Crowdsourcing Breakthroughs for Language Technology Applications
AAAI Technical Report WS-15-24

Rigorously Collecting Commonsense Judgments for Complex Question-Answer Content

Mehrnoosh Sameki*
Boston University
Computer Science Department
Boston, MA 02215
sameki@bu.edu

Aditya Barua
Google Inc.
340 Main St.
Los Angeles, CA 90291
adityabarua@google.com

Praveen Paritosh
Google Inc.
345 Spear St.
San Francisco, CA 94105
pkp@google.com

Abstract

Community Question Answering (CQA) websites are a popular tool for internet users to fulfill diverse information needs. Posted questions can be multiple sentences long and span diverse domains. They go beyond factoid questions and can be conversational, opinion-seeking and experiential questions, that might have multiple, potentially conflicting, useful answers from different users. In this paper, we describe a large-scale formative study to collect *commonsense properties* of questions and answers from 18 diverse communities from stackexchange.com. We collected 50,000 human judgments on 500 question-answer pairs. Commonsense properties are features that humans can extract and characterize reliably by using their commonsense knowledge and native language skills, and no special domain expertise is assumed. We report results and suggestions for designing human computation tasks for collecting commonsense semantic judgments.

1. Introduction

Community Question Answering (CQA) websites such as *Yahoo! Answers* and *StackExchange* have emerged over the past decade as a popular tool for information seekers to fulfill their information needs. These websites allow individuals to collect multiple responses from the community, choose one as the best answer, or wait for the community to select the best answer. The amount of user-generated content on these websites is growing rapidly and is available and accessible to people all over the world.

Due to the high volume of user-generated data and high user access rate to such websites, it is critical to propose methods and criteria to evaluate this massive corpus of data. Questions in this corpus elicit a variety of answers ranging from high quality and vital to low quality and abusive. This evaluation is useful as CQA websites have advantages that would make them richer than other archival media. These advantages include, but are not limited to, their rapid growth, coverage of diverse domains, easy and free user access, and availability in different languages. This evaluation of questions and answers, however, is challenging and goes beyond linguistic characteristics of the question or answer's body text.

In this work, we designed a human computation formative study to extract human annotators' commonsense judgments regarding fine-grained aspects of question-answer (QA) pairs. Instead of asking annotators for their expertise-based judgments on each QA pair, we targeted their intuition. These commonsense properties do not require any domain expertise, and are designed to capture commonsense knowledge of native speakers. The key contributions of this paper are:

- Designing a case study to demonstrate an iterative process for collecting fuzzy semantic judgments from crowdsourced workers.
- Demonstrating how social phrasing can be used to obtain more consistent, bias-free responses from annotators in human computation tasks.
- Demonstrating how standard evaluation metrics such as inter-rater agreement would fail to evaluate commonsense-based tasks, as some level of disagreement is expected and tolerable in collecting commonsense semantic judgments.

The paper is organized as follows. Section 2 describes the challenges in rigorously evaluating question answer content on CQA websites. Section 3 describes related work. Section 4 describes the motivations behind our proposed study and the structure of the commonsense properties human annotation task. Section 5 describes our formative study and quantitative and qualitative analyses of the data we collected. We discuss the lessons learned from each human computation experiment as well as our hypotheses on each empirical analysis in section 6. Finally, in section 7, we conclude the paper with challenges and suggestions for rigorously collecting commonsense knowledge from humans.

2. Challenges in Evaluation of CQA Content

The complexity of human evaluation methods includes the following factors:

2.1. Factoid Versus Subjective Questions: The evaluation of factoid questions has been started and tackled by venues such as TREC[1]. These venues have published

[1] http://trec.nist.gov/

a rich corpus of annotated questions and answers for factoid questions, which has been used to design machine learning models to detect correct and relevant answers and evaluate new QA pairs. This evaluation, however, is inadequate when the corpus consists of subjective and opinion-seeking questions, such as questions about personal experiences of people, reason and purpose of a social phenomenon, or individuals' thoughts on a specific situation. One example of a subjective question extracted from StackExchange's fitness category is: http://fitness.stackexchange.com/questions/7796/why-do-i-feel-more-out-of-shape-sometimes-doing-the-exact-same-workout*Why do I feel more out of shape sometimes doing the exact same workout?*

2.2. Answer Validity: Another challenge of evaluating the quality of an answer is the matter of whether or not it can be verified. Based on our observation, questions would fall into one of the following categories:

