Data Collection and Crowd-Work Quality Management for GLUCOSE (Generalized and Contextualized Story Explanations)

Nasrin Mostafazadeh, Lori Moon, Lauren Berkowitz, and Jennifer Chu-Carroll
Elemental Cognition
glucose@elementalcognition.com

1 Introduction

GLUCOSE, GeneraLized and COntextualized Story Explanations, is a large-scale common sense knowledge acquisition task. Given a short story and a sentence X that is contained in the story, the GLUCOSE task asks crowd workers to compose specific statements about the story and general rules about the world based on their specific statements. There are 10 specific statements and 10 general rules solicited per sentence, guided by ten dimensions of common sense causal explanation.

Elemental Cognition embarked on an effort to collect over 250,000 common-sense knowledge rules to help systems to reason better with regarding causes and effects in story understanding. The work is described in [ACL link], and this document describes the data collection schema and process and data quality control.

We collected a total of 337,636 pairs of specific statements and general rules. We used various data quality management techniques, described in this document, to classify the data into three quality levels. The quality ratings are: **3 = Highest quality rating** These data are characterized by having rules that display an accurate level of generalization from the story to other situations in the world. They are grammatical and not overly-complicated. **2 = Mid-quality rating** These data points are still very good rules that have a good balance of generality, but a higher percentage in this set may be overly specific, use less clear wording, or not be as conceptually concise. **1 = Lower-quality rating** These rules are still useful, but a higher percentage of them have, in addition to the issues in the 2-level ratings, highly specific general rules and some misunderstanding of how to word the rules and statements.

Of the collected data, 28.4% are given the highest quality rank, '3', 44.9% are given the second highest quality rank, '2', and 26.7% are given the quality rating

of '1'. All rules collected were done by workers who went through qualification and training for the task and our own results, reported in [ACL citation], used all work quality levels in training and development of language models.

This document describes the date collection task, the data collection process, the work quality management, and analyses of the collected data. The data download and pre-trained models using the data are available at `https://github.com/ElementalCognition/glucose`. For results from using the GLuCOSE rules in model training see [ACL 2020 paper].

2 GLUCOSE Data Schema

The GLUCOSE Task involves collecting human-generated common sense reasoning information related to understanding narrative text. Stories and other kinds of narrative text play a prominent role in human communication. Much of what humans understand in narratives is not stated directly in text. For example, humans have a basic understanding of where objects and individuals are relative to each other in a story, what events in the story are the causes or motivations for other events, and how people feel during these events. We prompted readers to make the statements and rules that go into understanding a story explicit by having them write sentences about the narrative.

There are many dimensions of common sense reasoning that go into understanding narratives, but we restrict the task to five types. The five types are (1) Events or states that cause or enable other events or states, (2) basic human drives and motivations that cause or enable other events or states, (3) locations of items, individuals, events, or states that cause or enable other events or states (4) possession of items or traits that cause or enable events or states and (5) other attributes, such as changes of states, that cause or enable events or states.

For each of these five dimensions, workers considered a sentence from a story. They were asked whether the sentence was connected to one of the five dimensions, either before or after the time of the target sentence, given the story context.[1]

An example of a story is below:

> My daughter was born last week. I have a ton of pictures to show you.
> She is so cute. She has a ton of hair. I am in love.

For each sentence in the story, workers are asked whether each of the 5 dimensions applied. For example, if X were the second sentence below, workers were asked whether each dimension applied to that sentence, given the containing story.

[1]The stories used in the task were from ROC stories [citation]. We restricted the set of stories used to those with a vocabulary no larger than that expected for an eight year old child.

I have a ton of pictures to show you.

Workers were asked first about the things that occurred before the sentence. They were asked if there was anything in the story that caused or enabled the events or states in the sentence. If there was something relevant, then they wrote a specific statement about it. For example, a worker might say that the fact that the author's daughter was born enabled him to have pictures of her to show. Next, the worker was asked to make a general rule about the world based on the specific statement that they wrote. For example, based on the previous statement, they might say that the fact that someone is born enables her to be photographed by her parents.

For the next dimension, workers were asked, about the same sentence, if there were any basic human desires or motivations that caused the events or states in the sentence, in this case 'I have a ton of pictures to show you'. A specific statement would be that the parent's feeling of love motivated him to show pictures of his daughter. A general rule about the statement could be that someone's feelings of love motivate him to brag about his loved one.

This process is repeated for locations, for possessions, and for any other attributes. The ten question used in the data schema are below:

Dimension 1 An event that directly causes or enables the sentence X

Question 1 Consider the events that happen before X (or are likely to happen). Does any of them directly cause X, or simply make X possible (i.e., enable X)?

Dimension 2 An emotion or basic human drive that motivates X

Question 2 Consider the likely emotions and basic human drives of the participants in X. Does any of these states of mind/feelings motivate the participant to do X?

Dimension 3 A location state that enables X

Question 3 Consider the likely locations of the story participants (people, things, etc.) before X. Does one of these location states make X possible?

Dimension 4 A possession state that enables X

Question 4 Consider which things each story participant possesses (or is likely to possess) at different times. Does any of these possession states make X possible? (This often happens when X is an event that involves physical transfer, change of ownership, or destroying something.)

