Journal of Artificial Intelligence Research 69 (2020) 143-189 Submitted 03/2020; published 09/2020

Annotator Rationales for Labeling Tasks in Crowdsourcing

Mucahid Kutlu M.KUTLU@ETU.EDU.TR
TOBB University of Economics and Technology
Ankara, Turkey

Tyler McDonnell TMCDONNELL@UTEXAS.EDU
University of Texas at Austin
Austin, TX USA

Tamer Elsayed TELSAYED@QU.EDU.QA
Qatar University
Doha, Qatar

Matthew Lease ML@UTEXAS.EDU
University of Texas at Austin
Austin, TX USA

Abstract

When collecting item ratings from human judges, it can be difficult to measure and enforce data quality due to task subjectivity and lack of transparency into how judges make each rating decision. To address this, we investigate asking judges to provide a specific form of *rationale* supporting each rating decision. We evaluate this approach on an information retrieval task in which human judges rate the relevance of Web pages for different search topics. Cost-benefit analysis over 10,000 judgments collected on Amazon's Mechanical Turk suggests a win-win. Firstly, rationales yield a multitude of benefits: more reliable judgments, greater transparency for evaluating both human raters and their judgments, reduced need for expert gold, the opportunity for *dual-supervision* from ratings and rationales, and added value from the rationales themselves. Secondly, once experienced in the task, crowd workers provide rationales with almost no increase in task completion time. Consequently, we can realize the above benefits with minimal additional cost.

1. Introduction

When asking human judges to provide item ratings, it can be difficult to measure and ensure data quality given task subjectivity and lack of transparency into how rating decisions are made. This is particularly true when using crowdsourcing (Kittur et al., 2013) platforms such as Amazon's Mechanical Turk (MTurk) (Barr & Cabrera, 2006; Vakharia & Lease, 2015) in which inexpert, remote, unknown annotators are provided only rudimentary communication channels and training. The annotation process is largely opaque, with only the final labels being observable. Such factors do little to inspire trust between parties and faith in the overall paradigm. Risks may be seen to outweigh potential benefits, limiting the scale and complexity of tasks for which crowdsourcing is considered viable, and thereby the number of jobs made available to workers. When the accepted practice to ensure data quality requires aggregating responses from multiple workers, the cost of data collection increases and individual worker wages may be reduced. As we consider more subjective

tasks, in which disagreement might be valid, measuring and enforcing data quality becomes even more challenging (Tian & Zhu, 2012; Nguyen et al., 2016).

We believe that *annotator rationales* (Zaidan et al., 2007) offer an opportunity for new traction on the above problems. The key idea of rationales is to ask human annotators to support their labeling decisions in a particular, constrained form: an excerpt from the example being judged. As with Zaidan et al., we emphasize that the idea of rationales generalizes beyond the particular annotation task or form of rationale used (e.g., Donahue & Grauman, 2011, investigate visual rationales for image annotation). However, while Zaidan et al. assumed trusted annotators and proposed rationales only to support dual-supervision in machine learning, we hypothesize that rationales offer a far broader range of potential benefits (Section 2) than has been capitalized upon in prior work (Section 3).

We ground our investigation of annotator rationales in the information retrieval (IR) task of *relevance assessment*, which calls on human judges to rate the relevance of *documents* (e.g., Webpages) to search topics (Cleverdon & Keen, 1966). Unlike simple labeling tasks, relevance is complex phenomenon that researchers continue to study (Saracevic, 2007). Consequently, annotator agreement is often low, even with simplified notions of relevance and trusted judges (Voorhees, 2000; Bailey et al., 2008). While crowdsourcing's potential for more efficient relevance judging has sparked great interest (Alonso et al., 2008), its use has tended to further exacerbate annotator agreement issues.

In this work, we ask *assessors* to provide a rationale for each judgment by copy-and-pasting a short document excerpt (2-3 sentences) supporting their judgment. Following Zaidan et al. (2007), we opt for textual excerpts instead of free-form rationales because this constrained form of rationales enables automatic verification and filtering methods based on textual overlap. To collect relevance judgments, we propose three task designs refined through pilot experiments (Section 4). While our *Standard Task* design collects relevance judgments without rationales, our *Rationale Task* outperforms the Standard Task simply by asking judges to provide rationales; the submitted rationales themselves are completely ignored. Moreover, we find that *experienced* workers (who we define as those completing 20 or more tasks) are able to complete the Rationale Task with almost no increase in average task completion time (29 vs. 27 seconds). In addition, we design two judgment filtering algorithms using textual similarity between rationales. Finally, our *Two-Stage Task* design utilizes the rationale to enable us apply iterative task design principles (Bernstein et al., 2010; Little et al., 2010) to our item rating task: one judge provides the rating with a rationale, then a second *reviewer* verifies or revises the rating based on the rationale.

In sum, we believe rationales offer a myriad of potential benefits for many annotation tasks (Section 2). For our relevance judging task, cost-benefit analysis over 10,000 MTurk relevance judgments suggests a win-win: *experienced* crowd workers provide rationales with almost no increase in task completion time while providing more reliable judgments and greater transparency for evaluating both human raters and their judgments. Further benefits include reduced need for expert gold, the opportunity for dual-supervision from ratings and rationales, and added value from the rationales themselves (Section 2.3).

In comparison to our earlier work (McDonnell et al., 2016), this article greatly expands our presentation of the rationale approach, including: conceptual framing (Section 2), generality of the approach (Section 2.2), and coverage of related work (Section 3), including extensive discussion regarding appropriate payment (Section 3.3). We also present results

for multiple aggregation methods and further analyze our analysis of experienced vs. inexperienced workers (Section 6.3). Our expanded qualitative analysis (Section 7.1) includes *negative rationales* (Section 7.2) and worker feedback (Section 8). We close with an expanded discussion of limitations and a detailed agenda for future work (Section 9).

Regarding contributions, we extend *annotator rationales* (Zaidan et al., 2007) beyond their original conception of supporting dual-supervision, identifying a myriad of further potential benefits. While we focus on use of rationales in crowdsourcing, we also discuss how traditional annotation processes and resultant their datasets stand to benefit from rationales. Beyond relevance judging with text, we discuss generalization of our rationale approach to other types of annotation tasks and use with non-textual data. We show how rationales can bridge previously disparate lines of prior work in crowdsourcing: "simple" labeling tasks (e.g., classification and ratings) and task design for iterative refinement of "complex" responses (e.g., free-text, Bernstein et al., 2010). On the annotation task of human relevance judging, cost-benefit analysis over 10,000 judgments shows a win-win: once familiarized with the rationale task, crowd workers provide far more accurate labels with almost no increase in task completion time. Moreover, accuracy of crowd work provides further evidence that crowdsourcing can be effectively utilized in lieu of traditional annotation practices, in IR and beyond. To improve data quality, many crowdsourcing studies restrict who is allowed to work via platform-specific filters (e.g., by past approval rating, quantity of prior work, or the worker's geographic region). We employ no such restrictions and still achieve high labeling quality. This represents another win-win: we further democratize global access to work and we increase the effective supply of available workers.

The remainder of this paper is organized as follows. Section 2 introduces annotator rationales and a range of potential benefits they offer beyond enabling dual-supervision. Prior work is next reviewed in Section 3. Following this, we present several alternative task designs for crowdsourced collection of human relevance judgments (Section 4). Hypothesizing that accurate judges will tend to converge on similar rationales, Section 5 describes two heuristic methods to exploit this correlation and filter judgments prior to aggregation. Our primary evaluation is presented in Section 6. Section 7 presents a qualitative analysis of collected rationales. Section 8 discusses feedback received from crowd workers regarding our task design. Section 9 discusses limitations of our work and proposes several directions for future work. We conclude in Section 10.

2. Annotator Rationales: What and Why?

In this section, we first provide a brief history of rationales (Section 2.1). Subsequently, we discuss how the rationale approach can be applied for a wide range of tasks (Section 2.2). Moreover, we explain further benefits of rationales such as enhancing transparency and enabling crowd verification (Section 2.3).

2.1 A Brief History of Rationales

The *credit-assignment* problem in supervised machine learning is that a model must infer which part(s) or feature(s) of input x best explain its assigned output label y. To address this, Zaidan et al. (2007) and Zaidan and Eisner (2008) proposed that annotators not only provide the label y for example x, but further explain each label by marking a supporting

rationale: a portion of x which best explains the assigned label y. Because the rationale is constrained to be part of the example x, this provides better focus in model training as to which part(s) of x to focus in learning the desired x to y input-output mapping. As such, rationales simplify the learning problem by narrowing data scope.

Zaidan et al. (2007) grounded this work in text-classification, specifically binary sentiment analysis of movie reviews. Annotators were instructed, beyond judging the sentiment of a review, to also mark words or phrases which supported their assigned label:

> You will be asked to justify why a review is positive or negative. To justify why a review is positive, highlight the most important words and phrases that would tell someone to see the movie. To justify why a review is negative, highlight words and phrases that would tell someone not to see the movie. These words and phrases are called "rationales". You can highlight the rationales as you notice them, which should result in several rationales per review. Do your best to mark enough rationales to provide convincing support...

It was expected that such constrained annotations might be easy to provide:

> ...an annotator who is categorizing phrases or documents might also be asked to highlight a few substrings that significantly influenced her judgment... As long as the annotator is spending the time to study example x and classify it, it may not require much extra effort for her to mark reasons for her classification.

In fact, rationales doubled the annotation time in their study. Zaidan et al. also used traditional (on-site, trusted, and expert) annotators – an author and two students – and did not consider the question of how asking annotators to provide these rationales might impact the quality of the annotations being collected, especially from laymen.

Recent work on neural models has continued to investigate use of such constrained rationales to both supervise models (Zhang et al., 2016; Bao et al., 2018) and evaluate unsupervised models (Lei et al., 2016; Bastings et al., 2019).

2.2 Generality of Rationale Approach

Like Zaidan et al. (2007), we investigate rationales in the context of a text-based task. However, the key idea of annotating rationales extends far beyond text-based tasks. For example, Zaidan and Eisner (2008) write that, "In the visual domain, when classifying an image as containing a zoo, the annotator might circle some animals or cages and the sign reading 'Zoo'." Donahue and Grauman (2011) put this into practice for image classification, with annotators marking one or more regions of each image as rationales for their labeling decisions. We imagine rationales could be similarly collected for audio and video classification tasks as well, where a short segment of a larger audio or video clip could serve as rationale for a labeling decision. Such tasks exemplify applicability of rationales to a broad class of *Where's Waldo?*[1] search problems of determining whether or not a given item contains something of interest (e.g., does Waldo appear in a given image or video clip, do we hear his voice in a given audio recording, is he discussed in a given text, etc.). The larger the example, the greater the search problem. For example, while Donahue and

1. https://en.wikipedia.org/wiki/Where%27s_Wally%3F

Grauman (2011) work with relatively small imagery, the approach could be extended to massive satellite or aerial imagery in which annotators might need to zoom far into images in order to find a suitable rationale (e.g., to locate Waldo).

Framing rationales as a search problem (i.e., for evidence to support a labeling decision) lets us relate rationales to a large body of related work mobilizing the crowd for distributed search of large search spaces. Classic examples include the search for extraterrestrial intelligence[2], for Jim Gray's sailboat (Vogels, 2007) or other missing people (Wang et al., 2009), for DARPA's red balloons (Tang et al., 2011), for astronomical events of interest (Lintott et al., 2008), for endangered wildlife (Rosser & Wiggins, 2019) or bird species (Kelling et al., 2013), etc. Attenberg et al. (2011) asked the crowd to find examples on which classifiers erred. Across such examples, what is being sought must be broadly recognizable so that the crowd can accomplish the search task without need for subject matter expertise (Kinney et al., 2008).

Whereas the cases above involve searching across domain instances, note that rationales filter *within* each instance, finding the key evidence in each example relevant to some task. The search problem may be explicit (e.g., *does an audio clip contain a bird call?*, which clearly necessitates searching the example) or implicit (e.g., *rate a product from its description*, where the primary task is to rate but the annotator must search for evidence to support their rating decision). We can thus view the *credit-assignment* problem as a search problem, with Zaidan et al. (2007) shifting this search problem from the model training procedure to human annotators.

Another class of related search problems from which we draw inspiration comes from computational complexity theory (e.g., P vs. NP[3]). In many cases, it is far easier to verify a candidate solution (e.g., a polynomial-time algorithm) than it is to find it. In such cases, the candidate solution provides a shortcut to solving the original problem by significantly narrowing the search space. Similarly, it may be easier for annotators to verify a label from a focused rationale than to determine the appropriate label given the full example. This serves as inspiration for our Two-Stage Task Design (Section 4.4): once the full example is reduced to a smaller rationale, we ease the work for a second-stage human assessor, who can restrict their attention to the rationale in verifying or revising the label.

2.3 Further Benefits of Rationales

Rationales appear to offer a myriad of potential benefits beyond their original conception of supporting dual-supervision (Zaidan et al., 2007). We further explore these here.

2.3.1 Enhancing Transparency

Annotator rationales offer a simple, concise, and light-weight explanation for a given answer and demonstrate it represents a thoughtful decision. Once acquired, rationales stand to benefit all future users of a dataset, not only those who originally collected it. When data from a given study is published, rationales would help others to inspect and assess data quality for themselves. When a worker disagrees with "expert" opinion or accepted gold for objective tasks, a rationale could help establish the validity of an alternative answer or reveal

2. https://setiathome.berkeley.edu/
3. https://en.wikipedia.org/wiki/P_versus_NP_problem

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 (Kutlu et al., 2018). While we follow most prior work in measuring accuracy wrt. a gold standard, we believe it will be particularly valuable in future work to explore the potential of rationales to ensure data quality for subjective tasks (Tian & Zhu, 2012; Nguyen et al., 2016), without reliance on a gold standard.

