Using Mechanical Turk to Automate Subjective Annotations in the Context of Rapport

Rae Lasko
Carnegie Mellon University
Pittsburgh, Pennsylvania. United States
rlasko@cmu.edu

Abstract
While very little literature currently exists on using Amazon Mechanical Turk (AMT) to crowd-source subjective annotations specifically in the context of rapport, there is significantly more research on subjective annotations in general, subjective annotations in rapport, and crowdsourcing subjective annotations in general. A large portion of the research on crowdsourcing subjective annotations involves non-verbal behavior, semantic categorization, and detecting sentiments through text. In order to discuss subjective annotations specifically in the context of rapport, I'll use a three-pronged approach involving the aforementioned research areas. The goal of this paper will be to discuss using AMT for subjective annotations in the context of rapport, and also present methods to increase inter-rater reliability and overall data quality.

Author Keyword
Mechanical Turk; subjective annotations; annotations; crowdsourcing, micro-tasks; inter-rater reliability, thin-slicing; rapport ratings

ACM Classification Keywords
H1.2 [User/Machine Systems]: Human Factors

Introduction
The rise of Amazon Mechanical Turk over the past few years has provided an opportunity for researchers to take advantage of cheap and quickly available labor from all over the world. While at the moment it's mainly used for tasks with an unambiguous answer, the service has shown promise in more subjective areas. The use of AMT for subjective annotations in research pertaining to rapport is especially of interest because while in-house annotations can take hundreds of person-hours, the crowdsourcing platform provides easy and quick access to as many workers as needed. But while the service is easily implemented for tasks with a reasonably obvious correct answer, it's reliability and scalability is less tested in more subjective areas. The service has shown promise in these areas, though concerns still remain about achieving inter-rater reliability and detecting and circumventing cheating from workers.

General Subjective Annotation
In social and behavior psychology, there is often a need to annotate qualitative data. In order to achieve inter-rater reliability, raters are often involved in training until agreement on sample data sets reaches a certain pre-prescribed threshold [9, 11].

Annotating in the Context of Rapport
Rapport is a quality of relationship that is present with increased empathy, attention, and understanding. [4] Since rapport is generally considered the result of a combination of qualities that must be observed through verbal and non-verbal behavior, research studying rapport generally involves studying these qualities separately and together. Non-verbal behavior, such as gaze-aways, head nods, and leaning are considered along side conversational strategies such as indirect deliver and delivery style. [13, 4]

Annotating Rapport Strategies
To annotate non-verbal and verbal behavior, annotators must achieve a certain degree of agreement before continuing. This generally consists of annotating a training set and meeting to discuss the results until a predetermined threshold is reached. This process to annotate for rapport-building strategies is very similar to processes used in other subjective annotations.

Thin-slicing

Design

In contrast, thin-slicing involves showing naïve, untrained, raters short (<5 min) slices of an interaction. Audio is generally included but not always. The raters are generally given a short definition of the target interaction, but no further training. This increases the subjectivity and may decrease the consistency of the ratings [1, 7].

Thin-Slicing Overview

The use of thin-slicing in rapport relies on the theory that people are able to form extremely accurate judgments based on short exposure to behavioral samples. A meta-analysis conducted by researchers at Harvard University of a broad range of studies found that there was a statistically significant accuracy and consistency across raters. All studies that were considered eligible consisted of videos of no more than 300 seconds and no judge rated more than 600 seconds of video. This ensured that judgments were formed based on a small exposure and not over a larger sample. They also found that increasing the length of the sample videos or adding audio did not correlate with increased accuracy. Overall, they concluded that affective and interpersonal dynamics can be judged extremely accurately and efficiently [1].

Analysis

Depending on the purpose of the thin-slicing, the results are analyzed in different ways. In some cases, the results are used on a slice to slice basis to quickly judge people's personality or detect deception. More often, the results are taken together and used as a baseline [1]. Methods of aggregating data range from taking weighed ratings based on raters' reliability through algorithmic means, taking the average, and using the population minus outliers [1, 11, 7].