Dimension 5 Other property (besides location, emotional state, or possession) that enables X

Question 5 Consider everything else about the participants (people, things, etc.) in the story. Does some likely state of a participant besides their location, emotional state, and possessions make X possible?

Dimension 6 An event that is directly caused or enabled by X

Question 6 Consider the events that happen after X (or are likely to happen). Does X directly cause any of them, or simply make it possible (i.e., enable it)?

Dimension 7 An emotion that is caused by X

Question 7 Consider the likely emotions of the participants in X and those affected by it. Is any of these emotions caused by X?

Dimension 8 A change of location that X results in

Question 8 Consider the likely locations of the story participants (people, things, etc.) after X. Does X directly result in any of these location states? (This often happens when X is an event that involves change of location or movement.

Dimension 9 A change of possession that X results in

Question 9 Consider which things each story participant possesses (or is likely to possess) at different times. Does X directly result in any of these possession states? (This often happens when X is an event that involves physical transfer, change of ownership, or creating something.

Dimension 10 Other change in property (besides location, emotional state, or possession) that X results in

Question 10 Consider everything else about the participants (people, things, etc.) in the story. Does X directly result in some participant being in some state? For this question, ignore locations, emotional states, and possessions.

For dimensions 6-10, the questions are about what happens after the target sentence, rather than before. For example, for question 6, a specific statement is that the author has pictures to show people causes people to see pictures. A general rule about that statement is that someone showing someone else something results in them seeing something.

Responses to each of the ten dimensions are gathered for each sentence, in a story context. Not all dimensions are relevant for every sentence, but if a worker does consider a dimension to be relevant, then the worker must fill out both the specific statement and the general rule about it.

The result of the task is that stories have associated with them a set of explicit statements about the world and the connections among events in the story. Furthermore, each story is the inspiration for a set of general rules that could apply to other narratives.

3 Data Collection and Quality Management

Data were collected through a user interface that was designed at Elemental Cognition and hosted by Amazon Web Services (AWS) Mechanical Turk (Mturk). Crowd workers were recruited exclusively through Mturk, but they were vetted with several layers of testing and training designed at Elemental Cognition. All data scoring, classifying, and analysis was done at Elemental Cognition with in-house tooling that retrieved crowd worker responses submitted through our user interface to AWS Mturk.

Mturk requires that tasks be broken down into units called HITs (Human Intelligence Tasks). These are the units that workers are paid to do. Because crowd workers accept and work on HITs from many different requesters, it is important to follow similar practices in terms of the amount of work per HIT and the rate of compensation. One of the challenges in our task design was in determining how to break down the GLUCOSE task into HITs and how much to pay for those HITs in order to get the work quality level needed for the task.

This section describes how we translated the GLUCOSE data schema into a user friendly interface that worked with the constraints of the Mturk crowdsourcing host. We also considered ways to promote a uniform structure in user responses to make the response data easier to use in NLP applications.

Although Mturk provides access to a large worker pool, the level of worker quality management that is supports is very limited. Therefore, in addition to task design, we designed and maintained data quality management strategies to ensure high quality responses to the task.

3.1 Qualification Task

In order to qualify workers for the main GLUCOSE task, crowd workers had to take a test. AWS Mturk allows for only very limited question types in their qualification framework, so we developed a HIT with an in-house user interface to thoroughly

test workers for entry to the main task.[2]

3.1.1 Participants

The GLUCOSE qualification task used three of these Mturk-internal ratings: (1.) the worker had done more than 100 HITs on Mturk, (2.) The worker had an acceptance rate of at least 95%, (3.) The worker had an IP address matching a list of country codes.[3]

We did not limit how many times workers could take the test, however, they could only take it once per launch. Each launch of the qualification task was a one-HIT launch with 1,500 workers able to do the HIT. The task was launched nine times over the course of running the main task. A pool of 1039 workers qualified for our main task via the Qualifying Exam.

3.1.2 Qualification Content

The qualification content consisted of the qualification HIT and training materials for doing the qualification HIT. Users were encouraged to carefully read the instructions in the qualification HIT UI, which included stories with examples of GLUCOSE style specific statements and general rules about the stories. The qualification HIT UI also provided a link to a text document that contained general guidelines for the task.[4] This document was revised and expanded in the course of the task to improve worker training. The general guidelines also contained a link to a FAQ page that was also continuously revised and updated as workers asked about the task.

The qualification test contained questions testing expertise in three areas. The first area involved identifying correct use of the UI slots for language expressions. The main task would use slots to guide annotators in selecting the essential expressions to use in a rule. These slots had labels like ”subject” and ”verb” and sometimes had drop-down lists of selections. For example, ”preposition” had a

[2]We piloted the qualification task two times before using it to recruit workers for the main task.