FigureEight (formerly `CrowdFlower.com`, now part of Appen) already employs rationales in the other direction. Because gold *honey-pot* questions used to evaluate workers are not always clear, correct, or complete, *Requesters* are encouraged to provide a textual "reason" for each correct answer in order to justify it to workers. Workers, in turn, are provided an avenue to appeal if they disagree. In this spirit, collecting rationales also provides a new means of scalable crowd-based creation of honey-pot questions for worker testing: even if the "gold" label is sometimes wrong, the rationale provides a basis on which workers who disagree can appeal (and thereby simultaneously checking correctness of both workers).

2.3.2 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. In addition, when one is paid per-task (rather than hourly) and any answer seems acceptable (e.g., subjective rating tasks), it can be tempting to answer quickly to increase one's effective pay rate. An established practice to discourage such behavior is to design tasks such that it is no easier to create "believable invalid responses" than to undertake the given task in good faith (Kittur et al., 2008). *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.* We test this hypothesis indirectly by evaluating the quality of judgments obtained in the presence or absence of rationales.

Moreover, because rationales can be checked relatively easily, *we hypothesize this will reduce the temptation to cheat* (due to greater perceived risk of getting caught). As above, we test this hypothesis indirectly. Significantly, we would expect that higher quality data might be obtained by simply requiring rationales, *even if they are discarded without inspection.* Because subjective tasks make it difficult to verify answers, Kittur et al. (2008) recommend creating additional, non-task verifiable questions to include (e.g., "What is the third word on this page?"). However, such questions are easily distinguished from real task questions, so they can be easily passed without undertaking the real task questions in good faith (Marshall & Shipman, 2013). In contrast, because rationales are tied to the real task questions of interest, they support more robust verification of data quality.

2.3.3 Enabling Crowd Verification

Rationales also create a new opportunity for utilizing sequential task design approaches in the spirit of Find-Fix-Verify (Bernstein et al., 2010) (Section 3.2). By themselves alone,

labels do not provide enough information for iterative refinement; verifying a label is no easier than completing the task from scratch. However, a rationale's "explanation" for a label could enable one worker's label and/or rationale to be easily verified or improved upon by a subsequent worker (See Section 4.4 and the end of Section 2.2). Rationales could thus help extend the generality of sequential task design to a broader range of common data labeling tasks than they have been traditionally considered applicable for iterative refinement. Moreover, because rationales help to verify worker answers, there is an increased opportunity for effectively delegating label verification to the crowd to reduce "expert" workload.

2.3.4 Improving Aggregation

As shown by 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 plentiful work on label aggregation (Sheshadri & Lease, 2013; Zheng et al., 2017), we are not familiar with any prior work proposing dual-supervision for aggregation. In this paper, we present two very simple heuristic algorithms to do so, filtering judgments based on rationale overlap, then aggregating remaining labels (Section 5).

2.3.5 Additional Domain Value

Finally, rationales themselves may provide additional value in the task domain. For example, whereas traditional document retrieval simply returns entire documents, *passage retrieval* and *question answering* seek to provide searchers with more focused search results than entire documents (Trotman et al., 2007). By requiring assessors to provide a document excerpt supporting a judgment of document relevance, judges effectively annotate relevant passages. While our task design encourages judges to converge on similar rationales (rather than find all relevant passages in a document), future work to support relevance judging for passage retrieval could relax this aspect of our task design and still realize many of the other benefits provided by collecting rationales.

3. Related Work

Many "best practices" have been proposed for effective task design. In early work, Kittur et al. (2008) propose designing tasks such that it is no easier to create "believable invalid responses" than to undertake the given task in good faith. Alonso (2015) describes a development process to collect crowd judgments at scale. Gadiraju et al. (2017) investigate task clarity and propose a method to predict the clarity of the task designs. Wu and Quinn (2017) investigate the effect of best practices for task design on the actual outcomes.

3.1 Justifying Answers

Whereas several prior crowdsourcing studies have asked workers to provide open-ended textual justifications for their answers (Alonso, 2009; Drapeau et al., 2016; Chang et al., 2017; Wang et al., 2019; Han et al., 2020), the *rationales* approach (Zaidan et al., 2007) is differentiated by requiring a far more restricted form of justification. For a given instance

x to be labeled, annotators identify a subset of x as the rationale for their label; there is no free-form text response. Zaidan et al. (2007) asked annotators to mark "important words and phrases" explaining overall movie review sentiment. For image classification, Donahue and Grauman (2011) asked annotators to indicate a region of the image justifying the overall image label.

In early work, Alonso (2009) recommends collecting optional, free-form, task-level feedback from workers. While Alonso found that some workers did provide example-specific feedback, the free-form nature of their feedback request elicited a variety of response types, difficult to check or to invalidate spurious responses. Alonso also found that requiring such feedback led many workers to submit unhelpful text that was difficult to automatically cull. Such feedback was therefore made optional rather than required.

In more recent work, Drapeau et al. (2016) propose a *Justify-Reconsider* method. In the Justify task, "workers provide reasoning with their answer in terms of the task guidelines taught during training." For the Reconsider task, "workers are shown an argument for the opposing answer and then asked to reconfirm their original decision or change their answer." The authors report that their Justify-Reconsider method generally yields higher crowd accuracy for an NLP annotation task, but that requesting justifications requires additional cost. Consequently, they find that simply collecting more crowd annotations yields higher accuracy in a fixed-budget setting. In contrast, with our more restricted rationales, we find that requesting rationales incurs nearly no extra judging time for experienced workers. Of course, this restriction comes with limitations; rationales do not let the worker reference the task guidelines or provide additional information outside the example to support their labeling decision. Consequently, we see a tradeoff between expressiveness and control.

Chang et al. (2017) propose a three-step approach in which crowd workers label the data, provide justifications for cases in which they disagree with others, and then review others' explanations. They evaluate their method on an image labeling task and report that requesting only justifications (without any further processing) does not increase the crowd accuracy. However, as noted above, they request free-form justifications whereas we require stricter document excerpts. While their open-ended text responses can be subjective and difficult to check, the restricted rationales we utilize are more amenable to manual and automatic verification. We have hypothesized that requiring rationales suggests greater accountability to workers than simple labeling tasks without rationales, and thus may reduce potential temptation to cheat.

Kulkarni et al. (2012) provide workers a chat feature that "helps workers maneuver around inadequate explanations in the task, suggest additional examples that could be given to requesters, teach other workers how to use the interface, and confirm their theories about what a task means." Schaekermann et al. (2018) investigate the impact of discussion among crowd workers on the label quality using a chat platform allowing synchronous group discussion. While the chat platform allows workers to better express their justification than text-excerpts, the discussion increases task completion times. In addition, chatting does not impose any restriction on topic, limiting discussion from unenthusiastic workers and efficacy. Manam and Quinn (2018) similarly propose synchronous Q&A between workers and Requesters to allow workers to ask questions to resolve any uncertainty about overall task instructions or specific examples.

Chen et al. (2019) also proposed a workflow allowing simultaneous discussion among crowd workers, and designed task instructions and a training phase to achieve effective discussions. While their method yields high labeling accuracy, the increased cost due to the discussion limits its task scope. While our restricted form of rationale limits the discussion among workers, we found that it causes no additional cost for experienced workers.

Recently, Han et al. (2020) collected relevance judgments and open-ended justifications from crowd workers. They more substantively distinguished between experienced and inexperienced workers and studied the strategies of experienced crowdworkers in task selection and completion such as use of scripts and keyboard shortcuts. They also found that all workers speed up as they complete more tasks, and they compare the speed up in experienced vs. inexperienced workers, in judging relevance and in justifying their judgments.

Most recently, Hasanain et al. (2020) collect rationales for an Arabic Web search IR test collection. Whereas we collected relevance rationales from secondary assessors using crowd judges, they collected rationales from the topic developers themselves and use in-house assessors. As we did, they requested text excerpts as rationales for relevant documents. However, for non-relevant documents, they provided a pre-defined list of possible rationales, based on (Kutlu et al., 2018): e.g., "The page has no relevant information" and "The page has no relevant information, but it has a link that might point to a page with relevant information". In addition, if the pre-defined list does not provide a suitable rationale, the assessor may express their rationale via open-ended text.

3.2 Sequential Task Design

The idea of sequential task design is that a piece of work can be improved through iterative revision (Little et al., 2010; Bernstein et al., 2010; Weld, 2010). For example, one worker drafts a text or a logo, a second worker further refines it, and so on. Such an iterative approach can be applied to any task in which a substantive work product passed along from worker to worker, such that it becomes more effective to revise existing product rather than start over from scratch. The work product effectively acts as an information channel, enabling earlier workers to communicate partial solutions to later workers, thereby iteratively reducing the amount of work remaining in order to converge on a solution.

With simple labeling tasks, however, it is not clear how to apply such iterative task design because a label by itself contains so little information. For example, with document classification, a candidate label cannot be verified or revised without reading the document, which is the same as the original task. A label by itself lacks any evidentiary support. Naively applying iterative design to simple labeling tasks would be effectively equivalent to having multiple independent annotators, since the first stage worker essentially passes along no actionable information to reduce the remaining work for subsequent workers.

A contribution of our work is forging a new bridge between past work on iterative task design and work on simple labeling tasks. Rationales enable simple labeling tasks to benefit from the wealth of work on iterative task design because they enlarge the communication channel between workers; the rationale as a work product serves as an information conduit to ease subsequent labor on the task. Unlike chat approaches (Kulkarni et al., 2012; Schaekermann et al., 2018; Manam & Quinn, 2018), however, the communication channel is intentionally bounded in expressiveness to streamline communications, as a sweet spot

between no communication and fully open-ended communication. Given the rationale, it becomes easier to verify a candidate label than it would be to classify the example from scratch because a worker can focus their attention on the shorter rationale instead of the full-length example.

Note that iterative task design, in which workers iteratively revise an answer, should not be confused with dynamic labeling, or iteratively requesting more worker responses when there is a disagreement (Kamar et al., 2012).

3.3 Appropriate Payment for Microtasks

The question of how much to pay for microtasks on MTurk has a long history of debate, revolving around distinct issues of practical impacts on production (quality, efficiency, and scalability), what is appropriate or ethical pay for "placeless" work completed different physical locales (Mason & Suri, 2012), and U.S. legal classification of independent contractors vs. employees (Wolfson & Lease, 2011). An oft-cited but now dated demographic study (Ross et al., 2010) reported that "On average, Turkers earn just under $2/hr, with Indian workers earning less..." A recent demographic study by Difallah et al. (2018) paid $6/hr (as we do): "The survey... takes on average 30 seconds... we pay 5 US cents."

With regard to work quality, payment can impact who chooses to work on a task and how well they perform the work (Mason & Watts, 2009; Ho et al., 2015). Horton and Chilton (2010) frame the question wrt. the economics notion of *reservation wage*: "...the minimum wage a worker is willing to accept ...for performing some task; it is the key parameter in models of labor supply." Thus as pay decreases, it could fail to match more workers' reservation thresholds and thus potentially bias the sample of workers who choose to perform the task (see discussion below regarding IRB and non-exploitative human subjects research). However, Horton and Chilton find mixed evidence for worker behavior confirming to predictions of the rational model: "workers are clearly sensitive to price but insensitive to variations in the amount of time it takes to complete a task."

While independent contractors are not subject to U.S. minimum wage requirements, debate over this legal classification rages on (Kappel, 2018). Salehi et al. (2015) and Silberman et al. (2018) argue for paying U.S. minimum wage regardless of legal requirements or worker location, and separate recent efforts have explored ways to grow wages by simply making it easier to provide higher pay (Mankar et al., 2017; Whiting et al., 2019).

Another important distinction exists between human subjects research, as overseen by U.S. university institutional review boards (IRBs), and non-human subjects research, such as collecting annotations to train or test AI models. With human-subjects research, participation exposes participants to risk and so selection cannot be exploitative. It would be unethical to reduce study costs by targeting a vulnerable demographic group whose economic need could make them more willing to incur participation risks for low payment.

On the other hand, data processing tasks such as annotation are increasingly being outsourced in our global economy to regions of the world with lower costs of living. Platforms such as MTurk face continued downward pressure on pay rates as competitor vendor workforces[4] set fixed hourly pay rates to workers below U.S. minimum wage while simultaneously providing more stringent data quality control and data privacy guarantees. For example,

4. `https://aws.amazon.com/sagemaker/groundtruth/pricing/`

Amazon SageMaker GroundTruth's popular vendor iMerit[5] charges Requesters $6.12/hour for engaging its managed, secure workforce. Given this, a typical Requester may be unlikely to use MTurk if they can pay a vendor workforce less for better data.

Studies have repeatedly shown that a primary request of MTurk workers is simply for more work (Gray & Suri, 2019). While the rise of AI is increasing overall data annotaton work, there is a risk that this work increasingly shifts away from MTurk (and similar public crowds) toward managed vendor workforces. For non-human subjects research (e.g., typical data annotation), imposing non-market price premiums on MTurk (Whiting et al., 2019) risks further accelerating this trend of driving work away from MTurk. Researchers might consider instead how to help public workforces such as MTurk better compete for data annotation work against managed vendor workforces, to ensure that non-managed working opportunities remain widely available for any capable Internet worker. For example, research might be directed toward showing the means, advantages, and capabilities by which a large, diverse public workforce can outperform a managed vendor workforce.