Mechanical Turk Overview

Amazon Mechanical Turk is an online labor marked that allows requesters to upload HITs (Human Intelligence Tasks) for workers (often referred to as Turkers) to complete. The HITs are generally very short and fairly simple in nature, but too complex to automate completely. Requesters are able to set a price per HIT, determine the required qualifications for workers, and approve or reject the submissions made by workers. After the data is collected, requesters are also able to provide bonuses for exemplary work, increasing their popularity among Turkers. Furthermore, AMT is designed to allow requesters to make one template HIT with variables and then create batches by uploading a CSV file with values for that variable(s). This makes it extremely easy to set up bulk annotations of videos, images, or transcripts. Requesters can also determine the number of assignments for each HIT. This determines the number of unique Turkers who will complete each HIT, allowing multiple workers to complete each individual task.

Crowdsourcing Subjective Annotations

When crowdsourcing subjective annotations, one of the primary concerns that researchers have is inter-rater reliability. Most studies approach this issue in multiple ways. Part of ensuring agreement begins before the HITs are even published to AMT by setting a high requirement for the percentage of HITs that a worker has been approved for. Researchers may include a small proportion of gold-standard data in the HITs. With this method, researchers are able to determine which Turkers are more accurate and are able to weight their ratings higher if aggregating data across HITs (such as in thin-slicing). Researchers also often set a high number of assignments so that multiple unique Turkers complete the same task for more post-processing flexibility [3, 8].

Other methods can include filtering the submissions by the time taken to complete them. For example, if the HIT requires the Turkers to watch a 30 second video and answer a few questions, we can reasonably say that any worker who took less than 35 seconds to launch the page, watch the video, answer the questions, and submit the results either didn't watch the video completely, or didn't pause to answer the questions carefully. In this case, their submission can be rejected.

In post-processing of the results, an important step is aggregating the data. If the researchers specified multiple assignments per HIT, there are multiple ways to process the data. Some studies remove outliers and take the average of the remaining, while others use more complex algorithms.

One paper that dealt only with processing collective annotations used social choice theory, "the systematic study of methods of aggregating information provided by individuals into a collective view of that information". Instead of only using only maximum likelihood estimation like most other studies, this study presented an approach that accounted for annotator bias and other noise in the data set. The study found their methods were more accurate than using only maximum likelihood estimations could provide, but cautioned that the methodology would need to be adapted to the specific dataset at hand if used in other studies [7].

Another study asked workers to make pairwise comparison of images and suggested that with inherently subjective tasks, outlier detection can fail since minority votes are labeled outliers and eliminated. Instead, they suggested that if local rankings are integrated in a global ranking, outlier detection is much more accurate and effective, since it is able to remove outliers that are locally consistent but cause global inconsistency [5].

Accuracy of crowdsourced subjective annotations

In natural language processing

Natural language processing is one of the most popular areas of research into using AMT for subjective annotations. One study by researchers at Stanford tested the accuracy of using AMT for affect recognition, word similarity, recognizing textual entailment, event temporal ordering, and word sense disambiguation. For each task, they compared the gold standard data set to the non-expert annotations from AMT and found that an average of 4 non-expert labels were needed to reach expert level labeling. Overall they concluded that the crowdsourced labels were sufficiently accurate compared to the gold standard, with most of the tasks having an agreement of 90% or higher [12].

Another NLP study used AMT to label quote and response text pairs as sarcastic, ambiguous, or not sarcastic. They then compared the results of the AMT data collection to a gold standard data set, using several different measures of weighting the Turkers' responses. Interestingly, in contrast to the previous two studies discussed here, they found that simple majority voting was more accurate than more complex methods of weighting the Turkers' responses.

Sentiment analysis has also been conducted extensively on AMT. In a study from IBM, researchers used Mechanical Turk workers to first identify snippets of the tweets relevant to a politician, classify the snippet as Positive, Negative, Both, or Neutral, determine if the content was subjective or not, and then if the snippet was written to support or oppose the named political candidate. The results were then compared to on-site expert annotators. While the researchers found a significant amount of noise in the crowdsourced data set, the researchers concluded that there was reasonable accuracy in using AMT for sentiment classification, and after eliminating the noisiest Turkers, the accuracy of the data set improved. Consistent with previous studies, they concluded that using multiple Turkers to annotate each tweet was essential in reaching accurate results [6].

