Data Collection and Worker Quality Control for Generalized and Contextualized Story Explanations (GLuCOSE)

Nasrin Mostafazadeh & Lori Moon
Elemental Cognition
crowd@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 mini-theories about the world based on the sentence/story pair. These mini-theories are in the form of semi-structured inference rules. There are 20 rules solicited per sentence, guided by ten dimensions of common sense causal explanation with one mini-theory for each dimension being specific to the story and another mini-theory generalized from the specific mini-theory to other situations in the world.

We collected a total of 337,636 pairs of specific and general rules. Of the rules, 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.

This document describes the task, the data collection process, the work quality management, and analysis of the data collection task. The data download is available at https://github.com/ElementalCognition/glucose. For results from using the GLuCOSE rules in model training see [ACL 2020 paper].

2 GLuCOSE Task

The GLuCOSE Task involves collecting common sense reasoning that is necessary for understanding stories. There are many dimensions to the common sense rea-

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

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] 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.[2]

The task was set up as follows, if the story in the task is:

> 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 apply before or after that sentence. Let us take the second sentence as an example:

> I have a ton of pictures to show you.

Workers are asked first about the things that occurred before the sentence. They are asked if there is anything in the story that causes or enables the events in the target sentence. If there is something relevant, then they write a specific statement about it. They are given slots for words to help them reduce the sentences, and the connective, in this case, 'Causes/Enables' is supplied:

> My daughter was born Causes/Enables I have pictures to show you

Next, the worker is asked to make a general statement about the world based on the specific statement that they wrote, for example:

> Someone_A has a baby Causes/Enables Someone_A shows Someone_B pictures of the baby

[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.

[2] 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.

Workers had a drop-down menu of expressions such as 'Someone_A' to help them generalize. Each pair of rules like this is a specific and general rule.

For the next dimension, workers were asked, about the same sentence, if there were any basic human desires or motivations that caused events the sentence, in this case 'I have a ton of pictures to show you'. The worker had more prompts for this response. For example, they were offered a drop-down list of verbs, shown in table1. A worker might well respond, 'Someone_A feels love Motivates Someone_A shows Someone_B pictures of Someone_C (who Someone_A loves). This process is repeated for locations, (Is there a location that enables the events in the sentence?), for possessions, and for any other attribute.

After all five dimensions are considered, the order is reversed and they are repeated. Dimension six asks if the events of the sentence under consideration cause or enable anything to follow. An example is below:

> I have pictures to show you Causes/Enables You see pictures

And a general statement could be:

> Someone_A shows Someone_B Something_A Causes/Enables Someone_B sees Something_A

This continues for the other four dimensions in this direction. So each of ten dimensions total are solicited 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 and the general statement about it.

3 Data Collection

Data was collected through an in-house user interface, hosted by Amazon Mechanical Turk.

3.1 Participants

A pool of 1039 workers qualified for our main task via our in-house Qualifying Exam. Of the workers, 373 contributed to the main task. The worker pool was constrained by Mturk-internal ratings (e.g., worker has a high acceptance rate, worker has done at least 100 HITs on Mturk), country codes, and by scores on our qualification task.[3] We did not limit how many times workers could take the test, however,

[3]Country codes were originally restricted to a standard list containing the US, the UK, New Zealand, Australia, South Africa, and Canada, but we opened it up to many other countries per worker

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.

they could only take it once per release, where each release had 1,500 HITs. In order to do well on the test, users were encouraged to carefully read instructions and the examples embedded in the test itself and referred to a shared document that contained general guidelines for the task.[4]

3.2 Qualification Content

The qualification test contained questions testing expertise in three areas: Identifying correct use of the UI slots for language expressions (see Figure 1), recognizing the right level of generalization (see Figure 2), and identifying causes and effects with proper temporal understanding of the stories (see Figure 3).

For the slot use, prospective workers were given correct slot examples and incorrect slot use examples and asked to choose the appropriate ones. The slots were simplified from part of speech tags to allow for easier understanding and to encourage simplification of the language of the stories.

