ingemar J Cox, Nataša Milić-Frayling, Gabriella Kazai, and Vishwa Vinay. On aggregating labels from multiple crowd workers to infer relevance of documents. In *ECIR*, pages 182–194. Springer, 2012.

[Kittur *et al.*, 2013] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The Future of Crowd Work. In *CSCW*, pages 1301–1318. ACM, 2013.

[McDonnell *et al.*, 2016] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. In *Proc. of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, pages 139–148, 2016. *Best Paper Award*.

[Sheshadri and Lease, 2013] Aashish Sheshadri and Matthew Lease. SQUARE: A Benchmark for Research on Computing Crowd Consensus. In *Proceedings of the AAAI Conference on Human Computation (HCOMP)*, pages 156–164, 2013.

[Voorhees *et al.*, 2005] Ellen M Voorhees, Donna K Harman, et al. *TREC: Experiment and evaluation in information retrieval*. The MIT Press, 2005.

[Voorhees, 2001] Ellen M Voorhees. The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems*, pages 355–370. Springer, 2001.

[Zaidan *et al.*, 2007] Omar F Zaidan, Jason Eisner, and Christine D Piatko. Using annotator rationales to improve machine learning for text categorization. In *HLT-NAACL*, pages 260–267, 2007.