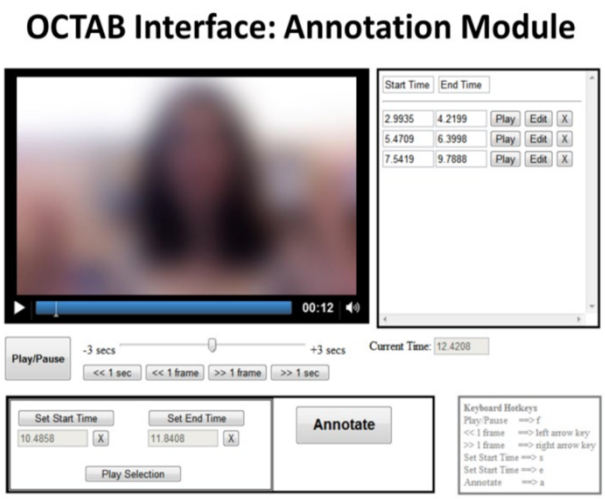

Figure 1. OCTAB interface for micro-level annotations using Amazon Mechanical Turk [10]

Annotating non-verbal behavior

To augment the tools that Amazon provides as part of Mechanical Turk, a group of researchers at the University of Southern California developed a specialized interface, OCTAB, to crowd source micro-level behavioral annotations (gaze-aways, headshakes, frowns, and pause fillers). Designed to be more task specific than the generic templates provided, OCTAB (Figure 1) includes video controls that are specific to its tasks (instead of forward/back by 10 or 30 seconds, it's fine-tuned to 1 second). The interface includes a fairly easy to use method of inputting start and stop times, and displays all of the annotations in one place. Less complicated than specialized tools used by experts in research, it's tailored for the use of the non-expert workers one finds on a service such as AMT.

Unlike the previously presented studies, this one included a training phase for all workers. The training modules were formulated after observing local coders reaching agreement, and focused on giving the annotators a visualization to compare their own annotations with ground-truth annotations (a bar-graph over a timeline), and quickly review all the ground-truth annotations and their own side by side. The researchers suggest using an iterative approach to training workers, whereby the workers annotate a single video and immediately receive feedback until the agreement is satisfactory.

The researchers then compared the annotations results of trained local annotators (2), untrained crowd-workers, and trained crowd-workers. They then also compared between crowdsourced majority (the majority voting of 3 raters) and

crowdsourced unique (annotation from 1 rater) annotations for trained and untrained workers. The most surprising result of the study was that agreement between the two local expert annotators was actually lower than the agreement between the trained crowdsourced majority and each of the individual local annotators for all behaviors. Furthermore, the trained crowdsourced unique was found to be sufficiently in agreement for two of the target behaviors (gaze-away and pause filler).

When comparing local annotators with untrained workers, the study found that while the agreement was lower than when compared with trained workers, the desired threshold was still reached in several of the behaviors. Notably, for both crowdsourced majority and crowdsourced unique for untrained workers, the threshold was met for gaze-away behavior. For other behaviors, the crowdsourced majority barely missed the agreement threshold with local annotators. The lowest agreement was with the headshake behavior, but researchers noted that local annotators also had trouble reaching agreement.

Overall the researchers found a statistically significant increase in annotation agreement with trained crowd workers as compared to untrained workers. They also suggest that when using trained workers, the necessity of multiple assignments of the same task decreases and may even become unnecessary. This is in notable contrast to the previously discussed NLP studies which all strongly asserted that multiple assignments are a must. However, this could be due to the fact that the behaviors coded for in this study involving OCTAB were somewhat less subjective than behaviors in the previous studies. [10]

More relatable to rapport, a study on analyzing the impression of vloggers used AMT workers to annotate for personality, attractiveness, and mood impressions. The HITs asked the workers to watch the one-minute video segments and complete questionnaires on the vlogger's personality, attractiveness, and mood, as well as the worker's demographic. All annotations of impressions of the vloggers were utilized the 7-point Likert scale. Inter-rater agreement ranged from .25 to .77. The highest inter-rater agreement was found when rating extraversion, happiness, and overall mood, while the lowest were found for nervous, smart, and emotional. The researchers were not sure why the agreement varied between these ratings but interestingly, they found statistically significant differences between annotations from males and females for ratings of personality (female raters tended to give higher scores). The researchers concluded that crowdsourcing can be used to collect interpersonal impressions, specifically in the context of vloggers, however more work needs to be done in how the data is aggregated is used. The main limitation of this study is while they reported inter-rater agreeability between crowdsource workers, the results weren't compared to a local dataset. As a result, the validity of they annotations they collected as compared to annotations from local annotators is uncertain [2].