3.4 Relevance Judging

While the concept of search relevance has been investigated for over 80 years, it remains a thorny phenomenon with many complex and interacting underlying factors (Saracevic, 2007). To create a useful gold-standard to train and evaluate IR systems, relevance judges are typically instructed to assess a more objective, simplified form of *topical relevance* which ignores various factors, such as redundancy in search results, the searcher's prior knowledge about the topic, and others (Voorhees, 2001). Since 1992, NIST TREC[6] has organized shared task evaluations, and collected and shared relevance judgment datasets to support IR evaluation (Voorhees et al., 2005). Trusted relevance judgments are a cornerstone of TREC, and we adopt TREC judgments as our gold standard (Section 6).

Historically, relevance judgments have only rarely been collected from multiple trusted judges for the same document, precluding measurement of inter-annotator agreement. When such data has been collected, relatively low agreement is typical, even with trusted judges (Voorhees, 2000; Bailey et al., 2008). The reasons for disagreement have been investigated using various methods such as interviewing the assessors (Sormunen, 2002; Wakeling et al., 2016) and asking them to think aloud during relevance judging (Al-Harbi & Smucker, 2014). Prior work reports various disagreement reasons such as human error (Grossman & Cormak, 2012), ambiguous topic descriptions (Sormunen, 2002), different perception of relevance (Kutlu et al., 2018), and long, incoherent text (Chandar et al., 2013).

In short, describing precise relevance criteria is difficult, even with simplified *topical relevance*. Precise quantitative comparisons among prior work are also difficult to make due to various judging scales, agreement measures, and datasets which have been reported.

Prior work has mixed findings about the quality of crowd relevance judgments. A number of studies report that crowd judgments can be a reliable alternative for relevance assessment (Alonso & Mizzaro, 2009, 2012; Kazai et al., 2012) or at least do not cause significant errors in ranking of IR systems (Blanco et al., 2011; Kutlu et al., 2018). For instance, Alonso and Baeza-Yates (2010) use crowd-sourced judgments for Spanish sub-

5. https://aws.amazon.com/marketplace/pp/B07DK37Q32
6. http://trec.nist.gov

collection of CLEF (Braschler & Peters, 2002) and report 70% agreement with experts as promising results. On the other hand, there are also studies suggesting being more cautious about when to use judgments of non-trained secondary assessors: Kinney et al. (2008) report that assessors who are not domain experts disagree on the underlying meaning of domain specific queries significantly more often than experts. Bailey et al. (2008) claim that assessors who are neither task nor topic expert might not be a reliable substitute for the topic experts. Clough et al. (2013) find that crowdsourced judgments appear unable to distinguish different levels of highly accurate search results the way expert assessors can.

In perhaps the closest work to our own, Hosseini et al. (2012) compare a sophisticated *full* task design to a baseline *simple* task design for collecting relevance judgments. For each task, 3 MTurk judgments were collected per document, followed by aggregating judgments using Dawid and Skene (1979)'s algorithm. The simple task includes a minimal quality control where a single test question is used to detect respondents providing random answers rather than undertaking the task in good faith. In the full task design, several quality control mechanisms are utilized. Firstly, in order to attract workers interested in or knowledgeable about a particular topic, search topic details are presented in the title, description, and keywords of the tasks. Secondly, each task included 2 test questions. Thirdly, and closest in spirit to rationales, "to enforce the requirement that the workers needed to read a page before deciding about its relevance, a captcha was included asking them to enter the first word of the sentence that confirmed or refuted the relevance of the page." However, the authors provide no detail regarding if and how this captcha was verified. Did the authors retrieve all page sentences beginning with that word, then manually judge themselves if any of those sentences indicated a suitable rationale that was consistent with the judgment rendered? Or did they ask workers to provide this word, to give the impression of work being checked, while actually ignoring the word provided? This point is not clarified. The designs achieve 75% and 82% accuracy, when aggregating 3 judgments per document in simple and full task designs, respectively.

4. Task Design

This section describes three different task designs developed and evaluated in this work: our Standard Task (Section 4.2), which does not collect rationales, our Rationale Task (Section 4.3), and a Two-Stage Rationale Task (Section 4.4). We begin by describing our initial pilot studies (Section 4.1) which iteratively experimented with different designs.

4.1 Pilot Studies

Our iterative design process involved deploying many small-scale relevance judgment tasks which varied key design features, such as the specificity of instructions for the crowd worker, the format of the grading scale, and most importantly, the definition of a rationale. In each iteration, we relied on manual inspection of work to evolve our design. We also included a free-form text box in which workers were encouraged to provide constructive feedback for the task with a possibility of bonus compensation (Alonso, 2009).

Before launching our various studies, we conducted a medium-scale pilot study judging 70 Webpages from ClueWeb09 (see Section 6.1) for 25 search topics drawn from the 2009 TREC Million Query Track (MQT) (Carterette et al., 2010). For each document, we

collected 8 judgments each for Standard and Rationale Tasks. One of the authors blindly judged each document, both as a benchmark for what level of agreement we might expect from the crowd, and to account for potential changes in content, since crowd workers were directed to judge live web pages rather than originally crawled versions of the Webpages judged by TREC (see Section 6.1). Essentially all of the same trends and findings reported in our main study evaluation (Section 6) were observed in this earlier pilot study. In this sense, our second, main study implicitly shows that our findings are reproducible, similar to Blanco et al. (2011)'s work.

However, besides the scale of this pilot study being relatively small and some parameters for filtering (Section 5) being tuned on pilot study data, a significant problem we encountered in the pilot was an inexplicable problem with the MQT gold judgments. While the crowd was internally consistent and consistent with the author's judgments, neither were consistent with the TREC judgments for reasons we could never explain. Consequently, for our main study, we abandoned the MQT data in favor of the Web Track's data (see Section 6.1). Of relevance to our broader motivation for collecting rationales, despite close analysis of the MQT gold standard by one of the authors, possessing only the judgments and topic narrative provided little insight, whereas if the gold data had included the sort of rationales we motivate here, we imagine this problem might have been resolvable. Section 7 provides further discussion on this issue.

4.2 Standard Task

This section describes the Standard Task used to collect crowdsourced relevance judgments. While Mechanical Turk makes it relatively easy to quickly collect large amounts of data, if one also wants this data to be of high quality, effective task design is both crucial and non-trivial. We believe the design of our Standard Task design represents a strong baseline for how relevance judging might be crowdsourced today. Below we discuss our iterative exploration of design alternatives in developing the task.

4.2.1 No Qualifications or Honey-Pots

An important design decision made from the outset was to avoid reliance on any platform-specific worker filtering. Crowdsourcing research too closely-tied to a particular commercial platform's capabilities (or addressing its peculiar limitations) risks reducing the generality and impact of its findings (Vakharia & Lease, 2015). Some platform features offer unknown black-box behavior (e.g., MTurk's *Master Workers*[7]), while others (e.g., MTurk's *Approval Rating*) have known and significant flaws (Ipeirotis, 2010; Gaikwad et al., 2016). Neither seems like a solid foundation.

In addition, while some Requesters impose geographic restrictions to exclude certain regions, presuming lower quality work, such geographic filtering can represent a biased crutch compensating for lazy design. We embrace crowdsourcing's ideal of providing anyone interested an equal opportunity to work and demonstrate their ability. Our responsibility in task design is to enable this vision.

Gold honey-pots require "expert" effort to create and typically assume objective tasks in which each question has a single, correct answer. Such tests are inherently less applicable for

7. https://www.mturk.com/worker/help#what_is_master_worker

subjective rating tasks. Ideal quality assurance methods would be effective and applicable to both task types.

4.2.2 Relevance Scale

At what granularity of relevance should judgments be collected? Binary judgments are simplest but least informative for system evaluation, and all-or-nothing relevance permits judges no way to indicate borderline relevance, which searchers often encounter in practice. On the other hand, a finer granularity scale such as {*Perfect, Excellent, Good, Fair, Bad*} (Sanderson, 2010) is more informative for system evaluation and more flexible for judges, but requires making more subtle distinctions and borderline decisions between categories. Prior literature appears to offer little guidance regarding how choice of relevance scale interacts with judges' efficiency and effectiveness in executing their work (Tang et al., 1999), though it does explore how the number of relevance categories can impact judgment quality. Tang et al. (1999) recommend using a 7-point scale for relevance judgments. Recent work has investigated very fine-grained scales (Roitero et al., 2018) and the impacts of transformations between judging scales (Han et al., 2019).

In our experiments, we did not seek to reproduce or argue against the results of their study, nor did we use the same relevance scale found in the TREC gold standard. Adoption of the TREC scale would have simplified final evaluation between the data we collected and the gold standard, and one might also assume this gold scale implicitly embodies best practice grown out of past trial-and-error experimentation. However, we found scant past work justifying *why* any particular scale was better than any other; we know of no evidence that the TREC scale is optimal or any set of factors for which it was optimized. While this question is secondary to our primary interest in rationales, we wanted to ensure that this aspect of our design was sensible. Moreover, TREC has traditionally assumed trained judges rather than the crowd, and prior work has suggested a mismatch (Alonso, 2009). Even if the gold standard had k categories, judges might still find it easier to judge on a $k + 1$ scale, and judgments could be post-processed to conflate categories for comparison. Finally, we did not want to tie our task design to an arbitrary gold standard we happened to be evaluating against here.

After iterative experimentation with MTurk judges using a variety of options for scale granularity and relevance category names, we ultimately selected a balanced, quaternary (4-point) scale with the following named categories: {*Definitely Not Relevant, Probably Not Relevant, Probably Relevant, Definitely Relevant*}. Having four degrees of relevance, uniformly spaced across the spectrum of relevance, appears to offer flexibility to judges without overwhelming them, satisfying the so-called "Goldilocks" criterion (offering neither too many nor too few options). Consistent with best practices (Alonso, 2009; Blanco et al., 2011), it was found the best to name relevance categories colloquially and avoid technical jargon (e.g., *marginally relevant*) familiar to trained judges but not intuitive or meaningful to laymen. In so doing, we sought to mitigate bias that might otherwise be incurred by unfamiliar terminology or perceived mismatch in connotation. Final category names used are, by design, both consistent with one another and symmetric with regard to adjectival descriptors and colloquiality. This helped us to simplify task instructions (see below) and appeared to improve judges' comprehension and distinction between relevance categories.

Finally, by excluding a neutral option, there is no "easy-out" for judges when they are unsure, forcing them to actively lean in one direction or the other.

4.2.3 Instructions

We ultimately converged on very concise and self-explanatory task instructions which were less specific. Our early designs experimented with different relevance category names and accompanying specific instructions which showed examples of pages that should fall into each category and why. However, we received feedback suggesting workers were frustrated by this level of instruction, not only because of the extent of reading required, but also because it made them feel unsure of whether a given document would fit under the strict, but ultimately ambiguous, definitions we provided. As a point of comparison, Google (2016)'s own extremely detailed judging guidelines still fall back on the phrase "use your judgment" 22 times. While we initially tried to address these concerns by further clarifying corner cases, the instructions quickly became unwieldy, while a long-tail of new corner cases continued to be found with each new batch of experiments. By adopting the colloquial and self-explanatory relevance scale above, we were able to provide very concise task instructions. **Figure 1** shows instructions for Rationale Task.

4.2.4 Better metrics for rationale similarity

In this work, we used simple character overlap between rationales to measure textual similarity. In the image domain one could similarly measure pixel overlap in bounding boxes (e.g., Donahue & Grauman, 2011). Intuitively, better measures of rationale similarity should lead to better validation and aggregation. To this end, future work on text might investigate embedding.

4.2.5 Payment

As discussed earlier (Section 3.3), the question of how much to pay is complicated. We expected the Rationale Task to take more time than the Standard Task; Zaidan et al. (2007) found it typically required two trusted annotators twice as long to collect rationales in addition to labels. Consequently, our pilot study paid $0.05 for the Standard Task vs. $0.10 for the Rationale Task. However, we were surprised to observe that *experienced* workers (who completed 20 or more tasks) converged to the same average completion time for both tasks, reproduced in our main study (**Figure 3**).

Consequently, our main study paid the same for both tasks, intending to render payment a fixed control variable in explaining any difference observed in work time or quality. That said, a potential confound is that some individual workers may have expected or found the Rationale task to be slower, and altered their behavior in response: avoiding the task, abandoning it, or rushing their work in order to achieve a desired *reservation wage* (Horton & Chilton, 2010). We leave for future work further investigation of this potential confound due to payment. Our task payment of $0.05 for all task types worked out to roughly $6.00 hourly wage for workers who completed at least 20 tasks (i.e., the *experienced workers*).