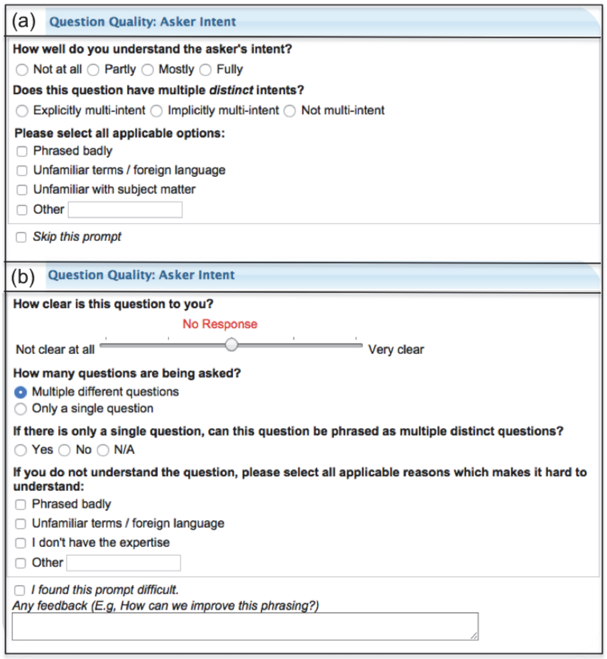
- One valid answer (e.g., *Who was the USA president in 2000?* which has a verifiable answer)
- No valid answer (e.g., *Do ghosts exist?* as there is no way one can prove whether ghosts exist or not)
- Multiple valid answers, consistent (e.g., *Does exercising make me healthier?* as people would tend to agree toward a positive answer, but might cover different aspects such as mental health, blood pressure, and weight control in their answers)
- Multiple valid answers, conflicting (e.g., *What can I do to encourage my 12 year old son to enjoy productive activities?* as people would likely have different personal experiences)

2.3. Time Sensitivity: Even if a verifiable answer exists, it might depend on the time the question is being asked. For instance, the question *when is Thanksgiving?* has a verifiable answer at this point of time that would be incorrect within a year.

2.4. Expertise: Given all the diverse domains for questions on CQA websites, it would be expensive to hire domain experts for each category. Thus, designing a general human computation task for evaluating these topically diverse QA pairs would be challenging.

3. Related Work

There are a small number of papers that have focused on the evaluation of the quality of questions and answers in CQA websites. Shah and Pomerantz (2010) focused on the quality of answers, selected a subset of Yahoo! Answers questions, and evaluated 13 different aspects of each answer including its relevance to the question using Amazon Mechanical Turk workers. They observed a mismatch between those human-extracted features and the prediction of whether the answer would be selected as the best answer. Liu, Bian, and Agichtein (2008) designed an answer quality prediction model using basic entities in a CQA website such as question, answer, asker and answerer profiles, category and textual features. Surdeanu, Ciaramita, and Zaragoza (2008) focused on non-factoid questions and designed an answer ranking system using mostly NLP features such as web correlation and word frequency features. Zhu, Bernhard, and Gurevych (2009) tackled the problem of evaluating answer quality by extracting 13 human-generated features including informativeness and politeness, and computed correlations between these features. Ravi et al. (2014) focused on different aspects of the question itself, and addressed the problem of question quality. They also investigated users' behavior and how it relates to different question contents. Chua and Balkunje (2012) performed a comparative evaluation of six CQA websites from three different perspectives: information organization, information quality, and website usability. To design semantic judgment collection questionnaires, Rattray and Jones (2007) investigated the systematic development of questionnaire design by presenting a framework that critically evaluates the theoretical and methodological issues of developing such questionnaires and suggests a list of heuristic decision-making strategies at each stage of development.



(a) Question Quality: Asker Intent

How well do you understand the asker's intent?

○ Not at all ○ Partly ○ Mostly ○ Fully

Does this question have multiple *distinct* intents?

○ Explicitly multi-intent ○ Implicitly multi-intent ○ Not multi-intent

Please select all applicable options:

□ Phrased badly
□ Unfamiliar terms / foreign language
□ Unfamiliar with subject matter
□ Other

□ *Skip this prompt*

(b) Question Quality: Asker Intent

How clear is this question to you?

No Response

Not clear at all — Very clear

How many questions are being asked?

● Multiple different questions
○ Only a single question

If there is only a single question, can this question be phrased as multiple distinct questions?

○ Yes ○ No ○ N/A

If you do not understand the question, please select all applicable reasons which makes it hard to understand:

□ Phrased badly
□ Unfamiliar terms / foreign language
□ I don't have the expertise
□ Other

□ *I found this prompt difficult.*

Any feedback (E.g, How can we improve this phrasing?)

Figure 1: A comparison of *Asker Intent* prompt from V1 (a), and V2 (b). Phrasing simplification and answer scale modification are the main changes applied on this prompt.

4. Methods

In this section, we begin by describing commonsense properties. Next we summarize the findings of the most related previous work which motivated us to perform a new formative study. After explaining the main motivation to design our new human computation task, we describe our dataset and other task properties.

4.1. Commonsense Properties

Our formative study was designed to collect commonsense semantic judgments around multiple aspects of QA pairs. We aimed to design a group of properties and features for questions and answers that are instinctive and intuitive for annotators and can be collected without any special training. An example of such a property is *helpfulness*, which can be assessed without having any type of domain expertise (Danescu-Niculescu-Mizil et al. (2009)). In this work, we refer to CQA websites (e.g., StackExchange) as communities. Also we refer to specific subjects (e.g., parenting, cooking) as domains. Our task was designed in a way that it can be reused across different domains and communities.