For generalization understanding, prospective workers were presented with a Specific Statement based on the target story. They then 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 understanding that, if something is too general, it does not make sense as a rule.

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.

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.

---

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

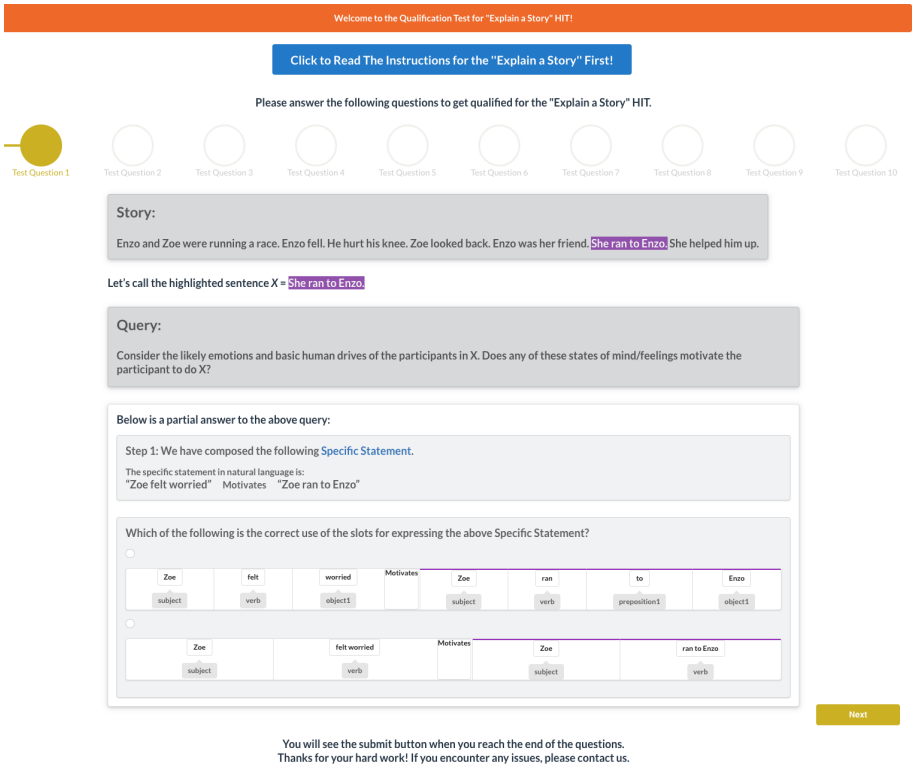

Figure 1: Qualification question about slot use

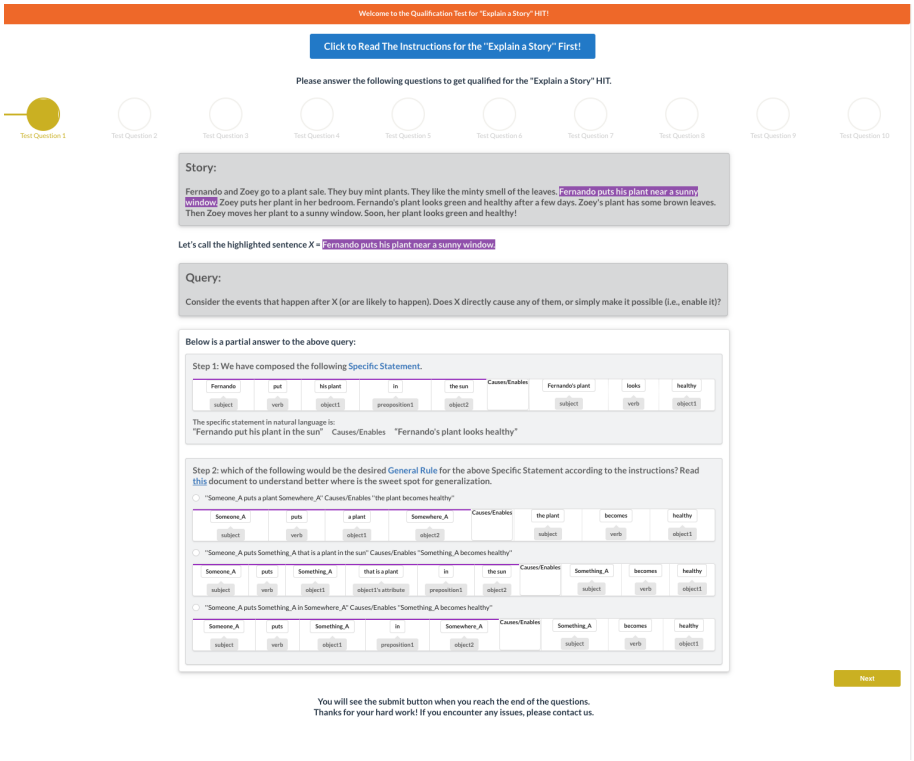

Figure 2: Sample qualification question about the right level of generalization

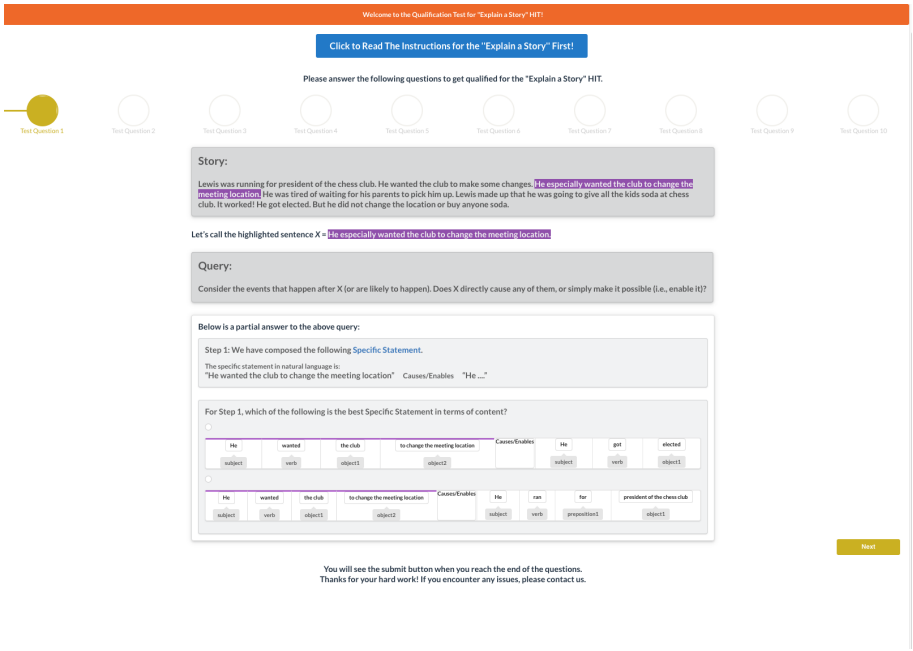

Figure 3: Sample qualification question about what events follow

3.3 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.

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.