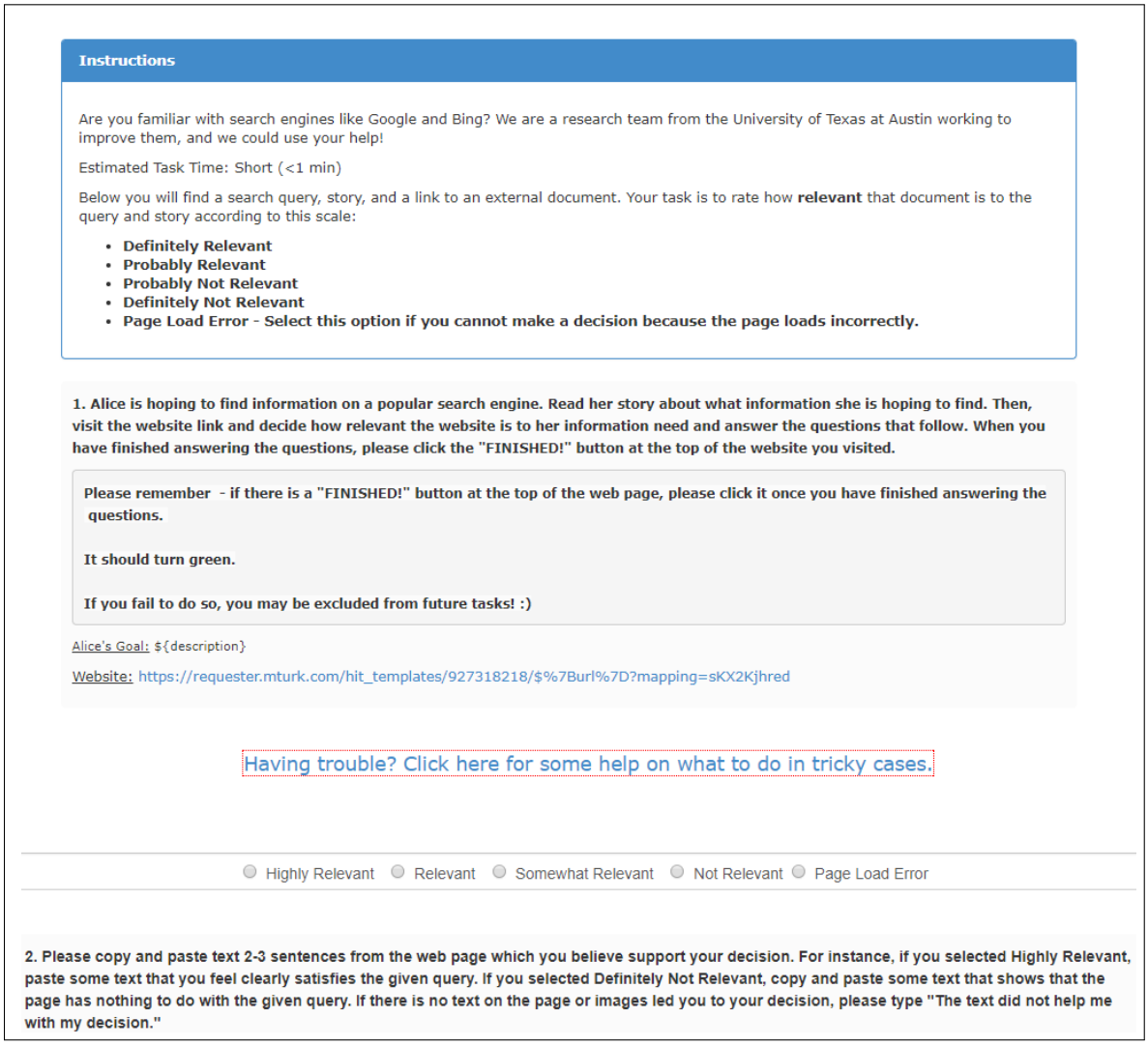

Figure 1: Instructions for Rationale Task.

4.3 Rationale Task

A major goal of our early pilot studies was to experiment with different definitions of rationale in order to determine what worked best. Regarding rationale length, for Zaidan et al. (2007), annotating a few keywords sufficed for the learner; for us, keywords do not provide meaningful insight into the thought process of the annotator. Moreover, keyword rationales might bias judges toward simple *keyword-spotting* (i.e., judging any document containing a query term as relevant). In contrast, we wanted assessors to reflect and provide more complete justifications.

On the other hand, extremely long excerpts would not provide focused insight into a judge's thought process and key elements of their decision-making process. Moreover, if wish to support multi-stage, sequential tasks in which one judge verifies or fixes another's

work (Section 4.4), an overly long rationale might save only minimal time for the second judge vs. simply re-reading the entire document.

Giving no guidance on expected length provided little task clarity for judges and tended to result in overly terse responses, insufficient for either understanding judges' thought processes or automatically analyzing their rationales for overlap-based filtering (Section 5). Consequently, we found that requesting rationales of roughly 2-3 sentences in length frequently provided clear, focused insight into the worker thought process and supported post-processing.

Requiring document excerpts rather than free-form feedback enables one to automatically check (strictly or approximately) if a submitted excerpt is actually found in the document. Moreover, excerpts permit dual-supervision (Zaidan et al., 2007) and can provide additional domain-specific value (e.g., implicitly marking relevant passages). However, textual excperts are not appropriate for all situations. Because resource-type searches and Webpages dominated by imagery provide few useful textual extracts for explaining relevance, we created special instructions for these cases: workers were asked to manually type a fixed string if the document's text did not support their judgment, which was then treated as their rationale. Workers provided this string in roughly 10% of cases.

While restricting rationales to being document excerpts prevents annotators from detailing their full thought process, allowing free-form rationales can yield many spurious responses that are difficult to check and often are not rationales for labeling decisions. One might predefine a set of rationales (e.g., "the document is a spam page, not a source of information") to be used when document excerpts seem insufficient, but this would further complicate matters. Ultimately, we decided to limit rationales to document-excerpts only. We provide further discussion about rationales for non-relevant documents (Section 7.2), free-form rationales (Section 8.1) and graphical rationales (Section 8.2).

4.4 Two-Stage Rationale Task

In the spirit of Find-Fix-Verify (Bernstein et al., 2010), we also designed a two-stage, sequential task which first collected a relevance judgment and rationale from a single judge, and then asked four subsequent *reviewers* to either confirm or modify the initial judgment. Second-stage reviewers were presented with the following scenario:

> Alice was looking for some information. She typed the following search query into a popular search engine. Tom looked at the page seen below and decided whether he thought the page was *Definitely Not Relevant*, *Probably Not Relevant*, *Probably Relevant*, or *Definitely Relevant* to Alice's search. Tom also provided the following quotation from the web page to support his decision. Please answer the following questions:
>
> 1. Do you agree with Tom's assessment of the page for Alice's search? If not, how would you rate the page? (*multiple-choice selection of relevance level*)
> 2. Please describe in words why you agree or disagree with Tom's decision. (*free-form feedback*)

To better understand how reviewers conducted the review process, our second question above requested open-ended justifications (rather than revised rationales). Because the second-stage task did not request a revised rationale, it was not possible to apply our overlap-based filtering methods Top-N and Threshold (Section 5) in conjunction with our two-stage design. As with other task designs, however, we could collect multiple judgments and then perform aggregation (Sheshadri & Lease, 2013; Zheng et al., 2017) in order to induce a single, combined *consensus* judgment for each example.

5. Using Rationales in Label Aggregation

Crowd annotation (Su et al., 2007; Snow et al., 2008) has stimulated a large body of work on *label aggregation*: assigning the same task to multiple annotators and then identifying the best label to keep from the group (Dawid & Skene, 1979; Sheshadri & Lease, 2013; Hung et al., 2013; Zheng et al., 2017). These methods typically assume simple labeling tasks in which the only information available to model is the labels, but what about when we have *rationales* as well? Zaidan et al. (2007) hypothesized that rationales could enable more efficient machine learning through *dual-supervision*: exploiting not only the label y but also its rationale. How might rationales be similarly exploited for better label aggregation?

Weighted voting mechanisms typically estimate each annotator's ability by label agreement: how frequently that annotator's labels match a gold standard (supervised, expert agreement) (Snow et al., 2008) or peer labels (unsupervised agreement) (Dawid & Skene, 1979). Rationales might be similarly utilized by assessing how closely an annotator's rationale for a given example's label matches a gold rationale or peer rationales.

5.1 Measuring Similarity of Rationales

Conceptually, we assume a monotonic text similarity function $\text{SIMILARITY}(r1, r2) \rightarrow [0, 1]$ to measure the similarity between each pair of rationales. Ideally, this function would embody full language understanding, recognizing if two different rationales expressed the same meaning in different ways (e.g., paraphrase detection). Of course, with perfect NLP the system might "read" each rationale and directly assess its support for a given label.

In this work, we adopt a far more modest and expedient approach based on string similarity. Assuming our task design motivates workers to quickly find plausible rationales to justify their labeling decisions (thereby maximizing their per-task compensation), we hypothesize annotators will tend to converge on selecting similar extracts as rationales: those found early in the document. When rationales are constrained to come from the document, follow a relatively prescribed length, and incentivized for selection early in the document, we simplify the problem in practice such that string similarity might be reasonably effective. Our experiments use the following two standard methods.

Jaccard Similarity. The Jaccard (1901) Index is defined as the intersection of n-grams divided by the union of n-grams in two strings. We experimented with a variety of n-gram sizes, which all produced similar results.

Ratcliff-Obershelp. Ratcliff and Metzener (1988) compute the similarity of two strings as the number of matching characters divided by the total number of characters, where matching characters are taken from the longest common subsequence (LCS) and then recursively from the regions on either side of the LCS.

5.2 Filtering Outlier Rationales

In this work, we pursue unsupervised, peer-based aggregation. Given similarity between two rationales, we adopt a two-stage approach to aggregation using labels and rationales.

First, we analyze rationales to detect outliers: rationales exhibiting low similarity with others for the same example. We describe simple two heuristic algorithms below to perform this outlier detection. These outliers are assumed to indicate lower quality labels, and these rationes and labels are discarded. In the second stage, remaining labels are then aggregated using a standard aggregation algorithm (e.g., majority voting, Dawid & Skene, 1979).

Top-N Filtering (Algorithm 1) keeps the top-N rationales which have the highest similarity with others for the same example. For a particular example, we compute the string similarity between all pairs of rationales for that example. We then select the Top-N judgments based on this sorting, breaking ties arbitrarily.

Algorithm 1 Top-N Filtering

```
1: procedure FILTER-BY-TOP-N(J_d, N)
2:     pairs = COMBINATIONS(J_d, 2)
3:     for each pair ∈ pairs do
4:         pair.sim ← SIMILARITY(pair.j_1, pair.j_2)
5:     Sort(pairs) by descending similarity
6:     return GETTOPJUDGMENTS(pairs, N)
```

Threshold Filtering keeps only rationales having similarity score $\geq T$ with at least one other rationale for the given example. The algorithm dynamically determines T for each example by finding the highest similarity among rationale pairs for the example, then rounding this down to the nearest 0.1. We arbitrarily chose this rounded-down value and found it worked well in our pilot study; it was never tuned. Our intuition was to keep labels in the vicinity of the maximum rationale similarity observed, rather than excluding all but the highest N, which our Top-N Filtering algorithm does.

6. Evaluation

We evaluate the utility of annotator rationales for the specific IR task of collecting document relevance judgments, a subjective task which tends to have relatively low annotator agreement (Section 3.4). While relevance in general is quite subjective (Saracevic, 2007), *topical relevance* is intended to be impersonal and more objective (Voorhees, 2001). We thus hypothesize that: 1) *high agreement is possible, provided one is willing to invest enough annotation effort to achieve it*; and that 2) *rationales require relatively little additional effort to achieve higher annotator agreement.*

We first investigate whether collecting rationales during crowdsourced relevance judging can improve the quality of judgments, even if the submitted rationales themselves are completely ignored. To evaluate this, we perform A/B testing of our Standard Task (Section 4.2) vs. our Rationale Task (Section 4.3). We measure inter-annotator agreement (Section 6.2) to