To collect commonsense properties, we designed an experiment in which annotators were first shown a question and were asked to provide their best judgment on multiple commonsense properties, which were designed to cover as many informative aspects of the question as possible. We did not show the answer at this point, to prevent bias. After providing their judgments on commonsense properties of the question, annotators were redirected to another page in which both the question and the best extracted answer were revealed. They were then asked to provide their best judgments on multiple fine-grained aspects of the answer and its relationship with the corresponding question. The prompts and related commonsense properties are described in Table 1.

4.2. Formative Study 1 (V1)

We build upon on a recent project by Barua and Paritosh (2015) that presented the results of a human annotation task to evaluate fine-grained properties of questions and answers. Table 1 (column 2) demonstrates the human computation task they performed to collect data from annotators. In this paper, we refer to that study as *V1*. To evaluate the results, they used *inter-rater agreement/reliability* (Hayes and Krippendorff (2007)) as a commonly-used metric in linguistic annotation and content analysis to measure the agreement among the annotators for each commonsense property.

For some commonsense properties, they observed a high inter-rater reliability suggesting that people could reliably agree on describing those fine-grained aspects of the QA pair. On the other hand, some properties had a low inter-rater reliability. They further inspected the properties with low inter-rater agreement by calculating the top-selected options for each property, and demonstrated that most of the annotations were selected from a small number of options they termed *top buckets*. This proves that the collected annotations are not evenly distributed and have skewed distributions. Besides considering the adverse effect of skewed distribution on inter-rater reliability, they also hypothesized that this metric is not sufficient to judge annotators' validity of work for many of the subjective commonsense properties, as personal features such as expertise might heavily affect semantic judgments.

4.3. Motivation to Design Formative Study 2 (V2)

To address the issues reported in V1, we conducted a formative study as a follow-up to the work of Barua and Paritosh (2015) in order to understand challenges with extracting semantic judgments and find the best practices for designing such an experiment. We refer to that study as *V2*.

As a skewed distribution of data was observed over most of the collected judgments in V1, we changed the scale on which responses are collected to make them more granular and investigate if this modification can make data more evenly-distributed.

Next, to make sure that annotators perceive the properties in the way we desired, we discussed the prompt questions in our research group. In case of a misunderstanding or a non-mutual perception about a commonsense property, we replaced the corresponding question with a more intuitive version by gathering participants' mutual intuition about the best way of communicating that property to annotators.

Also, motivated by the work of Gilbert (2014) on asking social versions of questions such as *Do you think other people would find this review helpful?* in addition to the traditional personal question of *Do you find this review helpful?* to get low-variance results, we explored this methodology by creating a social phrasing version and compared the results to the personal phrasing version. Our follow-up formative study is discussed in the following section.

4.4. Human Computation Task Design

Similar to Barua and Paritosh (2015), we chose 18 topically-diverse domains within the Life/Arts category in StackExchange, including Graphic Design, Cooking, and Parenting. These topics were either factual or subjective, containing technical domains in which expertise is required to provide a valid answer (e.g., Graphic Design), and experiential domains (e.g., Parenting). We selected 500 questions from all the domains. Question title, elaboration, and best answer were retained for each QA pair and presented to annotators.

We recruited crowdsourced workers through ODesk crowdsourcing platform. They were all native English speakers. We paid workers by the hour and approved all submitted tasks. Unlike V1, which had three unique annotators for each QA pair, we recruited five unique workers to create annotations on the same set of QA pairs for V2.

5. Experiments and Results

Inspired by the findings from V1, we designed another experiment with four major changes. First, we explored suggestions from related work on using social phrasing to collect commonsense judgments. We then simplified the phrasing and sentence structure to make the experiment more intuitive. Finally, we changed the range of the collected responses from radiobox options to range sliders, to test whether we would obtain less skewed data.

In this section, we describe our formative study and the iterative steps we took to design the task. We then discuss the impact of each step on the results.

5.1. Social Phrasing

The Bayesian Truth Serum, as described by Prelec (2004), is a technique for collecting subjective judgments from

Table 1: Human computation task to collect fine-grained commonsense properties on question and answer pairs (Column 2: V1 and Column 3: V2). Our modifications included applying *Social Phrasing* (e.g., Row 6), *Phrasing Simplification* (e.g., Rows 1 and 5), *Adding New Properties* (e.g., row 8), and *Answer Scale Changes* (e.g., row 2).