Conclusion and Future Directions

The studies discussed here strongly suggest that crowdsourcing is a reliable way of collecting subjective annotations. The research in using AMT for natural language processing relating to sentiment analysis, affect recognition, and word sense disambiguation indicate that subjective annotations from crowdsourced workers are sufficiently accurate, especially when combined with the correct data aggregation methods. Micro-level behavioral annotations further demonstrate that workers are able to notice minute details, especially when trained. Since rapport combines elements of the verbal behavior annotated for in NLP studies, and non-verbal behavior as demonstrated in the OCTAB study, a case can be made for using crowdsourcing for subjective annotations in the context of rapport, especially considering the relatively high inter-rater agreement found for interpersonal impressions in the last study. Moreover, for thin-slicing ratings, where researchers need a gut level reaction, AMT shows significant promise, though thin-slicing specific research should be done. Furthermore, more work needs to be done in the processing results from multiple assignments of the same HITs to gain even more inter-rater reliability and accuracy. This problem is especially evident in thin-slicing where naïve raters are used and inconsistent results expected, as simple outlier removal may dispose of wanted data. As for more specific timeline based annotations in rapport, such as conversational strategies and rapport building techniques, much more work needs to be done in worker reliability and accuracy. In particular, the large range of inter-rater agreement was unaccounted for, but the impressions that were being annotated are key parts of establishing rapport. The OCTAB study further suggests that using trained workers may be a feasible way to increase accuracy while decreasing the number of duplicate assignments (and cost). Though the studies here demonstrate that many of the parts of rapport can be accurately coded for, that doesn't necessarily mean that the sum of the parts will have the same results. With more research, the promise that Amazon Mechanical Turk

has already demonstrated can be verified and used to facilitate subjective annotations in the context of rapport.

References

[1] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychological bulletin, 111(2):256, 1992.

[2] Biel, J. I. and Gatica-Perez, D. The good, the bad, and the angry: Analyzing crowdsourced impressions of vloggers. Proc. ICWSM 2012, 407-410.

[3] C. Callison-Burch and M. Dredze. Creating Speech and Language Data with Amazon's Mechanical Turk

[4] G. 4. Rapport: Definitions and Dimensions. Advances in Consumer Research, 1984.

[5] Y. Fu, T.M. Hospedales, T. Xiang, J. Xiong, S. Gong, Y. Wang, and Y. Yao. Robust Subjective Visual Property Prediction from Crowdsourced Pairwise Labels

[6] P. Hsueh, P. Melville, V. Sindhwani. Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria

[7] J. Kruger, U. Endriss, R. Fernandez, and C. Qing. Axiomatic analysis of aggregation methods for collective annotation. In Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems

[8] W. Mason and S. Suri. Conducting behavioral research on Amazon's Mechanical Turk Behavior Research Methods 44, 1 (2012), 1-23.

[9] J.F. Marques and C. McCall. The Application of Interrater Reliability as a Solidification Instrument in a Phenomenological Study

[10] S. Park, P. Shoemark, and L. Morency. Toward Crowdsourcing Micro-Level Behavior Annotations: The Challenges of Interface, Training, and Generalization

[11] T. Sinha and J. Cassell. We Click, We Align, We Learn: Impact of Influence and Convergence Processes on Student Learning and Rapport Building.

[12] R. Snow, B. O'Connor, D. Jurafsky, A.Y. Ng. Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks

[13] R. Swanson, S. Lukin, L. Eisenberg, T. Chase Corcoran and M. A. Walker. Getting Reliable Annotations for Sarcasm in Online Dialogues

[14] Z. Yu, D. Gerritsen, A. Ogan, A.W. Black, J. Cassell. Automatic Prediction of Friendship via Multi-model Dyadic Features