Algorithm 2 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
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, 0.1)
```

test whether the crowd is internally consistent, regardless of their agreement with our gold standard. Next, we measure the accuracy of crowd judgments (individually and aggregated) vs. the TREC gold standard (Section 6.3).

Following this, we evaluate whether accurate judges select similar document extracts as rationales (Section 6.4). Specifically, we evaluate the two methods described in Section 5 for filtering out judgments with low rationale overlap prior to performing aggregation. In Section 6.5, we then report our cost-benefit analysis of Standard vs. Rationale Tasks. Finally, in Section 6.6 we evaluate the Two-Stage Rationale Task in which one assessor's judgment and rationale are reviewed by a second judge for verification or correction.

6.1 Experimental Setup

We collect *ad hoc* Web search relevance judgments for the ClueWeb09 Webcrawl (Callan et al., 2009) using the quaternary (4-point) scale described in Section 4.2. Search *topics* and judgments are drawn from the 2009 TREC Web Track (Clarke et al., 2010). Each topic includes a narrative for the user's *information need* which we provide to judges (See Table 6 for an example). TREC gold relevance judgments use a 3-point scale: *not relevant, relevant* and *highly relevant*.

We select 700 documents to judge from different topics covering 43 of the 50 topics in the Web Track. TREC gold judgments for our 700 documents are distributed as follows: 46% *not relevant*, 24% *relevant*, and 30% *highly relevant*. We evaluate collected crowd judgments against both this ternary gold standard (we collapse our *probably/definitely not relevant* distinctions) and a binarized version (we collapse TREC's *relevant* and *highly relevant* distinctions), yielding 46% *not relevant* and 54% *relevant* documents, and we collapse our own *probably/definitely relevant* distinctions.

While we had planned to judge ClueWeb09's crawled Webpages, images and style sheets associated with each page were often missing or rendered incorrectly, making the rendered pages difficult to assess. Consequently, we instead judged the live web pages associated

with each crawled URL. The 700 URLs we judge exclude all URLs yielding a `Page Not Found` error. Also, because live web pages today may differ from the versions crawled and judged in 2009, one of the authors blindly judged 200 of these URLs. Results closely mirrored the gold standard (95% binary accuracy, 88% ternary accuracy), suggesting that the existing gold judgments are reasonably accurate for the live web pages.

We collect 5 crowd responses per Webpage (700x5=3500 judgments) for each task design: Standard, Rationale, and Two-Stage. We set $N = 3$ for Top-N judgment filtering and rounding down to the nearest 10 for Threshold filtering based on pilot experiments (Section 4.1). For label aggregation, we apply Majority Voting (MV), and Dawid and Skene (1979)'s algorithm. For Dawid-Skene (DS), we adopt an existing open source package[8]. Aggregation is performed at the original quaternary scale, prior to any collapsing.

6.2 Annotator Agreement

We measure agreement using Fleiss' Kappa $\kappa_F = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e}$, where $1 - \bar{P}$ is the agreement attainable by chance and $\bar{P} - \bar{P}_e$ is the degree of agreement achieved above chance. Blanco et al. (2011) question use of Fleiss' Kappa with crowd annotations since it assumes a consistent set of annotators, while the set of crowd annotators per example is rarely consistent since workers come and go. However, Blanco et al. still report mean κ_F, and after exploring a variety of measures for measuring inter-annotator agreement, we found that each told a similar story and none more clearly or simply than Fleiss' Kappa, which we adopt.

Table 1 shows agreement between crowd judges. We observed much higher inter-annotator agreement among judgments collected with Rationale (binary $\kappa_F = 0.79$) and Two-stage (binary $\kappa_F = 0.85$) vs. Standard (binary $\kappa_F = 0.61$). The ternary agreement shows similar trends. *Near-perfect* binary agreement of second stage annotators in Two-Stage is particularly notable, suggesting its design emphasizes critical thinking elements central to making reasonable and consistent judgments. While common practice of aggregating results from 3-5 workers may be necessary with the Standard Task design to remedy its relatively low agreement seen here (a moderate binary κ_F of 0.61 and a fair ternary κ_F of 0.36), we see the Rationale design achieves 0.79 binary agreement with a single worker, and the Two-Stage design achieves 0.85 binary agreement with only two workers.

Judging Task	Binary Agreement	Ternary Agreement
Standard	0.61 (*moderate*)	0.36 (*fair*)
Rationale	0.79 (*substantial*)	0.59 (*moderate*)
Two-Stage	0.85 (*near-perfect*)	0.71 (*substantial*)

Table 1: Annotator Agreement using Fleiss' Kappa κ_F, canonically interpreted assuming 5 equisized bins: *slight, fair, moderate, substantial*, and *near-perfect* (Landis & Koch, 1977).

8. `https://github.com/dallascard/dawid_skene`

6.3 Individual and Aggregated Accuracy

				Binary	
Row	**Task**	**Filter**	**Judgments**	**Accuracy**	**Cohen κ_C**
1	Standard	-	Single Judge	0.65	0.36
2	Rationale	-	Single Judge	0.80	0.51
3	Two-Stage	-	Judge + Reviewer	0.85	0.58
4	Standard	-	5 Judges (MV)	0.84	0.51
5	Rationale	-	5 Judges (MV)	0.92	0.80
6	Rationale	Top-3	5 Judges (MV)	0.93	0.80
7	Rationale	Threshold	5 Judges (MV)	0.96	0.85
8	Two-Stage	-	Judge + 4 Reviewers (MV)	0.96	0.85
9	Standard	-	5 Judges (DS)	0.86	0.59
10	Rationale	-	5 Judges (DS)	0.92	0.80
11	Rationale	Top-3	5 Judges (DS)	0.93	0.81
12	Rationale	Threshold	5 Judges (DS)	0.96	0.85
13	Two-Stage	-	Judge + 4 Reviewers (DS)	0.96	0.85

Table 2: Binary 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.

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 & Poesio, 2008; Bailey et al., 2008), which accounts for chance in measuring agreement between two raters. We treat TREC gold as one rater and either a single crowd judge or aggregated crowd consensus as the other. Cohen's Weighted κ incorporates weights for treating disagreements differently, e.g., assigning partial credit for almost-correct answers in our ordinal judging scale. We adopt a squared weighting function without tuning. While Weighted Kappa seems most appropriate to us with ordinal judging, we note that regular Kappa (not shown) yielded similar results. κ_C agreement can be interpreted similarly to Fleiss' Kappa κ_F above: *slight* [0.00−0.20], *fair* [0.21−0.40], *moderate* [0.41−0.60], *substantial* [0.61−0.80], and *near-perfect* [0.81 − 1.00] (Landis & Koch, 1977).

Table 2 and 3 show 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.

Comparisons to prior work must be made with care to be meaningful, and it can be particularly challenging to fairly compare results of human computation experiments (Paritosh, 2012) because different results in two studies stem from many confounding factors, e.g., using different test collections, well-refined instructions through several pilot tests, col-

				Ternary	
Row	**Task**	**Filter**	**Judgments**	**Accuracy**	**Cohen κ_C**
1	Standard	-	Single Judge	0.47	0.34
2	Rationale	-	Single Judge	0.64	0.50
3	Two-Stage	-	Judge + Reviewer	0.75	0.60
4	Standard	-	5 Judges (MV)	0.74	0.46
5	Rationale	-	5 Judges (MV)	0.84	0.79
6	Rationale	Top-3	5 Judges (MV)	0.91	0.82
7	Rationale	Threshold	5 Judges (MV)	0.91	0.83
8	Two-Stage	-	Judge + 4 Reviewers (MV)	0.93	0.88
9	Standard	-	5 Judges (DS)	0.75	0.46
10	Rationale	-	5 Judges (DS)	0.84	0.80
11	Rationale	Top-3	5 Judges (DS)	0.91	0.82
12	Rationale	Threshold	5 Judges (DS)	0.91	0.84
13	Two-Stage	-	Judge + 4 Reviewers (DS)	0.91	0.85

Table 3: Ternary 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.

lecting different number of crowd judgments per document and others. Even the day and time when data is collected can greatly impact the set of annotators and corresponding results (Blanco et al., 2011; Arechar et al., 2017; Casey et al., 2017). While comparisons should be made with care, they still remain an important component of measuring progress.

Earlier we referenced Hosseini et al. (2012)'s work as perhaps the closest experimental design to our own. Their best "full" task design achieved 80-82% binary accuracy with MV or DS aggregation over three judges. Table 2 shows our Standard Task design with five judges achieves binary accuracy of 84% with MV aggregation and 86% with DS. Acknowledging differences in study designs, results are suggestive that our Standard Task design represents a strong baseline vs. prior work, with results on par with Hosseini et al.'s best full design. Moreover, it is notable that our Standard Task achieves this parity without any honey-pot questions or platform-specific worker filtering, which prior work has typically relied on as foundational design to ensure data quality.

Next, we compare results from the Rationale Task vs. the Standard Task. We observed notable improvement for both conditions of individual judging (Row 2 vs. 1) and aggregate consensus (Row 5 vs. 4, 9 vs. 10), as well as binary vs. ternary evaluation. With individual judging (Row 2 vs. 1), Rationale outperforms Standard for both binary judging (80% accuracy & $\kappa_C = 0.51$ vs. 65% accuracy & $\kappa_C = 0.36$) and ternary judging (64% accuracy & $\kappa_C = 0.5$ vs. 47% accuracy & $\kappa_C = 0.34$). For consensus, we see aggregated

judgment quality for Rationale beats Standard (Row 5 vs. 4) for both binary judging (92% accuracy & $\kappa_C = 0.8$ vs. 86% accuracy & $\kappa_C = 0.59$) and ternary judging (84% accuracy & $\kappa_C = 0.8$ vs. 75% accuracy & $\kappa_C = 0.46$), though we do see shrinking gains as we aggregate judgments and collapse to binary relevance.

A few other trends are interesting to note. First, we see improvements in accuracy even at the binary scale using the Rationale Task design. If we were to assume the gap between the Standard and Rationale Tasks was purely the result of nuance in the relevance grades (e.g., a worker judging a document *highly relevant* when the gold standard is merely *relevant*), we would expect the accuracy and κ gains to disappear as we collapse to binary, where κ is effectively no longer weighted and black-and-white decisions (e.g., relevant or not relevant) override nuance. However, this is not the case.

Another point to note is the effect of aggregation scheme in the Standard vs. Rationale Task. In both **Table 2 and 3**, the results in Rows 4-8 are aggregated using MV, while Rows 9-13 are aggregated using DS. We observed (Rows 5 and 10) that the more sophisticated DS aggregation technique offered no additional gains in either binary (92%) or ternary (84%) accuracy for the Rationale Task. In contrast, we saw gains in both binary (86% vs. 84%) and ternary (75% vs. 74%) accuracy for the Standard Task (Rows 4 and 9). These results reinforce the idea that the Rationale Task helps to filter out at least some low quality responses which might otherwise be collected by traditional means, as DS is thus unable to improve upon the filtering. Furthermore, the use of DS is ultimately unable to bridge the accuracy gap between the Standard and Rationale approaches.

In order to further insights into the results, we inspected whether the ratio of experienced workers are different in Standard and Rationale Tasks because experienced workers are more likely to provide high quality judgments than inexperienced ones. We first defined workers who completed 20 or more tasks as "experienced" workers and then partitioned workers into two groups: "experienced" and inexperienced (everyone else). We conducted this analysis on a subset of the original dataset (43%: 1481 judgments for Standard Task and 1487 judgments for Rationale Task) due to a data loss subsequent to publishing the original conference version of this work (McDonnell et al., 2016). Results are shown at **Table 4**. We observe comparable ratios of experienced vs. inexperienced workers for each task design. Both groups achieve roughly similar accuracy, with experienced workers being 1.5% more accurate on the Standard Task and 3.8% more accurate on the Rationale Task. Both groups achieve higher accuracy on the Rationale Task than the Standard Task. Overall, we do not see evidence to support a hypothesis that more accurate the Rationale Task performance can be explained away by greater worker experience.

Worker Group	**Standard Task**		**Rationale Task**	
	% Tasks completed	**Accuracy**	**% Tasks completed**	**Accuracy**
Inexperienced	25%	61.5%	25%	74.7%
Experienced	75%	63%	78%	78.5%

Table 4: Experienced vs. Inexperienced Workers.

6.4 Exploiting Rationales in Label Aggregation

Section 5 described a two-stage approach to use rationales and labels in aggregation: 1) identify and discard outlier rationales and their labels; and 2) aggregate remaining labels. Two string similarity functions were evaluated for comparing rationales (Section 5.1) and yielded similar results. We thus present results of Ratcliff and Metzener (1988) only.

Two heuristic algorithms were proposed for detecting outlier rationales: Top-N (Algorithm 1) and Threshold (Algorithm 2). Consensus results using Threshold filtering (Row 7) 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 suggests accurate assessors do select similar document extracts as rationales, correlating *overlap* in rationales and label accuracy.

Comparing algorithms, Threshold (Row 7) consistently outperforms Top-N (Row 6). Whereas Top-N always uses a fixed number of judgments, Threshold tunes the number of judgments kept per document according to the level of observed overlap. This suggests that adapting the number of filtered judgments for each document is important. We leave further exploration of this idea for future work.

Why might such pre-filtering improve upon DS weighted voting? DS models each worker as having a constant (i.e., global) expertise parameter for labeling each category. This is estimated over the entire dataset by maximum-likelihood. Rationale filtering in contrast is local to each example. It infers the quality of individual labeling decisions using an additional signal of agreement – the rationale – which is not directly encoded in the label and so not visible to DS. Intuitively, the local information captured by filtering might account for some "slip-ups" by workers on particular examples which their global DS parameters do not capture. Another potential intuition is that there exists a long tail of tasks completed by workers who do not perform many tasks, and so their sparsity of work likely means inaccurate estimates of their DS parameters.

Kutlu et al. (2018) also assessed these two algorithms but conversely find this filtering reduces accuracy. While this might be explained by various differences in their study design vs. ours, future work should further explore this to assess the generality and robustness of proposed algorithms. Overall, exploiting rationales and labels in tandem appears to have potential but mixed results thus far. Fortunately, the relatively simple methods proposed in this work leave great space and opportunity for further advancement.

6.5 Cost-Benefit Analysis of Rationales

While **Table 2** shows simple accuracy for the binary relevance of Standard vs. Rationale Tasks using either 1 judgment (individual judging) or 5 judgments (aggregate consensus), **Figure 2** 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 (DS results were similar), averaging over 20 random trials for each judgment count. Binary accuracy of Standard Task exhibits fairly consistent gains as judgments increase, achieving 86% accuracy with 5 judgments. In contrast, the Rationale Task approaches 90% accuracy with only three judgments, then shows rather modest gains thereafter.

As discussed earlier (Section 4.2), workers were paid the same amount for both tasks based on task completion times observed in our pilot study, in which *experienced* workers