Prompt header	V1 Prompt	V2 Prompt
1. Asker Intent	(a) How well do you understand the asker's intent? (Radiobox options: Not at all, Partly, Mostly, Fully) (b) Does this question have multiple distinct intents? (Radiobox options: Explicitly multi-intent, Implicitly multi-intent, Not multi-intent)	(a) How clear is this question to you? (Range slider from: Not clear at all to Very clear) (b) How many questions are being asked? (Radiobox options: Multiple different questions, Only a single question) (c) If there is only a single question, can this question be,phrased as multiple distinct questions? (Radiobox options: Yes, No, N/A)
2. Question Clarity	(a) How well written is the question? (Radiobox options: Not well written, Moderately well written, Very well written)	(a) How well written is this question? (Range slider from: Not well written to Very well written)
3. Question Type	(a) This question can be broadly categorized as one seeking: (Checkbox options: Factual information, Opinions, Not really a question) (b) Further, this question is seeking one or more: (Checkbox options: Definition, Procedure, Reason/Explanation, Time, Amount, Word seeking, Location, Entity, Reason for being notable, Yes/No, Choice, Other)	(a) How do you describe the goal of this question? (Radiobox options: Looking for facts, Looking for opinions, Looking for both facts and opinions, Not really a question) (b) Further, this question is seeking one or more: (Checkbox options:Definition, Procedure, Reason/Explanation, Time, Amount/Measurement, Looking for a specific word, Location, Entity, Reason for being famous, Yes or No, Choice (which does not have a yes-no response), Other (c) Is this question asking about multiple subjects? (Radiobox options: Yes, No)
4. Answerability	(a) This question can be satisfactorily answered by a: (Radiobox options: Single phrase, List of phrases, Short answer (1-3 sentences), Long answer (>3 sentences)) (b) Will different valid answers to this question be consistent with each other?: (Radiobox options: Yes, No)	(a) How long do you expect the answer to be? (Radiobox options: Words, Sentences, Paragraphs) (b) How many valid answers do you think this question has? (Radiobox options: No valid answer, One valid answer, Multiple valid answers that are consistent, Multiple valid answers that conflict) (c) Can the answer to this question be verified? (Radiobox options: Yes, No)
5. Open-endedness	(a) Is this question open-ended? (Radiobox options: Not open-ended, Open-ended)	(a) Is the asker of this question looking for a specific piece of information or interested in general informaion which enables him/her to discover further, and ask other related questions? (Radiobox options: Looking for specific information, Looking for general information to explore the domain)
6. Interestingness	(a) Do you find this question interesting? (Radiobox options: Yes, No) (b) This question could be interesting to: (Checkbox options: Almost everyone, People interested in the specific domain of the question, People in a specific location, Children (ages 4-12))	(a) Do you find this question interesting? (Range slider from: Not interesting at all to Very interesting) (b) Are you curious to see the answer of this question? (Range slider from: Not curious at all to Very curious) (c) Would you like to share this question with other people? (Range slider from: Would not share at all to Would like to share) (d) Do you think other people might find it interesting? (Range slider from: Not interesting at all to Very interesting)(e)Which group(s) of people might find this question interesting? (choose all applicable options): (Checkbox options: Almost everyone, Specific community, People in a specific location, Specific age range (kids, teenagers, adults), Specific gender (female, male), other)
7. Expertise	(a) How familiar with the domain of this question are you? (Radiobox options: Unfamiliar, Familiar, Very familiar) (b) How familiar with the domain of this question do you think the writer of the question is? (Radiobox options: Unfamiliar, Familiar, Very familiar)	(a) How familiar with the domain of this question are you? (Range slider from: Unfamiliar to Very familiar) (b) How familiar do you think general internet users are with the domain of this question? (Range slider from: Unfamiliar to Very familiar) (c) How familiar with the domain of this question do you think the writer of the question is? (Range slider from: Unfamiliar to Very familiar)
8. Time Sensitivity	-	(a) Do you think the answer to this question would be changed depending on the time it is being asked? (Radiobox options: No, Maybe, Yes)
9. Answer Relevance	(a) How relevant is the answer to this question? (Radiobox options: Off topic/ useless, Slightly relevant, Relevant, Useful, Vital)	(a) How clear is the answer to you? (Range slider from:,Not clear at all to Very clear) (b) How close is the answer to the subject of the question? (Range slider from: Not on-the-topic at all to Very on-the-topic)
10. Answer Clarity	(a) How well written is the answer? (Radiobox options: Not well written, Moderately well written, Very well written)	(a) How well written is the answer? (Range slider from: Not well written to Very well written)
11. Level of Information	(a) The answer contains: (Radiobox options: Less information than required, Right level of information, More information than required, Extraneous information which is not useful)	(a) The answer contains: (Radiobox options: Less information than required, Right level of information, More information than required, Extraneous information which is not useful)
12. Answer Plausibility and Helpfulness	(a) Is the answer plausible? (Radiobox options: No, Maybe, Yes) (b) Is the answer helpful? (Radiobox options: No, Maybe, Yes)	(a) Is the answer plausible? (Radiobox options: No, Maybe, Yes) (b) Is the answer helpful? (Radiobox options: No, Maybe,Yes)
13. Answer Satisfaction	(a) How well does the answer satisfy the question's intent? (Radiobox options: Fails to meet, Slightly meets, Moderately meets, Highly meets, Fully meets)	(a) How well do you think the answer satisfies the asker? (Range slider from: Fails to satisfy to Fully satisfies)
14. Expertise	(a) How familiar with the domain of this question do you think the writer of the answer is? (Radiobox options: Unfamiliar, Familiar, Very familiar)	(a) How familiar with the domain of this question do you think the writer of the answer is? (Range slider from: Unfamiliar to Very Familiar)