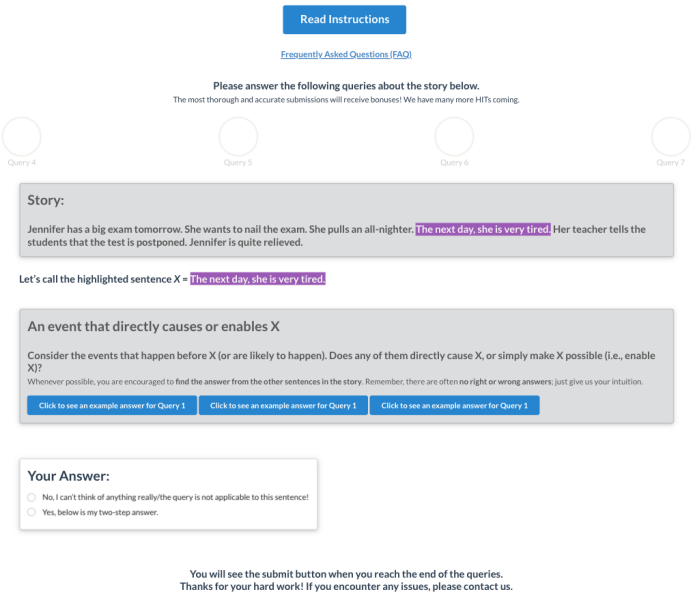

Figure 4: Main UI

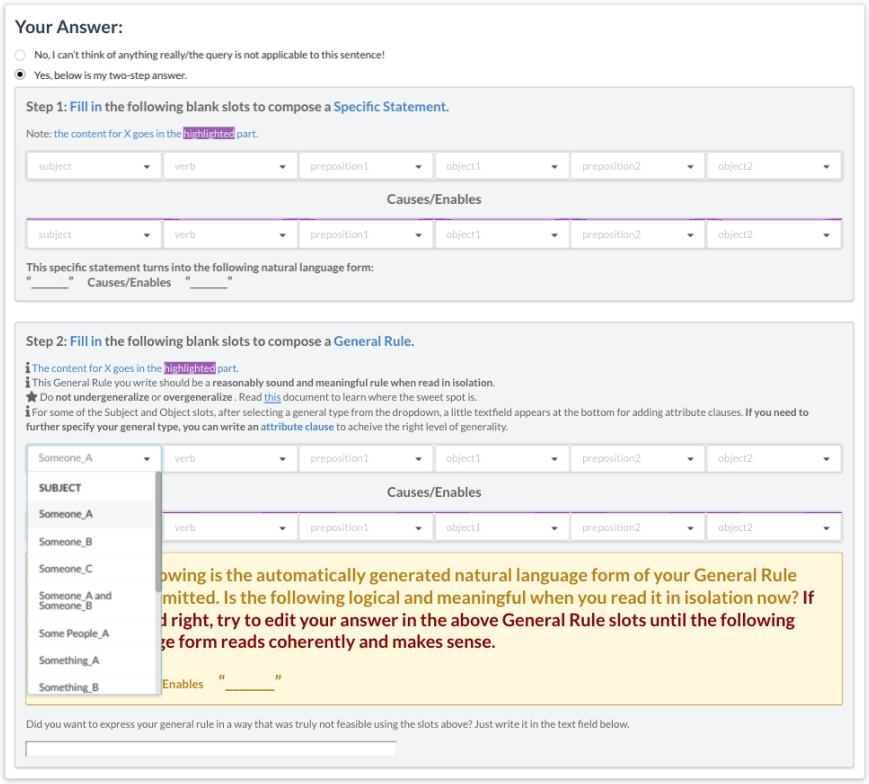

Figure 5: When "Yes" is selected for "Your Answer:" on the main UI, workers can input answers in slots

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

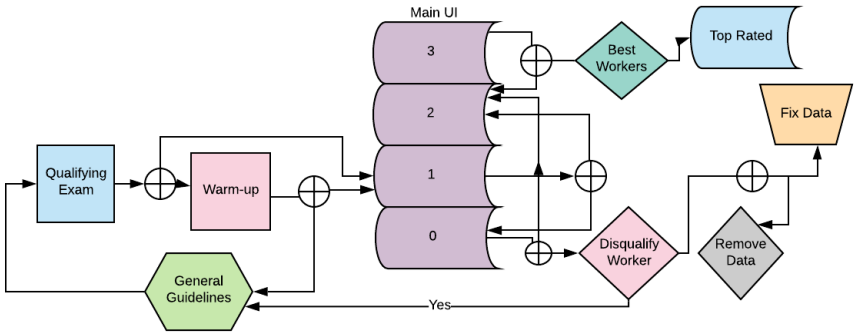

Figure 6: Data collection pipeline. Numeric values in the Main-UI (lavender) correspond to worker ratings.

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.