(who completed 20 or more tasks) were remarkably seen to converge to nearly the same average completion time for both tasks (27 seconds for Standard and 29 for Rationale; Two-Stage's reviewer task took 26 seconds, not shown). Equal pay was intended to reflect these comparable task completion times and to simplify comparison of effects across tasks, but as noted in Section 4.2, it is possible equal pay introduced other confounds that future work could further explore. In total, we had average of 415 unique workers for each task type, with 8% of workers completing 20 or more tasks. These experienced workers accounted for approximately 50% of all the judgments in our study.

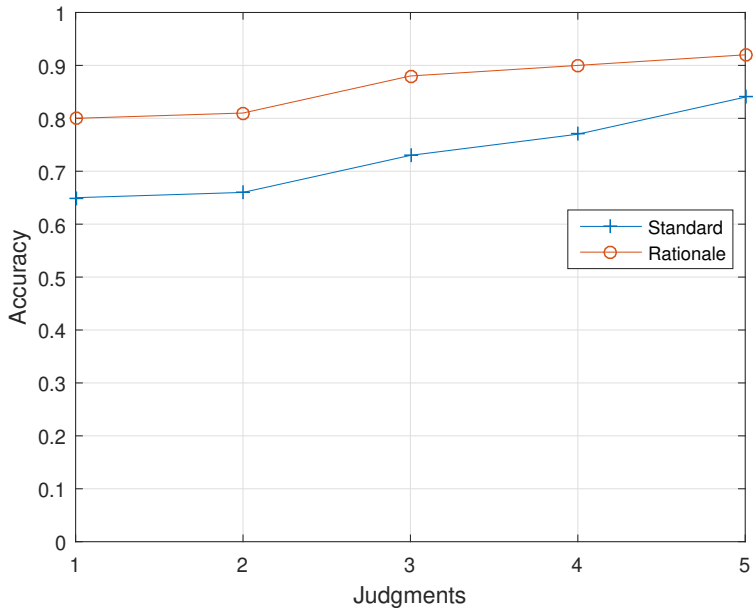

Figure 2: Judging accuracy vs. number of judgments, with MV for aggregation in the case of multiple judgments.

Why does average completion time decrease? We originally analyzed the average time *all* workers spent on each of their first 20 tasks (up to the number of tasks each completed). However, a confound of this analysis was that it was unclear whether completion time decreased with experience or if experienced workers were always faster, and increasing the number of tasks simply filtered out slower workers. Said another way, a worker might initially "triage" a task, to see if it can be completed quickly enough relative to payment in order to meet the worker's *reservation wage* (Horton & Chilton, 2010). If not, the worker would abandon the task, and average completion time might be explained by the departure of these slower workers, or workers with a higher reservation wage.

To remedy our analysis, we produced **Figure 3** by plotting the average time the subset of *experienced* workers (who we define as those completing 20 or more tasks) spent completing each of their first 20 tasks. The plot clearly shows a decrease in task time with more work, suggesting development of individual expertise and/or proceduralization efficiency

in completing the task. Our original analysis over all workers (not shown) was essentially identical[9].

Intuitively, both Standard and Rationale Tasks involve overhead for reading instructions and task familiarization. For Standard Task, we see task time rapidly fall off after this early phase, whereas Rationale Task time drops more slowly. We speculate that this is due to relevance judgment tasks being more common and familiar on MTurk. However, task time critically converges in both cases for *experienced* workers. We hypothesize that once familiarized with the task, both tasks effectively require the same cognitive effort: reviewing document text to make a relevance decision; the Rationale Task simply makes this explicit.

Zaidan et al. (2007)'s original rationale study found it typically required trusted annotators (an author and two students) twice as long to collect rationales in addition to labels. We posit this difference in completion times between our studies may stem from various factors: their rationale task requiring greater cognitive effort wrt. the number of rationales annotators were instructed to label, the marking of suggestive individual terms being less intuitive, and/or their annotators being more thoughtful in how they identified rationales based on provided instructions or interaction with the researchers.

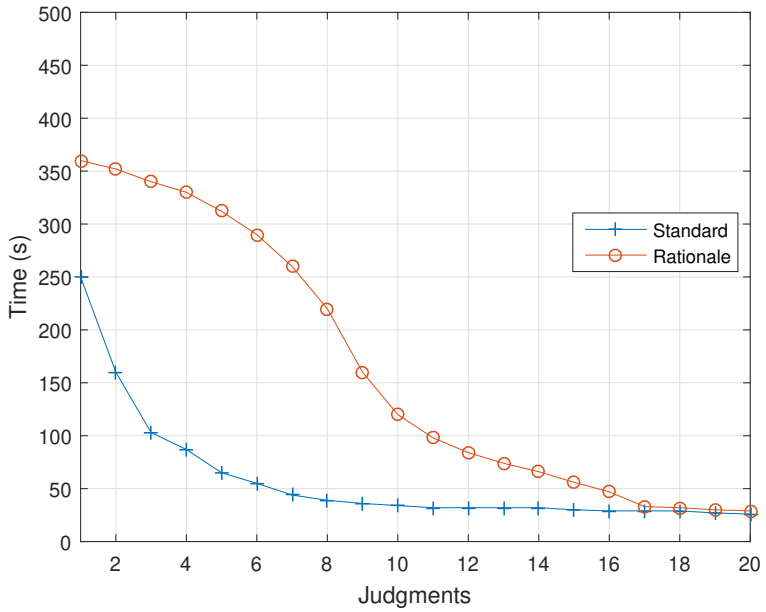

Figure 3: Average completion time vs. completed task count on Standard vs. Rationale Tasks for experienced workers.

6.6 Two-Stage Task Results

Our Two-Stage Task (Section 4.4) collects a judgment and rationale from a single assessor, then asks 4 subsequent *reviewers* to either confirm or modify the initial judgment. **Table 5**

9. Our coining and definition of 'experienced' workers (20 or more tasks) arose from this original analysis.

		Stage 2 Reviewer	
		Incorrect	Correct
Stage 1 Judge	Incorrect	4%	18%
	Correct	0%	78%

Table 5: Binary accuracy with respect to NIST contingency table of relevance judgments from Two-Stage Task. $\frac{18}{4+18} = 82\%$ of first stage errors are corrected without introduction of any new errors.

shows that the second-stage reviewer never introduced new judgment errors and fixed an error made by the initial judge 82% of the time. Further, recall the near-perfect binary agreement of second stage annotators in Two-Stage (Section 6.2). These results suggest that Two-Stage may provide high-quality data with only one judge and reviewer. **Table 2** also shows that Two-Stage with 2 judgments where non-relevant judgment is selected in cases of ties matches Standard's performance with 5 judges (Row 4) using 3 fewer judgments and with higher ternary κ_C: 0.60 vs. 0.46.

Next, we consider the case of consensus with 5 judgments. We aggregate judgments from 4 second-stage reviewers so that the 1 judge + 4 reviewers = 5 judgments matches the 5 judgment count (and cost) of Rationale with consensus shown in **Table 2**. Two-Stage (Row 8) is seen to match the accuracy and κ_C of Rationale with Threshold filtering (Row 7) while incurring the exact same cost.

7. Qualitative Analysis of Rationales

We now explore in detail the level of transparency provided by rationales. We begin by an inspection of common ways in which they were used in the decision-making processes of our reviewers in the Two-Stage Task. We then discuss how such insight into the annotator decision process provided by rationales can be used to confidently establish alternative truths in the case of disagreement between the aggregate crowd and the NIST gold standard. Finally, we discuss the case and potential shortcomings of *negative rationales*, or instances in which an annotator must provide a rationale in defense of a non-relevant label.

7.1 Two-Stage Task: Rationales as Insight

Two-Stage Task reviewers used rationales in several common ways. **Table 6** presents a search topic and extended narrative on pet adoption for which we collected relevance judgments in our pilot studies. We discuss a subset of judgments collected for various documents falling under this search narrative. For each example, the table contains a judgment and rationale from an initial annotator and a subsequent reviewer collected using the Two-Stage Task design, respectively. The gold standard is taken from one of the author's blind judgments on the same 4-point scale used by the workers. Recall that the first-stage annotator is always referred to as Tom in the Two-Stage Task (Section 4.4).

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. A site providing general information on dog adoption is also relevant.

Table 6: Sample search topic and narrative.

7.1.1 Rationale Affirmation

	Document 1
Worker 1 Judgment (Tom)	Probably Not Relevant
Worker 1 Rationale (Tom)	*Rooterville Sanctuary. For adoption: pets, pig, pigs, piggy, piggies, pork.*
Worker 2 Judgment	Probably Not Relevant
Worker 2 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, although they mention dogs
Gold Standard	Probably Not Relevant

Table 7: Rationale Affirmation: Two-Stage Task with worker responses.

Recall from Table 5 that the Stage Two action in 78% of cases did not result in the binary change of a relevance judgment; in these situations, the reviewer simply provided an affirmation of the relevance decision offered by the initial annotator. Often, the reviewer would *directly* reference the initial annotator's rationale in their reasoning why they support the original judgment. **Table 7** shows one such case of a rationale affirmation. The initial judge rated the document to be *Probably Not Relevant* and cited a rationale which suggests that though the website is indeed for a pet adoption sanctuary, it appears to only specialize in pigs. The reviewer affirmed this judgment and explicitly cited agreement with Tom's (i.e.,

the initial annotator's) rationale that the sanctuary was primarily focused on pigs, not dogs.

7.1.2 Rationale Error Correction

The second most common action (18% of cases) was a modification of the original relevance decision. We noticed that these modifications fall into three broad categories: (1) minor tweaking of the original answer; (2) a simple correction of an apparently blatant error in the original label; and (3) a refutation of a non-relevant label using an illustrative rationale. We discuss each of these cases in turn.

Tweaking. The most common form of error correction that we noticed in our experimental results was *tweaking*, in which an annotator modifies the original relevance judgment by only moving it up or down one point on the four-point relevance scale (e.g., *Definitely Relevant* → *Probably Relevant*). **Table 8** shows an example of tweaking. The initial annotator labeled the document at hand *Definitely Relevant*, and their rationale indicates that this is 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 for an organization in the US."

	Document 2
Worker 1 Judgment (Tom)	Definitely Relevant
Worker 1 Rationale (Tom)	*View our rescue dogs - visit our organization or contact us directly to see what is available.*
Worker 2 Judgment	Probably Relevant
Worker 2 Reasoning	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.
Gold Standard	Probably Relevant

Table 8: Tweaking: Two-Stage Task with worker responses.

Interestingly, although there is no preferred locale specified in either the Topic or Narrative (Table 5), our reviewer may have derived such conclusions from American spelling or vocabulary in the Narrative. More generally, their reasoning brings up a compelling point: namely, location is likely an important component of the information need and may have been overlooked when the topic was formulated. Such transparency of thought is invalu-

able since there is nothing explicit in the Narrative supporting the reviewer's supposition, though American vocabulary or spelling may be the culprit.

Blatant Error Correction. 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 (See **Table 9**).

	Document 3
Worker 1 Judgment (Tom)	Definitely Not Relevant
Worker 1 Rationale (Tom)	*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.*
Worker 2 Judgment	Definitely Relevant
Worker 2 Reasoning	Tom provided a lot of information that shows why this page should be useful for Alice.
Gold Standard	Definitely Relevant

Table 9: Blatant Error Correction: Two-Stage Task with worker responses.

Rationale Refutations. In this case, a judge provides a rationale in support of a relevant decision.

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. In all cases, the reviewer was empowered to take a different, confidently informed action: affirming, tweaking, or correcting the original judgment, respectively. Reviewer justifications further suggest that the Two-Stage Task design requires a more involved critical thinking process in which reviewers understand that their duty is not only to form a strong justification for their subjective judgment, but also grounding their decision-making process in tandem with reasoning about the original judge's opinion.

Summary. In any of these cases, we see the unique insight that rationales provide into the decision-making process and the responsibility that the Two-Stage Task demands of workers, who must weigh their relevance decisions critically against the logic of previous human annotators and/or be subject to additional oversight from the task creator. Our experimental results indicate that this results in higher-quality decision-making in the second stage of the process.

7.2 Negative Rationales

When Zaidan et al. (2007) proposed rationales, they considered a binary labeling task (movie review sentiment) in which annotators would mark indicative terms for either positive or negative reviews. We similarly collected rationales for both relevant and non-relevant documents, but our classification task had a subtle asymmetry that theirs did not: a relevant document excerpt is sufficient to judge a document as relevant, but an excerpt of non-relevant content is not proof that the rest of the document is non-relevant. The non-relevant content may simply be located elsewhere in the document.

Thinking beyond the relevance judgment task we focus on in this study, Section 2.2 discussed generality of rationales as a mechanism for filtering to narrow scope in search problems (e.g., filtering satellite imagery to find potential locations for Jim Gray's missing sailboat). For any such general search problem, this logic seems applicable: identifying a subset of the search space that does not satisfy search criteria does not mean that search criteria cannot be satisfied elsewhere in the search space. In this work, we distinguish *negative rationales* as rationales provided in support of a *non-relevant* document label.

We explored alternative designs which respect this difference, such as one not requiring rationales when labeling a document non-relevant. However, this might potentially bias workers towards choosing non-relevant, as the decision would require less total effort (Kittur et al., 2008). Furthermore, we would lose any insight into the worker thought process, a useful by-product of our rationale-based task design. Our ultimate decision to require negative rationales represents a design choice that future work might further explore.

Table 10 presents a selection of sample negative rationales from our experiments which we include as representative of their frequent forms and interpretability. Row (1) illustrates the most common form of the negative rationale seen throughout our experiments, in which the crowd worker selects an excerpt which summarizes the main content of the page, which is presumably at odds with the given topic. In this case, the web page was a home for the Adobe Flash Player download, which is clearly not related to any Minnesota Public Radio station. While we cannot be sure that other sections of the page do not contain information relevant to the topic, the rationale does partially justify the worker's label.

In Row (2), we see the worker quote a programming error that was visible on the web page. Unfortunately, it is difficult to draw complete conclusions from this rationale. The source code on the page may indeed be broken, making it irrelevant to the topic. Alternatively, perhaps only a particular subset of non-vital material failed to load appropriately on the page, which precluded the worker from making an informed decision, or it is the result of an error in our own framework. In this case, it was an error in the source code of the page, but confirming this required additional manual analysis of the page. Nonetheless, the rationale proved to be a useful tool in reaching this understanding.

The rationale shown in Row (3) appears to be Latin gibberish. Obviously, the excerpt does not appear to hold any relevant information. In practice, however, these Latin excerpts are used frequently in web development to fill in space when showcasing web design templates. Indeed, this particular web page was an unfinished page acting as placeholder for future content. Here we see again how negative rationales may be less trivially interpretable, but still provide a useful avenue towards understanding worker behavior.

Row	Query	Rationale
(1)	I'm looking for the homepage of The Current, a program on Minnesota Public Radio.	The Adobe Flash Player runtime lets you effortlessly reach over 1.3 billion people across browsers and OS versions with no install - 11 times more people than the best-selling hardware console.
(2)	Find songs, lyrics, music, and information about the musical The Music Man.	Fatal error: Route Error: The class 'FrkysController' doesn't appear to be valid.
(3)	What is the effect of excessive heat on dogs?	Illum secundum exerci erat plaga illum,enim, venio. Tamen causa ut diam torqueo gaciter inhibeo si quae exerci lobortis.
(4)	Find songs, lyrics, music, and information about the musical The Music Man.	The text did not help me with my decision.
(5)	Find information on diversity, both culturally and in the workplace.	N/A

Table 10: Sample of Negative Rationales.