crowds. In this technique, annotators are asked two questions: one with *personal* phrasing that targets their personal subjective opinions on a specific matter, and another with *social* phrasing that targets their best judgments on how other people would approach that matter. As previous studies (Gilbert, 2014; Shaw, Horton, and Chen, 2011) have shown, traditional methods of eliciting personal reviews often introduce a level of individual biases. Although this level of personal bias is helpful to personalize different services for a specific user, this is not tolerable in systems which rely on collective intelligence. For those types of systems where future predictions would affect all people, the best general

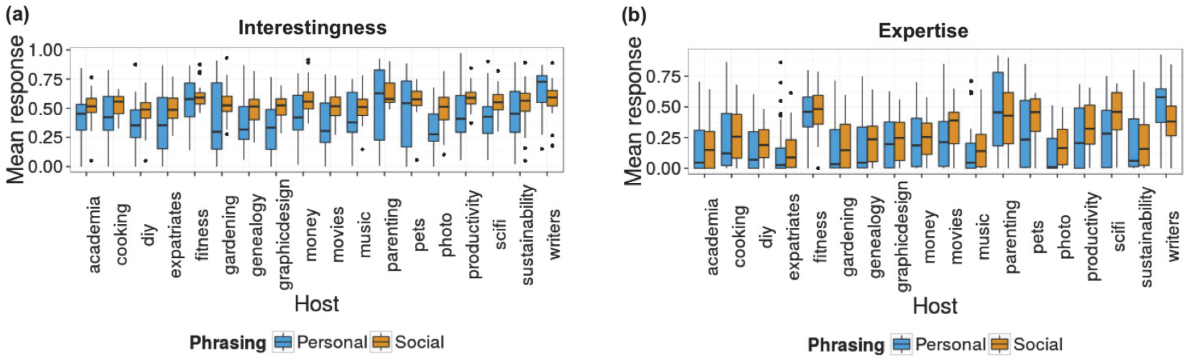


Figure 2: Variability within personal and social phrasing scores obtained for 18 domains (personal phrasing in blue; social phrasing in brown) for (a) question interestingness, and (b) domain expertise. Social phrasing's mean scores are mostly higher than personal phrasing across all domains. Annotators are more consistent in responding to the social version as opposed to the personal version.

opinion of a large group of people is required to be elicited. Inspired by previous findings on how social phrasing would help to exploit people's meta-knowledge of other people, we chose two of our questions for which the social phrasing made sense (*Question interestingness* and *Expertise*), and asked both personal and social phrasing questions. In other words, we first asked the annotators to answer prompts based on their personal opinion and perception, then asked them to provide their best guess about how other people (general internet users) would answer the same question. So, for the questions *Do you find this question interesting?* and *How familiar with the domain of this question are you?* we added two more questions that we call social phrasing versions, asking for annotators' best guess on how interesting they think general internet users would find the question (*Do you think other people might find this question interesting?*) and how familiar with the domain of the question they think other people are (*How familiar do you think general internet users are with the domain of the question?*). Their responses were collected on a scale from 0 to 100 for both personal and social phrasing questions.

5.1.1. Social Phrasing Results: Figure 2 represents the mean scores of all responses across multiple domains for both *Question interestingness* and *Expertise* prompts. It demonstrates that social phrasing led to statistically significantly different results than personal phrasing.

The mean scores of the annotators' responses for the social phrasing version were mostly higher than the same personal phrasing version across all domains for both prompts. It shows that annotators could predict that other people would likely be more interested and familiar with the domain of the question.

In addition, the standard deviation of the social phrasing's responses was significantly lower for *Question interestingness*, indicating that annotators tend to agree more when responding to the social version as opposed to the personal version. The social phrasing version of *Expertise* responses led to more consistent results as well. These results demonstrated that social phrasing lead to lower variance and increased convergence toward a common answer. This proves that people can successfully exploit their meta-knowledge about other people and make a consistent guess about others' behavior. It also strongly suggests that social phrasing can be used successfully to obtain more consistent, bias-free responses from annotators in human computation tasks.

5.2. Phrasing Simplifications

Our phrasing changes included the modifications we applied on prompt questions and answer choices, as well as the addition of new properties. In this section, we explain each modification and investigate the results.