[3]Country codes were originally restricted to a standard list Mturk provides, containing the US, Great Britain, Ireland, New Zealand, Australia, South Africa, and Canada. Half way into the project, we opened it up to many other countries per worker requests, including India, China, Germany, Jamaica, the Dominican Republic, Belize, and Nigeria. Many countries that are not in the standard list are already English speaking, and, furthermore, the qualification test itself would be challenging to pass if someone did not already have a good command of English. Since we were not limited to a particular dialect of English, we attempted to add the country code of nearly 50 countries which are English-speaking, but the Mturk interface limited us to 35 countries.

[4]The shared document is at: `https://docs.google.com/document/d/1W8S7y97G9yoAO5qVCHrBB7CTctVszTOrLCw1gP0dw-U/edit`

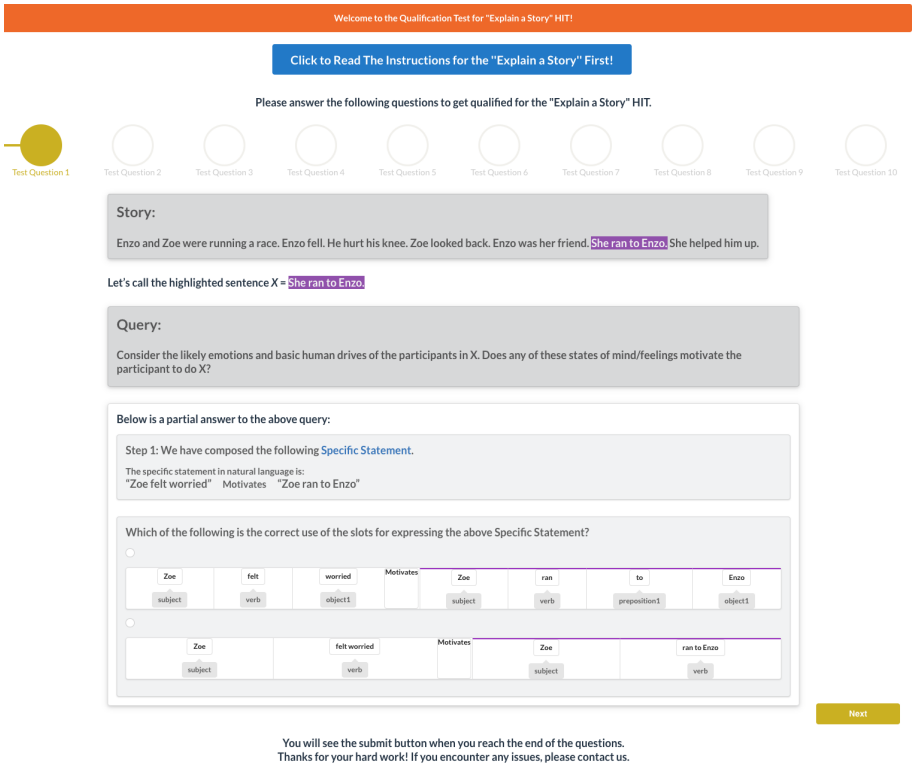

Figure 1: Qualification question about slot use

drop-down list of prepositions that users could choose. In anticipation of this format, we put multiple-choice questions on the qualification test HIT that showed correct and incorrect uses of the slots. An example is in Figure 1). In this example, the crowd-worker had to choose which slot use was correct for the sentence, 'Zoe felt worried'. One choice was 'Zoe (subject) felt (verb) worried (object1)' and the other was 'Zoe (subject) felt worried (verb)'. The correct answer was the former, in which 'worried' takes the object1 slot. The slots were simplified from part of speech tags to allow for easier understanding and to encourage simplification of the language of the stories.

The second area covered in the test involved recognizing the right level of generalization (see Figure 2). For generalization understanding, prospective workers were presented with a specific statement based on the target story. Then they had to choose the best general rule to derive from that statement. This type of question tested their understanding of the variable format (e.g., Someone_A), their understanding of the fact that proper names do not belong in general rules, and their