Finally, Rows (4-5) are rationales which both were observed repeatedly throughout experiments and emphasize some of the ambiguity present with negative rationales. In (4), the worker's selected rationale was "*The text did not help me with my decision*". Recall from Section 4 that this was a special rationale that we allowed workers to use when there was no textual information on the page which supported their decision. Though we intended for this special rationale to only be used in cases where only visual information was relevant to the topic, in retrospect it is understandable why some users would resort to it in cases of negative rationales: how would they prove that the document is non-relevant without copying and pasting the entire content of the page? In (5), we see an even more direct form of this critique, in which a worker felt that rationales simply made no sense for non-relevant document labels and used "*N/A*" as their rationale in such cases.

8. Worker Feedback

As outlined in Section 4, each iteration of our task design included an open-ended feedback form through which workers could communicate questions or offer suggestions. In fact, this open-ended worker feedback prompted our transition away from *exhaustive* special-case instructions towards a *simple*, *streamlined*, and *colloquial* set of instructions and relevance options, as workers in our pilot studies were simultaneously always able to uncover new corner cases we had not previously considered and became frustrated by the length and complexity of instructions. While these suggestions directly influenced our final design, we

also received a wide variety of other useful feedback which did not directly affect our design, but may inspire future work or further improvements.

8.1 Rationale Expressiveness

Following Zaidan et al. (2007), we adopted excerpt-style rationales, which proved to be both powerful and easy to process. Nevertheless, some workers expressed a desire for free-form rationales, suggesting it would allow for more powerful expressiveness and justification, as other prior work has allowed (Drapeau et al., 2016). As one worker commented:

> Please add an additional box where the Turker can input comments, rather than copy and pasting from the reference website.

Recall from Section 4 that we purposefully disallowed free-form rationales: by constraining rationales to exact in-document excerpts, we were able to exploit rationale similarity comparisons methods that operate on simple string overlap. Future task designs might simply incorporate a free-form rationale box *in addition to* the excerpt selection to allow for more flexible justifications from workers and to provide them with peace of mind, even if the free-form explanation was never utilized. We have discussed earlier how excerpt-style rationales provide a benefit, even if ignored after collection.

Additionally, one area of future work we previously mentioned revolves around the exploration of more sophisticated string similarity metrics which respects notions of *semantic similarity*, rather than simply string overlap (See Section 5). In addition to better modeling similarity in excerpt-based rationales, such semantic similarity metrics might also be able to derive utility from free-form rationales, a win-win for both task designers as well as annotators who desire more flexible expression of justifying labeling decisions.

Zaidan et al. (2007) motivated rationales as useful for gaining traction on the *credit-assignment* problem: helping a training model to better understand which subset features best explained the annotator label assigned. If free-form rationales were allowed, it is an open question how well they would support such use for dual-supervision. It would seem that there may be a "chicken-and-egg" situation here: if we understood free-form language well-enough to solve the credit-assignment problem, then our NLP models might already be strong enough that such dual supervision was no longer beneficial.

8.2 Graphical Rationales

As mentioned earlier, textual excerpts cannot convey *visual information* in a document that may be central to a worker's justification. In fact, we did realize this limitation during our pilot studies and incorporated this knowledge into our task design: recall that our final design allowed workers to explicitly use the following string as their rationale in cases where visual, rather than textual information, prompted their decision: *The text did not help with my decision.* This simplifying design decision allowed us to maintain a notion of rationale string overlap even in cases of visual information.

Nevertheless, we had several workers comment on this limitation of rationales and offer suggestions for such cases. For example, consider the following piece of feedback:

> If the intent of the searcher is to find graphical information - i.e., 'Find maps of the United States' - there should be additional options for justifying the response, such as enclosing link URLs.

The crowd worker makes a strong case and offers a simple solution for the problem, albeit it one that clashes with the simplest notion of rationale similarity used in this study. Alternatively, future work might consider other methods of incorporating visual information into the rationale, such as allowing workers to directly draw bounding boxes around relevant visual information on a page. Rationale overlap could then be computed as the area of overlap between bounding boxes drawn by different workers. Such an approach would afford workers greater expressiveness while preserving the notion of rationale similarity exploited here.

9. Limitations and Future Work

We believe annotator rationales (Zaidan et al., 2007) offer a wide range of potential benefits, beyond their original motivation to support faster machine learning relative to annotator effort. While we have begun to explore this potential, we recognize a variety of limitations of our work and imagine a number of ideas for future work.

Generality of findings. As articulated in Section 2.2, we perceive broad applicability for rationales. Nevertheless, our study only considered one annotation task: text-based relevance assessment. Moreover, a follow-on study (Kutlu et al., 2018) found various differences in empirical results which require further analysis. While prior work has considered text and image classification tasks (Zaidan et al., 2007; Donahue & Grauman, 2011), this is still a small set of tasks, and those works did not investigate rationales in the context of improving annotation processes or quality. Thus a clear direction for future work is to further investigate and assess the generality of findings reported in this study.

Sequential task design. In our Two-Stage Task, we referred to the first assessor as "Tom" in order to make our task more personable. However, some secondary reviewers might have interpreted this to mean that Tom is a real person. The name "Tom" may have introduced unanticipated impacts regarding culture or gender assumptions. For instance, the secondary reviewer may infer that "Tom" is a male from an English-speaking country, suggesting cultural expertise or bias. In addition, a secondary reviewer completing many tasks may form a mental profile of Tom's judgments and their average reliability. Such inferences, implicit or explicit, could influence the secondary reviewer's beliefs about the reliability of Tom's judgments. Alternative designs might be considered. More broadly, because rationales enable multi-stage, iterative refinement of labels through creative workflow design, a wide range of multi-stage workflows could be investigated (Lin et al., 2012).

What do annotators think? We hypothesized that requiring rationales could reduce temptation to cut corners in judging relevance, but we could only evaluate our hypothesis via submitted responses. User study techniques, such as surveys, observation, talk-aloud, and retrospection could yield more in-depth findings to better understand how the rationales design influenced annotator conceptualization and approach to the annotation task.

Better metrics for rationale similarity. In this work, we used simple character overlap between rationales to measure textual similarity. In the image domain (Donahue &

Grauman, 2011), one could similarly measure pixel overlap in bounding boxes. Intuitively, better measures of rationale similarity should lead to better validation and aggregation. To this end, future work on text might investigate embedding or transformer models, such as word2vec (Mikolov et al., 2013) and BERT (Devlin et al., 2019). When working with imagery instead of text, similar "visual word embeddings" might be explored instead.

Suggesting rationales to annotators. While our design required human annotators to find suitable rationales completely on their own (manually), future work might automatically suggest one or more potential rationales for annotators to select between and/or revise. For our relevance judging task, passage-retrieval (aka focused-retrieval) IR (Trotman et al., 2007) or extractive summarization approaches might suggest potentially relevant passages to assessors. Ideally, such hybrid labeling would reduce human effort, but poor automatic rationales could be worse than annotators starting from scratch (Ramírez et al., 2019), and even good rationales may still bias human workers toward lower quality rationales than they might have picked if working independently.

What rationales do annotators select, and why? When annotators have latitude in selecting a small rationale from a larger example to provide for a given label, how do they choose? With our task, a document may have a variety of relevant content to choose from, as well as freedom in task instructions regarding rationale length. While our task design likely biased assessors to select rationales near the start of the document, did they actually do so? Is there any correlation between judgment accuracy and the rationale position in the document? Such analysis was further complicated in our study of web pages due to their complex page structure; the order of text written in HTML may differ from the order in which the web page is designed to be rendered.

Measuring human agreement on rationales. Our heuristic filtering methods (Section 5) employed rationale similarity statistics, but we have not performed detailed study of rationale inter-annotator agreement statistics. As noted earlier, whereas simple labels permit standard annotator agreement measures, more open-ended rationales require an appropriate metric for measuring similarity, complicating analysis. The broader significance here is that there is great interest today in eXplainable AI (XAI) systems that generate rationales to explain model predictions. For example, recent work on neural models has continued to investigate use human rationales to supervise models (Zhang et al., 2016; Bao et al., 2018) and evaluate unsupervised models (Lei et al., 2016; Bastings et al., 2019). Unfortunately, human rationales are largely being used to train and evaluate model without any inter-annotator statistics. A challenge in studying this issue systematically is that the form of rationale being used in different studies varies greatly. We have also discussed how annotation task design can also bias annotators in which rationales they select.

Better label aggregation via rationales. Zaidan et al. (2007) motivated rationales as enabling dual-supervision over rationales and labels. While there has been plentiful work on label aggregation (Sheshadri & Lease, 2013; Zheng et al., 2017), we know of no prior work proposing dual-supervision for aggregation via rationales. In this paper, we presented two heuristic methods, filtering judgments based on rationale overlap prior to aggregation (Section 5). Future work could usefully investigate principled approaches to jointly exploit label agreement with rationale similarity in order to better aggregate annotator responses. As above, a metric for measuring rationale similarity is required.

Creating gold labels. With objective tasks, we assume each question has a single correct answer, often referred to as the "gold" label and generated by trusted and/or expert personnel. Because such personnel are typically busy and/or expensive, it would be valuable if we could effectively engage the wider crowd in gold label generation. Oleson et al. (2011) propose generating gold standard judgments using an initial set of manually created judgments, based on detecting common worker errors and performing spot checks. The additional quality and verifibiability of rationales suggest their use has potential here.

Improving gold labels. Another interesting direction is use of the crowd to find errors in existing gold labels (Chang et al., 2015), to try to further improve the gold standard (Wilbur & Kim, 2011). While gold labels and annotators are typically assumed to be infallible, in practice gold datasets often hide underlying impurities, especially when created without multiple annotators or quantified measures of inter-annotator agreement. With rationales, when a worker disagrees with an accepted gold label, the rationale could help establish the validity of an alternative answer or reveal error in the gold standard.

Going beyond Requester expertise. While many crowdsourcing studies assume interchangeable workers, an original motivation for crowdsourcing was to discover and utilize appropriate talent for a given task (beyond expertise within one's own organization) (Howe, 2006). For example, if we want to identify the (subjective) best sushi restaurant in Amarillo, TX, we might seek someone in the crowd having local expertise (Paiement et al., 2010). If we imagine this restaurant problem in a foreign country with reviews written in a foreign language, could rationales ease a Requester's job in verifying reasonable crowd responses describing an unfamiliar foreign region and language (Chen & Dolan, 2011)?

Supporting subjective tasks. With objective tasks, in which we assume each question has a single correct answer, quality control often involves checking crowd responses vs. experts (supervised, gold agreement) (Snow et al., 2008) or peers (unsupervised, peer agreement) (Dawid & Skene, 1979). Subjective tasks, on the other hand, assume a diversity of valid responses. For example, one might wish to poll the crowd to determine the distribution of popular opinion while ensuring data quality (e.g., not random clicking). With such subjective tasks, rationales may be useful for both ensuring data quality and providing further insight into responses. For example, if we want to identify the (subjectively) best restaurant(s) to visit, movie(s) to watch, or product(s) to buy, based upon online reviews, a rationale-based task might require not only selecting the restaurant, but also citing specific evidence from the review(s) to support the answer. Future work might probe respondents for likert-scale agreement with ethical, political, or medical arguments, citing key text they identify as the rationale for their responses. As discussed in Section 2.2, we can also generalize beyond text, e.g., polling likert agreement with a position statement based on audio or video evidence, asking respondents to indicate a short segment of the audio or video that most influenced their response.

10. 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 contrast with most prior work, we invite anyone interested to work on our tasks (we perform no worker filtering), and we require no labeled data to test workers (i.e., on questions with known answers). Despite this, our Standard Task serves as a strong baseline, achieving 82% binary accuracy. Our Rationale Task design further improves data quality while entirely ignoring the collected rationales. With only two workers, our sequential, Two-Stage Task design achieves 85% binary accuracy. Aggregating judgments from 5 workers provides further improvement, and by exploiting degree of overlap in judges' rationales, we can achieve 96% binary accuracy.

In addition to improved label quality, rationales provide further benefits. Rationales enable future users of a dataset to verify the label quality. Thus, any problem in datasets can be fixed and inaccurate evaluation of systems can be prevented. Furthermore, the rationales themselves can be used as labels for various tasks such as passage retrieval. Moreover, rationales enable a broader range of simple labeling tasks to benefit from iterative task design by providing a communication channel between crowdworkers.