As our goal to collect properties was meant to be commonsense, prompts were simplified by the removal of terms which could be complex for a lay person. For instance, we replaced the question *Is this question open-ended* with a more intuitive version, *Is the asker of this question looking for a specific piece of information, or interested in general information which enables him/her to discover further and ask other related questions?*.

We changed the phrasing of not only prompt questions, but of the collected answer choices. For instance, instead of asking people to predict the answer length to be *Single phrase*, *List of phrases*, *Short answer (1-3 sentences)*, or *Long answer (more than 3 sentences)*, we asked them to choose between three options of *word*, *sentences*, or *paragraphs*.

5.2.1. Unintended Meaning Changes: We observed that what we had considered as simplification of the phrasing often led to unintended changes in the inherent meaning of the prompt questions. For instance, this phenomenon was observed in the *asker intent* prompt which was changed from *How well do you understand the asker's intent?* to *How clear is this question to you?* Although unexpected, it was observed that the annotators provided lower ratings in V2 than in V1. We hypothesize that this happens because

V2 measures how well annotators understand the question, which requires them to comprehend all the details of the problem. Thus, unlike V1, the goal is not just understanding the asker's general purpose.

An example of this phenomenon is the response to this StackExchange question: http://sustainability.stackexchange.com/questions/689/how-to-measure-the-embodied-energy-of-a-regulation-is-data-available*How to measure the embodied energy of a regulation. Is data available?* While the responses were uniformly and unanimously *Mostly clear* for V1, the ratings in V2 were more varied. Three annotators voted for *Mostly clear* but with some variance on the range slider values, and one annotator voted for *Fully clear*. Since the question contains unfamiliar and technical words, it is not possible for every annotator to clearly understand what exactly the question is looking for. We also collected additional information in the form of flags which were supposed to indicate why the question was not understandable to annotators (Figure 1). We then used the selected flags to test our hypothesis regarding the change in the meaning of the asker intent prompt. As shown in Table 2, the Pearson correlation coefficients between the asker intent scores and the selected flags have increased negatively in study 2. In other words, in the second study, low asker intent scores tend to be more commonly associated with the selection of flags indicating the presence of unfamiliar terms.

For the sake of comparing the V1 and V2 prompts in which the answer options have been changed, we mapped the new options to the previous ones and calculated annotations' variation and IRR among multiple annotators in each study separately. We than calculated the variance of three annotations from V1 and five annotations from V2 separately over each query (QA pair). We next took an average over the dataset for both studies. The quantitative results suggested that the variance of annotators' responses have been increased for some of the prompts such as *Answer consistency* (V1: 0.07 and V2: 0.10), and remained about the same for some other properties such as *Open-endedness* (0.03 for both V1 and V2).

5.2.2. Reliability-reusability trade-off: These findings motivated us to take a closer look at the annotations qualitatively to understand what made the difference. We observed that our technical phrasing in V1 led to more consistent results. For instance, in our *Open-endedness* prompt, the word *open-ended* is a technical word which can be interpreted differently as we did not clarify what we meant by that. Thus, many annotators treated it as if questions with a single or multi-word answers are not open-ended, and questions which require more explanation are open-ended. For instance, the question *Postdoc position: should I stay for one year or two years?* is one which we expected to see rated as *open-ended*, as the person is looking to explore opportunities regarding his or her career. This question was marked as *not open-ended* by all three annotators of V1. The same question, however, was marked as *open-ended* by four annotators and *not open-ended* by one annotator from V2. This added to the variance calculated for V2, although the results were more accurate. The same pattern happened for the question *How can I keep my split pea soup from becoming too thick during storage?* which was marked as having *inconsistent answers* by all V1 annotators, yet marked as having *consistent answers* by four out of five V2 annotators and having *inconsistent answers* by the remaining annotator. We observed that although the variance has dropped in V2, the problem is defined as a more commonsense collection task in which a fair degree of disagreement is expected and tolerable. This improvement, however, not can be captured by inter-rater agreement, as it suggests a decrease in reliability. We call this observation *reliability-reusability trade-off.*

Table 2: Correlations of annotation scores and asker intent selected flags. It demonstrates that there is a higher correlation between the selection of the flags and marking a question as unclear in V2.

Flag / Correlation	**V1**	**V2**
Phrased badly	-0.23	-0.34
Unfamiliar terms	-0.14	-0.48
Unfamiliar subject matter	-0.24	-0.44

5.2.3. New Commonsense Properties: We also included more commonsense properties, as well as asking additional questions on previous ones. We added a new prompt, *Time Sensitivity*, and more questions to several prompts (e.g., added *Can the answer to this question be verified?* to the *Answerability* prompt).