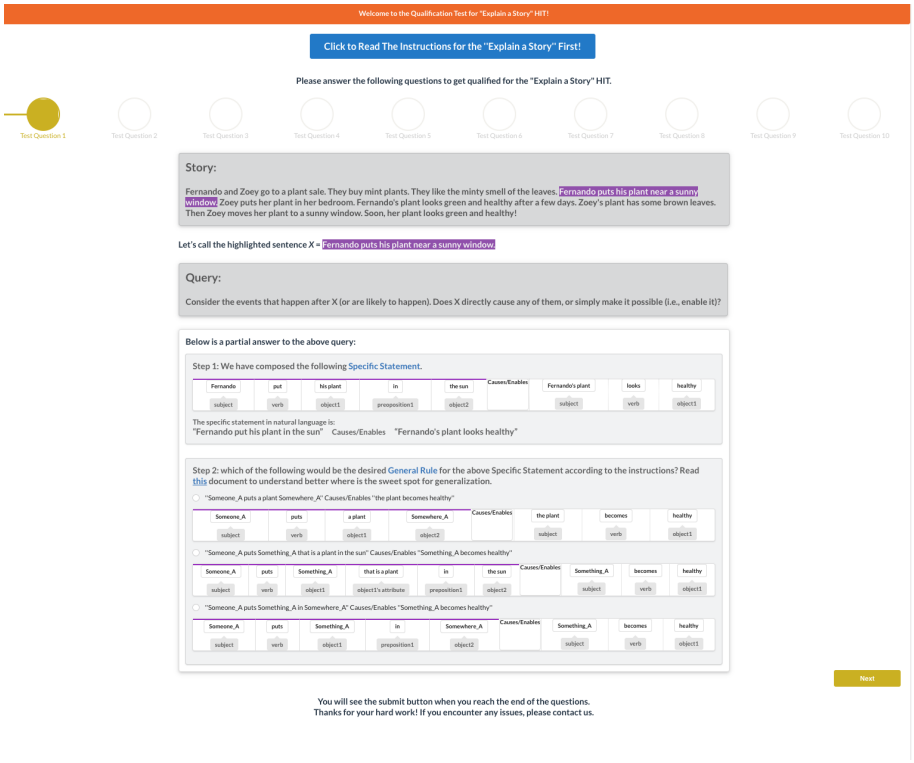

Figure 2: Sample qualification question about the right level of generalization

understanding that, if something is too general, it does not make sense as a rule.

The third area was identifying causes and effects with proper temporal understanding of the stories (see Figure 3). For understanding causes and effects, users were presented with sentences in the cause and effect structure, one of which was a valid cause-effect in the story and the others which were not.

3.1.3 Results

We found that the qualification task was sufficient to populate the main task with qualified workers. We found that workers were willing to engage with lengthy training documents in order to do well on the qualification HIT. One big motivating factor for workers was the size of the main task batches. Because passing the qualification test HIT allowed them to do an unlimited amount of the thousands of GLUCOSE main task HITs that were available, they were very motivated pass.

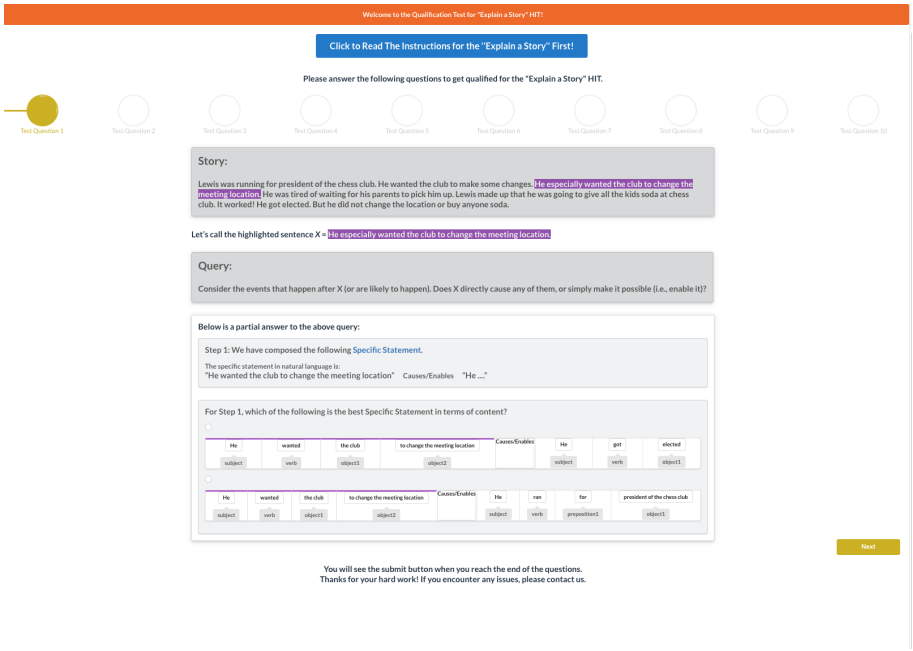

Figure 3: Sample qualification question about what events follow

3.2 Warm-up Task

In our mid-rounds, we started an additional level of testing for entrance called 'the warm-up task' as a way to ease workers into the large-scale task, if they did not get 100% on the qualification test, but got over 70% correct.

The warm-up was a sample of HITs like those seen in the main task, but it had simpler sentences and only seven HITs. We requested that workers submit three of the seven HITs, so that we had an ample sample-size to give feedback and see if they were ready for the large-scale task.

3.3 Main GLUCOSE Task

3.3.1 Participants

Of the 1039 workers who qualified for the main GLUCOSE task via the qualification task and warm-up, 373 contributed to the main task.

3.3.2 Main UI Content:

Qualifying workers were able to access large batches of data with no limit on how many HITs they could complete. The main UI displayed a page like in figure 5 for each of the 10 GLUCOSE dimensions in the same order each time. When they said that the dimension was relevant for the selected sentence (by clicking 'Yes..' in the "Your Answer:" box), they were taken to a screen where they could input answers, as shown in figure 5. We choose not to randomize the dimension order because the task was already long, and the predictable order of questions was something that helped workers form a strategy for doing the task in a manageable amount of time.[5]

For the Specific Statements, users freely entered text within the constraints of some part-of-speech guidance for some dimensions, but, for others, there were constraints on verbs or emotions, as shown in table 1. All preposition slots contained a drop-down list of English prepositions.