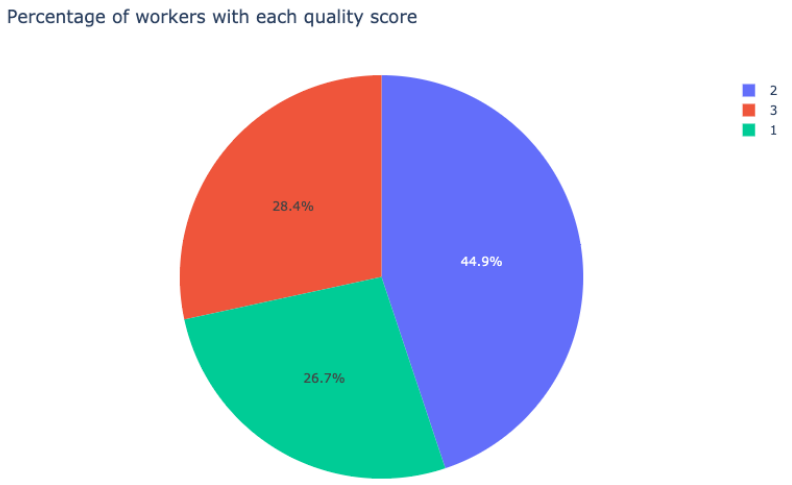

Figure 7: Workers by quality score rating, Red = '3'(high), Purple = '2'(mid), Green = '1'(lower)

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.

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.

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.

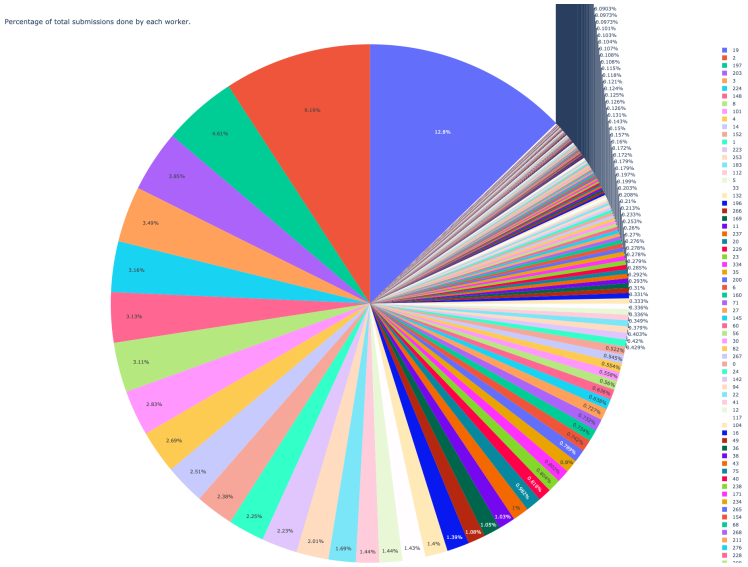

Figure 8: The data was annotated by a diverse set of workers with no one worker doing more than 12.8% of the data

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.