We used Spearman's correlation rank test to examine the association between old and new properties. Our new feature *Answer verifiability* had a negative correlation with the *Opinion seeking* option (-0.41), and a positive correlation with *One valid answer* (0.34). Moreover, the new feature *Multidisciplinary* did not have a noticeable correlation with any of the existing properties. This pattern was observed for *Specific age range* and *Specific gender* as well. The lack of strong correlation for these new properties approves their valuable addition to the experiment, showing that their impact is not covered by previous properties. On the other hand, *Shareability* and *Curiosity to see the answer* had strong correlations with *Interestingness* (0.86 and 0.94 respectively). This suggests that these two items are being covered with *Interestingness*.

5.2.4. Instructions Removal: We also removed the detailed instructions, guidelines, and examples we had provided in the previous study to give annotators the freedom of selecting options based solely on their common sense. We hypothesized that as there is no domain expert for commonsense-based tasks, this would help in eliciting the biases that might be introduced by detailed guidelines.

As it can be observed in Table 3, the inter-rater reliability has dropped for most of the commonsense properties, indicating that for semantic judgments, the presence of some guidelines would help in calibrating people's opinions.

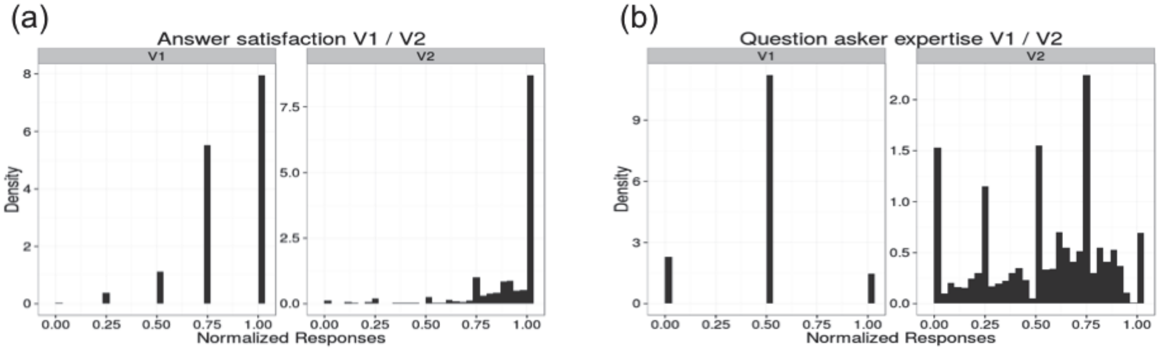


Figure 3: A side-by-side comparison of the answer distribution for V1 and V2 for (a) answer satisfaction and (b) question asker expertise. Adding more granularity to the response scale did not improve the response skew for (a), whereas it improved the distribution in (b).

5.3. Answer Scale Change

We also changed some of the radiobox options to range sliders to have a more granular range of 0 to 100. We chose to put five indicators on the sliders to make them similar to the Likert scale. We hypothesized that this would provide us with more granular data points, which would potentially solve the skewed data, and yet could be easily compared to the radiobox options of the previous study by discretizing the range and mapping each radiobox option to a specific data range on slider. This also helped us to reduce the number of undefined terms such as *Partly* and *Mostly*.

This change applied caused effects that heavily depended on the nature of the property. For some prompts such as *Answer satisfaction*, where there is some notion of a correct response, adding more granularity to the response scale did not lead to the expected improvement in response skew (Figure 3a). Responses in V2 continued to remain skewed, with a high proportion of responses being identically 100. However, for prompts such as *Expertise*, where there is less notion of a correct answer, we found annotators using the full range of the scale (Figure 3b).

Another issue arose from increased granularity in the response options' calibration, as we did not mark the five indicators of the range slider with the values they represent. Instead, we defined only the end points of the range slider, and left it to annotators to choose where in the range they wanted to select. This led to the problem of differing perceptions regarding where to place the slider hover. For example, considering the *Question clarity* prompt, one annotator's 25 on the range slider might be the same as another one's 70, since they both did not find the question clear but did not have a criterion to know how to mark their responses.

Figure 1 presents a snapshot of the *Asker intent* prompt.

6. Discussion

In this section, we describe our observations on inter-rater reliability of the commonsense properties. We further describe how the distribution of responses on properties vary across the eighteen domains within StackExchange. This is valuable, as it allows us to observe the differing information requirements within different domains. As seen in Table 3, *Interestingness*, *Expertise* and *Answer satisfaction* are the only properties for which the agreement has been increased in V2. In all other cases, however, we observed that it dropped. We hypothesize that the granular range sliders provided a suitable common ground for information extraction on the properties in which strong personal bias is desired and needed (e.g., describing the level of question interestingness, or estimating self-expertise in a domain). However, for the rest of the commonsense properties there is a more verifiable answer, which can be extracted by the aggregation of multiple annotations. For those cases, radiobox options worked better, as they helped focus people's diverse opinions on finite options.