For General Rules, the subject position had a drop-down menu of variables for people, places, and things, such as "Someone_A". The verb slot was constrained in the same way as with the corresponding Specific Statement for that dimension.

3.4 Data Quality Control Pipeline

For work contributed through the main UI, data quality was controlled through daily monitoring of a percentage of incoming submissions and statistics on average dimensions filled out and time. The percentage of answers reviewed by hand were used to modify worker ratings. Figure 6 shows the strategic flow of worker ratings. Workers enter the tasked with a score of "-1" then advance to "2" as they become more proficient, getting a bonus increase. The top numeric rating is "3", which has an additional bonus increase. Select workers with a "3" rating were also moved into "top rated" batches that paid more per HIT and included higher bonuses and incentives. If work quality dropped, workers' ratings were reduced. If their work was at a risk of degrading corpus quality, they were given a "0" and disqualified from the task. All incoming submissions were approved but did not receive a bonus. Several workers were disqualified and then worked to re-qualify and became top rated workers. The General Guidelines and Qualifying Exam, combined with individual feedback on answers, often provided sufficient remedial learning to get good responses again. Most data quality issues were due to workers trying to rush at the task and not reading the General Guidelines carefully. Figure

[5]The average time it took a worker to complete one HIT was 8 minutes and 2 second, and the median was 5 minutes and 8 seconds. These numbers, however, do not account for the fact that workers often open multiple HITs and complete them sequentially.

dimension	task connective	Slot Constraints
dim 1 An event that directly causes or enables X	Causes/Enables	none
dim 2 An emotion or basic human drive that motivates X	Motivates	verb [feels, wants, likes] object [curiosity, independence, competition, honor, approval, power, status, romance, success, friendship, belonging, health, safety, livelihood, happy, stressed, angered, disgusted, sad, surprised, fearful, trusting, love, obedient, amazed, disappointment, regret, worthless, aggression, optimistic]
dim 3 A location state that enables X	Enables	verb [am is are] preposition [above, across from, at, below, far from, in, in front of, inside of,near, next to, on top of, outside of]
dim 4 A possession state that enables X	Enables	verb [possesses]
dim 5 Other property besides location, emotional state, and possessions make X possible	Enables	verb [am, is, are, has, have, want, wants, need, needs]
dim 6 An event that is directly caused or enabled by X	Causes/Enables	none
dim 7 An emotion that is caused by X	Motivates	verb [feels, wants, likes] object [curiosity, independence, competition, honor, approval, power, status, romance, success, friendship, belonging, health, safety, livelihood, happy, stressed, angered, disgusted, sad, surprised, fearful, trusting, love, obedient, amazed, disappointment, regret, worthless, aggression, optimistic]
dim 8 A change of location that X results in	Enables	verb [am is are] preposition [above, across from, at, below, far from, in, in front of, inside of,near, next to, on top of, outside of]
dim 9 A change of possession that X results in	Enables	verb [possesses]
dim 10 Other change in property (besides location, emotional state, or possession) that X results in	Enables	verb [am, is, are, has, have, want, wants, need, needs]

Table 1: This table lists each dimension that workers were asked about for each sentence/story pair. Some dimensions contained drop-down choices.