Acknowledgments

We thank the many talented crowd workers whose annotations enabled our research. We also thank our reviewers, at both HCOMP and JAIR, for their helpful feedback and suggestions. This work was made possible by generous support from NPRP grant# NPRP 7-1313-1-245 from the Qatar National Research Fund (a member of Qatar Foundation), the National Science Foundation (grant No. 1253413), the Micron Foundation, and UT Austin's Good Systems Grand Challenge Initiative to design a future of responsible AI (`http://goodsystems.utexas.edu`). Any opinions, findings, and conclusions or recommendations expressed by the authors are entirely their own and do not represent those of the sponsoring agencies.

References

Aiman L Al-Harbi and Mark D Smucker. A qualitative exploration of secondary assessor relevance judging behavior. In *Proceedings of the 5th Information Interaction in Context Symposium*, pages 195–204. ACM, 2014.

Omar Alonso. Guidelines for designing crowdsourcing-based relevance experiments, 2009. CiteSeerX DOI 10.1.1.149.6649.

Omar Alonso. Practical lessons for gathering quality labels at scale. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1089–1092. ACM, 2015.

Omar Alonso and Ricardo Baeza-Yates. An analysis of crowdsourcing relevance assessments in spanish. In *Spanish Conference on Information Retrieval*, 2010.

Omar Alonso and Stefano Mizzaro. Can we get rid of trec assessors? using mechanical turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, volume 15, page 16, 2009.

Omar Alonso and Stefano Mizzaro. Using crowdsourcing for trec relevance assessment. *Information Processing & Management*, 48(6):1053–1066, 2012.

Omar Alonso, Daniel E Rose, and Benjamin Stewart. Crowdsourcing for relevance evaluation. In *ACM SigIR Forum*, volume 42, pages 9–15, 2008.

Antonio A Arechar, Gordon T Kraft-Todd, and David G Rand. Turking overtime: how participant characteristics and behavior vary over time and day on amazon mechanical turk. *Journal of the Economic Science Association*, 3(1):1–11, 2017.

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

Josh M Attenberg, Pagagiotis G Ipeirotis, and Foster Provost. Beat the machine: Challenging workers to find the unknown unknowns. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

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 *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 667–674. ACM, 2008.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. Deriving machine attention from human rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1903–1913, 2018.

Jeff Barr and Luis Felipe Cabrera. Ai gets a brain. *Queue*, 4(4):24, 2006.

Joost Bastings, Wilker Aziz, and Ivan Titov. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1284. URL `https://www.aclweb.org/anthology/P19-1284`.

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.

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.

Martin Braschler and Carol Peters. The clef campaigns: Evaluation of cross-language information retrieval systems. *UPGRADE (The European Online Magazine for the IT Professional)*, 3:78–81, 2002.

Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. The ClueWeb09 Data Set, 2009. Presentation Nov. 19, 2009 at NIST TREC. Slides online at `boston.lti.cs.cmu.edu/classes/11-742/S10-TREC/TREC-Nov19-09.pdf`.

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

Ben Carterette, Virgiliu Pavlu, Hui Fang, and Evangelos Kanoulas. Million query track 2009 overview. In *Proceedings of NIST TREC*, 2010.

Logan S Casey, Jesse Chandler, Adam Seth Levine, Andrew Proctor, and Dara Z Strolovitch. Intertemporal differences among mturk workers: Time-based sample variations and implications for online data collection. *SAGE Open*, 7(2):2158244017712774, 2017.

Praveen Chandar, William Webber, and Ben Carterette. Document features predicting assessor disagreement. In *Proceedings of the 36th ACM SIGIR conference on Research and development in information retrieval*, pages 745–748. ACM, 2013.

Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2334–2346. ACM, 2017.

Nancy Chang, Praveen Paritosh, David Huynh, and Collin F Baker. Scaling semantic frame annotation. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, page 1, 2015.

David L Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics, 2011.

Quanze Chen, Jonathan Bragg, Lydia B Chilton, and Dan S Weld. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

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

Cyril W Cleverdon and Michael Keen. Aslib cranfield research project-factors determining the performance of indexing systems; volume 2, test results. 1966.

Paul Clough, Mark Sanderson, Jiayu Tang, Tim Gollins, and Amy Warner. Examining the limits of crowdsourcing for relevance assessment. *IEEE Internet Computing*, 17(4):32–38, 2013.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics*, 28(1):20–28, 1979.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh acm international conference on web search and data mining*, pages 135–143. ACM, 2018.

Jeff Donahue and Kristen Grauman. Annotator rationales for visual recognition. In *ICCV*, pages 1395–1402. IEEE, 2011.

Ryan Drapeau, Lydia B Chilton, Jonathan Bragg, and Daniel S Weld. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 5–14. ACM, 2017.

Snehalkumar Neil S Gaikwad, Durim Morina, Adam Ginzberg, Catherine Mullings, Shirish Goyal, Dilrukshi Gamage, Christopher Diemert, Mathias Burton, Sharon Zhou, Mark Whiting, et al. Boomerang: Rebounding the consequences of reputation feedback on crowdsourcing platforms. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 625–637. ACM, 2016.

Google. Search quality rating guidelines. *Inside Search: How Search Works*, March 28 2016. `www.google.com/insidesearch/`.

Mary L Gray and Siddharth Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Eamon Dolan Books, 2019.

Maura R Grossman and Gordon V Cormak. Inconsistent responsiveness determination in document review: Difference of opinion or human error. *Pace L. Rev.*, 32:267, 2012.

Lei Han, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. On transforming relevance scales. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 39–48, 2019.

Lei Han, Eddy Maddalena, Alessandro Checco, Cristina Sarasua, Ujwal Gadiraju, Kevin Roitero, and Gianluca Demartini. Crowd worker strategies in relevance judgment tasks. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 241–249, 2020.

Maram Hasanain, Yassmine Barkallah, Rees Suwaileh, Mucahid Kutlu, and Tamer Elsayed. Artest: The first test collection for arabic web search with relevance rationales. In *The 43rd International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 2020.

Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, pages 419–429. International World Wide Web Conferences Steering Committee, 2015.

John Joseph Horton and Lydia B Chilton. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 209–218. ACM, 2010.

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.

Jeff Howe. The rise of crowdsourcing. *Wired Magazine*, 14(6):176–183, 2006.

Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. An evaluation of aggregation techniques in crowdsourcing. In *International Conference on Web Information Systems Engineering*, pages 1–15. Springer, 2013.

Panos Ipeirotis. A plea to amazon: Fix mechanical turk! *Blog: Behind Enemy Lines*, Oct. 21 2010. October 21, 2010. `www.behind-the-enemy-lines.com`.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 467–474. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

Mike Kappel. The end of an era? how the abc test could affect your use of independent contractors. *Forbes*, 2018. August 8.

Gabriella Kazai, Nick Craswell, Emine Yilmaz, and Seyed MM Tahaghoghi. An analysis of systematic judging errors in information retrieval. In *Proceedings of the 21st ACM conference on Information and knowledge management*, pages 105–114, 2012.

Steve Kelling, Jeff Gerbracht, Daniel Fink, Carl Lagoze, Weng-Keen Wong, Jun Yu, Theodoros Damoulas, and Carla Gomes. A human/computer learning network to improve biodiversity conservation and research. *AI magazine*, 34(1):10–10, 2013.

Kenneth A. Kinney, Scott B. Huffman, and Juting Zhai. How evaluator domain expertise affects search result relevance judgments. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 591–598, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458160. URL `http://doi.acm.org/10.1145/1458082.1458160`.

Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM, 2008.

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.

Anand Kulkarni, Philipp Gutheim, Prayag Narula, David Rolnitzky, Tapan Parikh, and Björn Hartmann. Mobileworks: Designing for quality in a managed crowdsourcing architecture. *IEEE Internet Computing*, 16(5):28–35, 2012.

Mucahid Kutlu, Tyler McDonnell, Yassmine Barkallah, Tamer Elsayed, and Matthew Lease. Crowd vs. Expert: What Can Relevance Judgment Rationales Teach Us About Assessor Disagreement? In *Proceedings of the 41st ACM SIGIR conference on Research and development in Information Retrieval*, 2018.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, 2016.

Christopher H Lin, Mausam Mausam, and Daniel S Weld. Dynamically switching between synergistic workflows for crowdsourcing. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

Chris J Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M Jordan Raddick, Robert C Nichol, Alex Szalay, Dan Andreescu, et al. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189, 2008.

Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. Turkit: Human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 57–66. ACM, 2010.

C. V. K. Manam and Alexander James Quinn. WingIt: Efficient Refinement of Unclear Task Instructions. In *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2018.

Akash Mankar, Riddhi J. Shah, and Matthew Lease. Design Activism for Minimum Wage Crowd Work. In *5th AAAI Conference on Human Computation and Crowdsourcing (HCOMP): Works-in-Progress Track*, 2017.

Catherine C Marshall and Frank M Shipman. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *5th Annual Web Science Conference*, pages 234–243. ACM, 2013.

Winter Mason and Siddharth Suri. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

Winter Mason and Duncan J Watts. Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 77–85. ACM, 2009.

Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. Why is that relevant? collecting annotator rationales for relevance judgments. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

An Thanh Nguyen, Matthew Halpern, Byron C. Wallace, and Matthew Lease. Probabilistic Modeling for Crowdsourcing Partially-Subjective Ratings. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, pages 149–158, 2016.

David Oleson, Alexander Sorokin, Greg P Laughlin, Vaughn Hester, John Le, and Lukas Biewald. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. *Human computation*, 11(11), 2011.

Jean-Francois Paiement, James G Shanahan, and Remi Zajac. Crowdsourcing local search relevance. *Proceedings of the CrowdConf 2010*, 2010.

Praveen Paritosh. Human computation must be reproducible. In *CrowdSearch*, pages 20–25, 2012.

Jorge Ramírez, Marcos Baez, Fabio Casati, and Boualem Benatallah. Understanding the impact of text highlighting in crowdsourcing tasks. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 144–152, 2019.

John W Ratcliff and David E Metzener. Pattern-matching-the gestalt approach. *Dr Dobbs Journal*, 13(7):46, 1988.

Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. On fine-grained relevance scales. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 675–684. ACM, 2018.

Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*, pages 2863–2872. ACM, 2010.

Holly Rosser and Andrea Wiggins. Crowds and camera traps: Genres in online citizen science projects. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

Niloufar Salehi, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, et al. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1621–1630. ACM, 2015.

Mark Sanderson. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc, 2010.

Tefko Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, 58(13):2126–2144, 2007.

Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–19, 2018.

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.

M Six Silberman, Bill Tomlinson, Rochelle LaPlante, Joel Ross, Lilly Irani, and Andrew Zaldivar. Responsible research with crowds: pay crowdworkers at least minimum wage. *Communications of the ACM*, 61(3):39–41, 2018.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

Eero Sormunen. Liberal relevance criteria of trec-: Counting on negligible documents? In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 324–330. ACM, 2002.

Qi Su, Dmitry Pavlov, Jyh-Herng Chow, and Wendell C. Baker. Internet-scale collection of human-reviewed data. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 231–240, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. doi: 10.1145/1242572.1242604. URL `http://doi.acm.org/10.1145/1242572.1242604`.

John C Tang, Manuel Cebrian, Nicklaus A Giacobe, Hyun-Woo Kim, Taemie Kim, and Douglas Beaker Wickert. Reflecting on the darpa red balloon challenge. *Communications of the ACM*, 54(4):78–85, 2011.

Rong Tang, William M Shaw Jr, and Jack L Vevea. Towards the identification of the optimal number of relevance categories. *Journal of the Association for Information Science and Technology*, 50(3):254, 1999.

Yuandong Tian and Jun Zhu. Learning from crowds in the presence of schools of thought. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 226–234. ACM, 2012.

Andrew Trotman, Nils Pharo, and Dylan Jenkinson. Can we at least agree on something. In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, pages 49–56, 2007.

Donna Vakharia and Matthew Lease. Beyond mechanical turk: An analysis of paid crowd work platforms. *Proceedings of the iConference*, 2015.

Werner Vogels. *Help Find Jim Gray*, 2007. `https://www.allthingsdistributed.com/2007/02/help_find_jim_gray.html`.

Ellen M Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management*, 36(5):697–716, 2000.

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

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

Simon Wakeling, Martin Halvey, Robert Villa, and Laura Hasler. A comparison of primary and secondary relevance judgements for real-life topics. In *Proc. of the ACM on Conf. on Human Information Interaction and Retrieval*, pages 173–182, 2016.

Bing Wang, Bonan Hou, Yiping Yao, and Laibin Yan. Human flesh search model incorporating network expansion and gossip with feedback. In *2009 13th IEEE/ACM International Symposium on Distributed Simulation and Real Time Applications*, pages 82–88. IEEE, 2009.

Nai-Ching Wang, David Hicks, Paul Quigley, and Kurt Luther. Read-agree-predict: A crowdsourced approach to discovering relevant primary sources for historians. *Human Computation*, 6(1):147–175, 2019.

Peng Dai Mausam Daniel S Weld. Decision-theoretic control of crowd-sourced workflows. In *Proceddings of the Twenty-Fourth Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pages 1168–1174, 2010.

Mark E Whiting, Grant Hugh, and Michael S Bernstein. Fair work: Crowd work minimum wage with one line of code. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 197–206, 2019.

W John Wilbur and Won Kim. Improving a gold standard: treating human relevance judgments of medline document pairs. *BMC bioinformatics*, 12(3):S5, 2011.

Stephen Wolfson and Matthew Lease. Look before you leap: Legal pitfalls of crowdsourcing. In *Proceedings of the 74th Annual Meeting of the American Society for Information Science and Technology (ASIS&T)*, 2011.

Meng-Han Wu and Alexander James Quinn. Confusing the crowd: Task instruction quality on amazon mechanical turk. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, pages 206–215, 2017.

Omar F Zaidan and Jason Eisner. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 31–40. Association for Computational Linguistics, 2008.

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.

Ye Zhang, Iain Marshall, and Byron C Wallace. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 795–804, 2016.

Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5): 541–552, 2017.