Another observation is the variance of annotation results calculated for each collected commonsense property over all 18 domains. In most cases, this average variance was about the same number for both studies. This negligible difference suggests that we have not changed the question meaning completely, and two different versions still represented the same notion to annotators. However, the annotators' responses showed more disagreement in V2. Qualitative inspection of the results from V2 revealed that the results were more reusable, as more people converged toward the correct values of verifiable commonsense properties. The increase in inter-rater agreement shows that people had different personal biases as they formed judgments on the commonsense properties. We call this the *reliability-reusability trade-off*.

7. Conclusions and Future Work

In this paper, we describe a large-scale formative study which aims to collect and evaluate commonsense features of question answer pairs. Designing such a commonsense semantic judgment collection task is challenging, as there is no expert on commonsense to verify the results. We used different task design paradigms to learn what is the most reliable way to design such a task. We then inspected the inter-rater agreement to find the cases where humans could reliably converge on specific judgments. Lower agreement

Table 3: Inter-rater agreement across the two versions of the study. Higher inter-rater agreement is observed for most of the commonsense properties in V1.

Property	IRR V1	IRR V2
Asker Intent	0.552	0.391
Question clarity	0.809	0.449
Fact	0.836	0.532
Opinion	0.907	0.597
Question Type: procedure	0.901	0.757
Question Type: reason explanation	0.745	0.478
Question Type: location	0.726	0.568
Question Type: time	0.807	0.764
Question Type: yes/no	0.86	0.6
Question Type: entity	0.75	0.55
Question Type: word seeking	0.72	0.71
Question Type: choice	0.776	0.533
Question Type: definition	0.846	0.735
Open ended	0.81	0.524
Interestingness	0.306	0.358
Expertise answerer	0.25	0.43
Expertise asker	0.53	0.41
Answer relevance	0.527	0.238
Answer clarity	0.847	0.333
Answer plausible	0.559	0.294
Answer helpful	0.54	0.37
Answer satisfaction	0.462	0.476
Expertise answerer	0.24	0.436

values might not be considered as failure, since some level of genuine semantic ambiguity around commonsense notions is expected. We call this dilemma the *reliability-reusability trade-off*. Improving the reliability in task design results from increasing convergence in a small pool of annotators. Operationalizing commonsense notions like “helpfulness” in this manner can hurt the reusability of the collected annotations. We show some qualitative examples from our formative study to demonstrate this trade-off.

Collecting results on both personal and social phrasing provided evidence that social phrasing has less variance and faster convergence to aggregated annotators’ scores. Our future work includes a new formal study to investigate more aspects of social phrasing over all commonsense properties, and applying the Bayesian Truth Serum to both social and personal phrasing collected annotations.

8. Acknowledgments

We would like to thank Michael Tseng, Ka Wong and Eric Altendorf for insightful feedback on prior drafts. We are also thankful to the participants who provided annotations and feedback on the task.

References

Barua, A., and Paritosh, P. 2015. Using commonsense for deeper understanding of complex question answer content. In *ACM SIGIR Workshop on Web Question Answering: Beyond Factoids*.

Chua, A. Y. L., and Balkunje, R. S. 2012. Comparative evaluation of community question answering websites. In *Proceedings of the 14th International Conference on Asia-Pacific Digital Libraries*, volume 7634, 209–218. Springer.

Danescu-Niculescu-Mizil, C.; Kossinets, G.; Kleinberg, J.; and Lee, L. 2009. How opinions are received by online communities: a case study on amazon. com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*, 141–150. ACM.

Gilbert, E. 2014. What if we ask a different question?: Social inferences create product ratings faster. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2759–2762. New York, NY, USA: ACM.

Hayes, A. F., and Krippendorff, K. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 1(1):77–89.

Liu, Y.; Bian, J.; and Agichtein, E. 2008. Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’08, 483–490. ACM.

Prelec, D. 2004. A bayesian truth serum for subjective data. *Science* 306(5695):462–466.

Rattray, J., and Jones, M. C. 2007. Essential elements of questionnaire design and development. *Journal of Clinical Nursing* 16(2):234–243.

Ravi, S.; Pang, B.; Rastogi, V.; and Kumar, R. 2014. Great question! Question Quality in Community Q&A. In *International AAAI Conference on Weblogs and Social Media*.

Shah, C., and Pomerantz, J. 2010. Evaluating and predicting answer quality in community QA. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 411–418. ACM.

Shaw, A. D.; Horton, J. J.; and Chen, D. L. 2011. Designing incentives for inexpert human raters. In *ACM Conference on Computer Supported Cooperative Work*, 275–284.

Surdeanu, M.; Ciaramita, M.; and Zaragoza, H. 2008. Learning to rank answers on large online qa collections. In *In Proceedings of the 46th Annual Meeting for the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT*, 719–727.

Zhu, Z.; Bernhard, D.; and Gurevych, I. 2009. A multi-dimensional model for assessing the quality of answers in social q & a sites. In *Proceedings of the 14th International Conference on Information Quality, ICIQ 2009, Hasso Plattner Institute, University of Potsdam, Germany, November 7-8 2009*, 264–265.