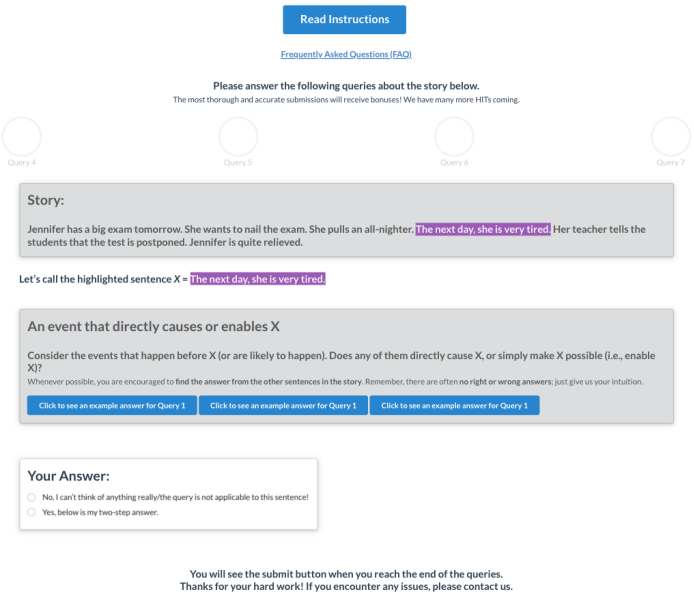

Figure 4: Main UI

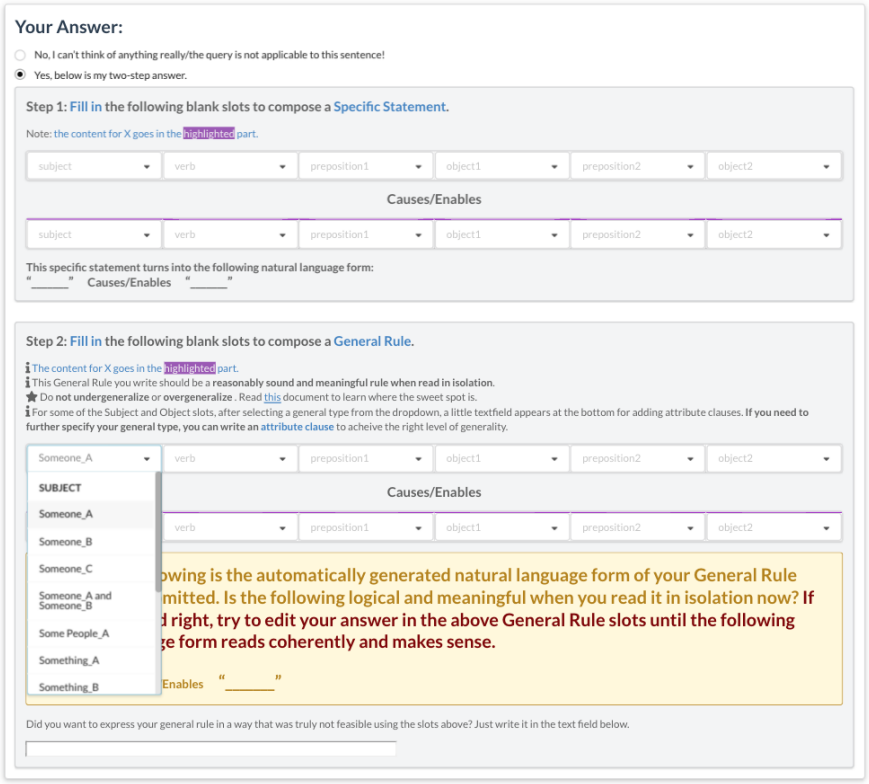

Figure 5: When "Yes" is selected for "Your Answer:" on the main UI, workers can input answers in slots

7 shows a state of the percentage of contributions to the data by worker score.

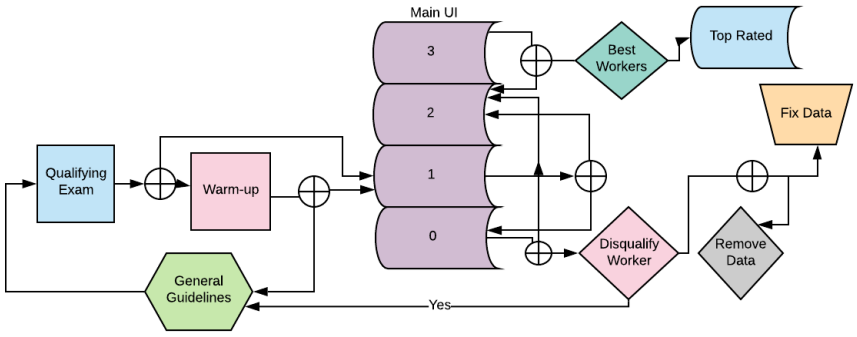

Figure 6: Data collection pipeline. Numeric values in the Main-UI (lavender) correspond to worker ratings.

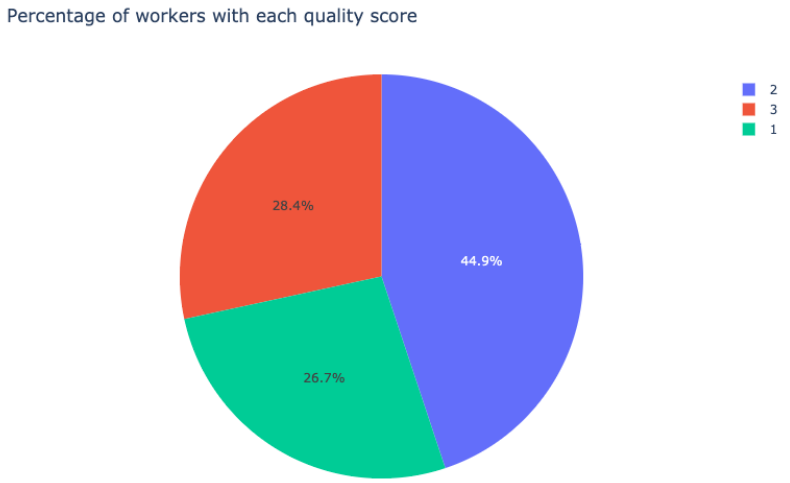

Figure 7: Workers by quality score rating, Red = '3'(high), Purple = '2'(mid), Green = '1'(lower)

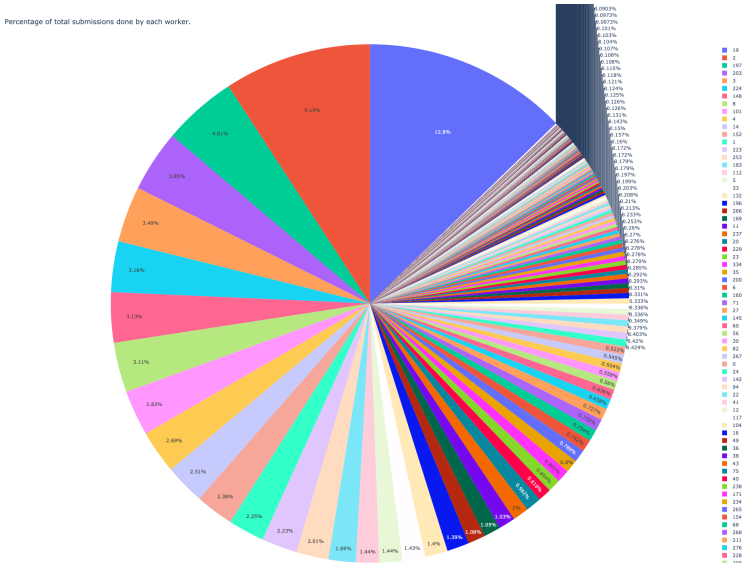

Figure 8: The data was annotated by a diverse set of workers with no one worker doing more than 12.8% of the data