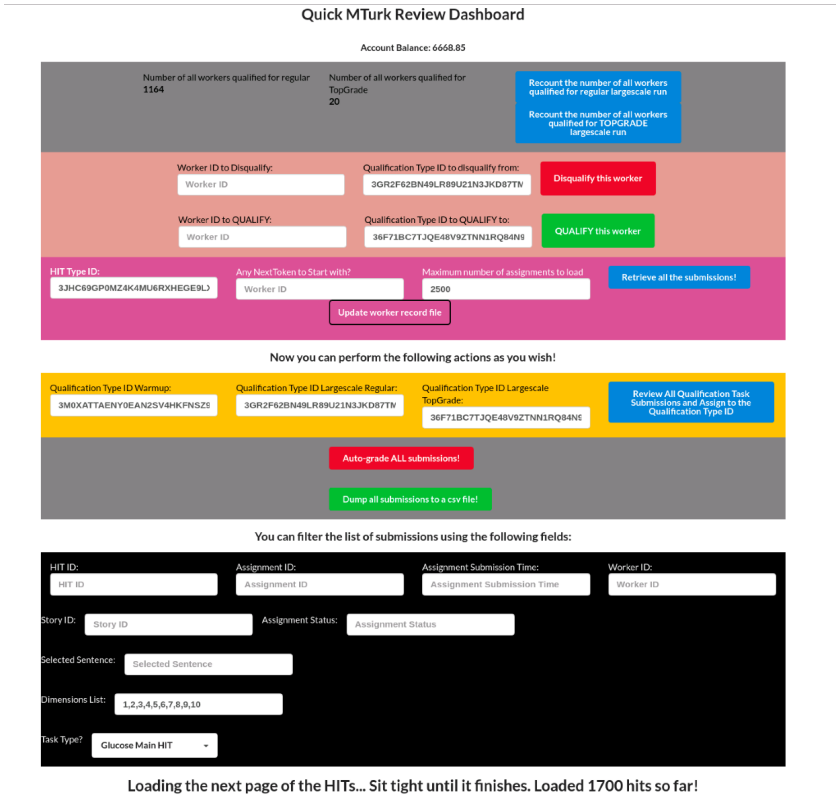

Figure 9: Data review dashboard allows reviewers to see user responses, sorted by desired parameters, and give feedback to workers

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

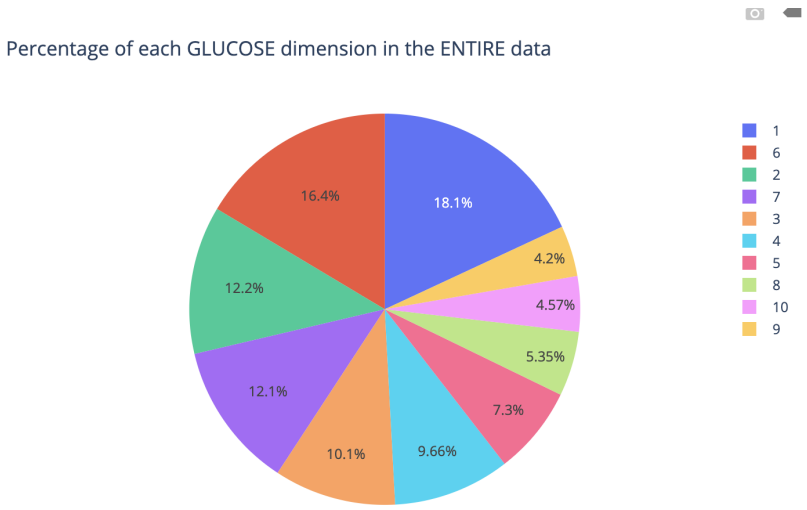

Figure 10: Data collected by dimension in the total data

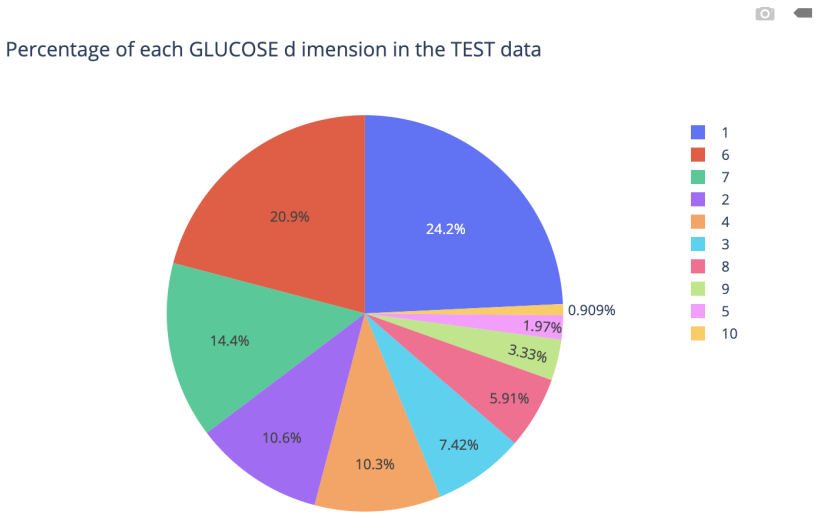

Figure 11: Data collected by dimension in the test data

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.

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.