3.4.1 Review Dashboard

Incoming submissions to the large batches were monitored daily through the in-house review dashboard for quality control. The dashboard is shown in Figure 9.

Submissions displayed through the review UI could be filtered by batch, submission status (e.g., Submitted or Approved), worker identification number, date of submission, story identification number, dimension, and sentence. When work in need of approval was displayed, it showed the worker's current quality rating and any notes on the worker's history.

These ratings were maintained in a spreadsheet that was uploaded to the UI as changes were made.

3.4.2 Worker Records File

4 Preparation and Analysis of the Final Data Set

After we completed the collection task, we did additional post-processing of the data. This section discusses that process and the results of the final data.

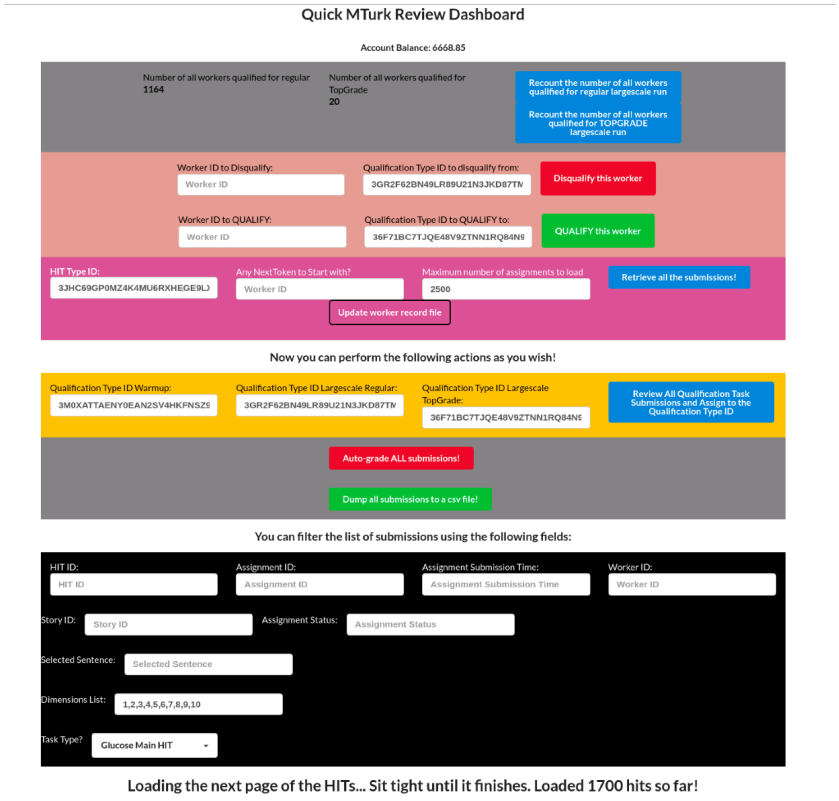

Figure 9: Data review dashboard allows reviewers to see user responses, sorted by desired parameters, and give feedback to workers

4.1 Data Post-Processing

In order to get more granularity in the final worker quality ratings, we coordinated the commit date of worker quality rating changes with the submission date of the HITs. This increased granularity improved out quality level to showing 80% very high quality answers when filtering a random 50 samples from workers rated as 3.

To clean the data even more, we got rid of many ratings used in the initial classification. Of the ratings '3', '2', '1.5', '1', '0', '-1', and '-2', we kept only '3', '2', and '1'. Ratings were changed with the following policies: For workers who contributed more than 200 HITs to the data ranked as '3' (20 workers total), we reviewed the quality of their work by rating and moved the ratings, as appropriate. For instance, if a review of their '3' rated work showed it to be of lower quality than expected, we included a policy in a cleaning script for moving all of their '3'

rated work to '2' rated work. For the ratings we were not using, '1.5', '0', '-1', and '-2', we review the set of work and moved it by scripted policy to the appropriate rating. For workers with 100-200 HITs in the '3' rating, we checked only their '3' rated work and demoted it, if necessary. For the remainder of workers, we created a policy for moving their '-2' ratings to '2'. This choice was because the '-2' rating was used when workers were going too fast or not filling out more than 1 or 2 dimensions. The content of their mini-theories was not bad, but rather they weren't contributing as many mini-theories per HIT. For the remaining scores, '1.5', '0', and '-1', the scores were moved to '1'. After running the policy script, we reviewed the data again and found much better consistency across ratings. A random sample of 100 answers ranked as '3' showed 92% of the data to be up to the highest quality standard.

4.2 Data Quality Results Analysis

In the total data, workers filled in an average of 4.7 dimensions, with the median being 4 dimensions.

Figure 10 shows the distribution of dimensions filled out by type. We expected dimensions 5 and 10 to be less frequent (Is there anything else that causes or enables the sentence?), due to the fact that they refer to rare changes of state, such as becoming wet. We saw some bias towards answering more of the first questions than the later ones, offset by an overall preference for dimensions 1 and 6, the causality dimensions.

Dimensions 8 and 9 were less frequent than their counterparts 3 and 4. The main reason is because a location state is often part of the scene setting or background of a story, whereas a location state that is the result of the highlighted sentence only occurs when a change of location occurs. Similarly, there are things possessed at the outset of a story, as part of the scene it occurs in, that enable things that are talked about in the story, however, for a sentence to lead to a possession state, it needs to be the case that something is acquired.

Of the final data, we used a portion of the data ranked as quality level '3' to construct a test set. This test set was judged by a group of human annotators and determined to be 'correct' or 'nearly correct' for all dimensions (see [ACL citation]). We consider the test data to be representative of the larger data set in terms of which dimensions are relevant. Figure 11 shows the percentage of each dimension filled out in the test data.

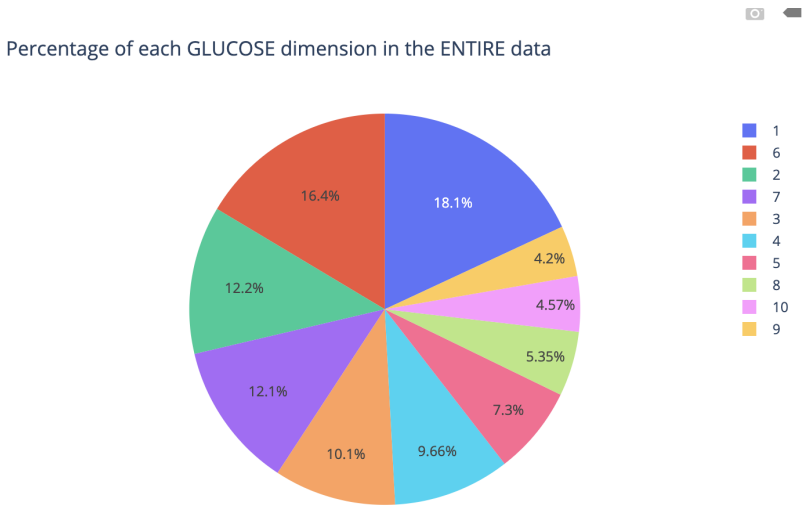

Figure 10: Data collected by dimension in the total data

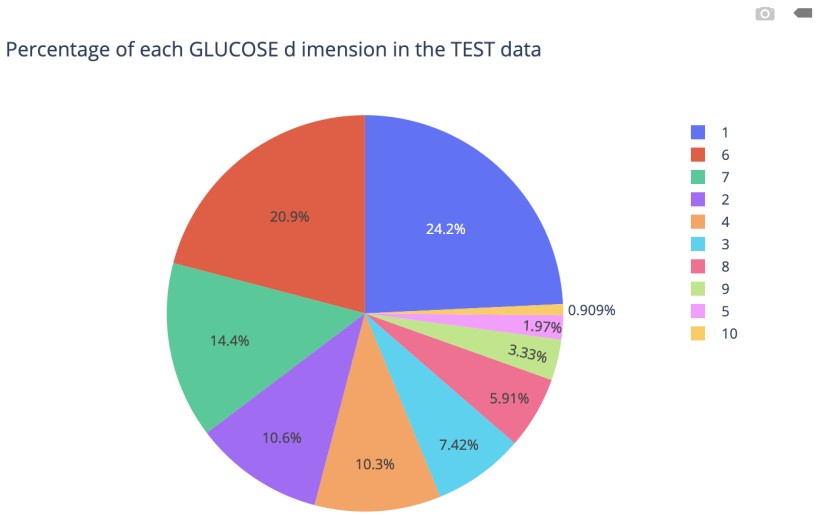

Figure 11: Data collected by dimension in the test data

5 Conclusions

The GLuCOSE task is a large-scale effort to collect common sense mini-theories from crowd workers. We successfully collected rules that prove to be useful in applications such as model training ([cite ACL]).

Using Amazon Mechanical Turk requires breaking tasks down into HITs (Human Intelligence Tasks), and the interface has an economy all its own. There is an expected range of the amount of work per HIT and an expected pay rates for that work. Breaking a large task like GLuCOSE into appropriate tasks required numerous iterations. Having a visually-appealing UI was essential for workers doing this task. Having slots to help with forming mini-theories kept the data in a format that made it easier for applications and provided guidance to workers. Inevitably, this task involved a lot of training for workers and for us as requesters.

6 Acknowledgements

This work could not have been completed without the group of Mturk workers who contributed their time and effort. We are thankful for their patience as we iterated the task instructions and quality control.