USING THE AMAZON MECHANICAL TURK FOR TRANSCRIPTION OF SPOKEN LANGUAGE

Matthew Marge, Satanjeev Banerjee, and Alexander I. Rudnicky

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{mrmarge, banerjee, air}@cs.cmu.edu

ABSTRACT

We investigate whether Amazon's Mechanical Turk (MTurk) service can be used as a reliable method for transcription of spoken language data. Utterances with varying speaker demographics (native and non-native English, male and female) were posted on the MTurk marketplace together with standard transcription guidelines. Transcriptions were compared against transcriptions carefully prepared in-house through conventional (manual) means. We found that transcriptions from MTurk workers were generally quite accurate. Further, when transcripts for the same utterance produced by multiple workers were combined using the ROVER voting scheme, the accuracy of the combined transcript rivaled that observed for conventional transcription methods. We also found that accuracy is not particularly sensitive to payment amount, implying that high quality results can be obtained at a fraction of the cost and turnaround time of conventional methods.

Index Terms— crowd sourcing, speech transcription

1. INTRODUCTION

Accurate manual speech transcriptions are crucial for nearly all aspects of spoken language research (speech recognition, speech synthesis, etc.). However, manual speech transcription is a demanding task. Conventional transcription methodology requires training transcribers on transcription guidelines, which last from hours to days. Once trained, a worker is expected to be available to transcribe sizable amounts of spoken data. Recruiting, training and retaining transcribers represents a sizable endeavor even for university-based efforts that can make use of undergraduate labor. Costs can run upwards of $100 per hour of transcribed speech (assuming the transcriber is being paid $10 per work hour and that transcription and checking takes 10 times the length of the audio being transcribed).

Amazon's Mechanical Turk[1] (MTurk) can potentially reduce the cost of manual speech transcription. Dubbed "artificial artificial intelligence", MTurk is a marketplace for human workers ("turkers") who perform tasks for pay; these tasks are submitted and paid for by other humans ("requesters"). The requester determines when work is satisfactory. The tasks targeted by the system are ones that are simple for humans to perform but are challenging for computers (e.g., determining if a person is facing to the right or left in a picture). Speech transcription fits well into this framework. In this paper, we investigate whether MTurk can be used to perform transcription according to the conventions and quality level expected by the speech research community.

MTurk has been previously used by others to transcribe speech. For example, [1] and [2] report "near-expert accuracy" when using MTurk to correct the output of an automatic speech recognizer. MTurk has been used for other natural language annotation tasks. For example, [3] used MTurk to carry out several different annotation tasks (such as word sense disambiguation) and found strong agreement with gold standard annotations. [4] used MTurk to evaluate machine translation output. [5] evaluated MTurk as a venue for conducting user studies. While these results show that MTurk appears useful for language annotation tasks, to our knowledge a systematic study of MTurk as a means for general speech transcription has not yet been reported. This paper presents such a study.

2. THE MTURK RESOURCE

Requesters post Human Intelligence Tasks (HITs) on Mechanical Turk by uploading tasks onto Amazon's web portal. MTurk maintains each turker's performance history, which requesters may use to specify who is eligible to perform the HITs. Eligibility may include the turker's location (country), HIT completion rate (fraction of tasks completed among those he signed up for in the past) and approval rate (fraction of tasks accepted by requesters among those he completed in the past). The requester must also specify the amount of payment a turker can receive if he completes the task and his work is accepted by the requester. Once a requester posts a HIT on MTurk, as in our transcription task, eligible turkers can immediately view the HIT, and sign up for it. For the transcription task, we restricted eligibility to turkers with a previous HIT approval rate of 95% or better.

[1] http://www.mturk.com

All audio was posted on a third-party website[2] that provides streaming audio through a widget embedded in our transcription HIT. Streaming makes it difficult for turkers to download or otherwise manipulate the audio in unanticipated ways. Each transcription HIT posted on MTurk was associated with a single unique streaming audio file. In our case, each transcription HIT could be performed by multiple turkers. MTurk removes completed tasks from a turker's list of available HITs, precluding a single turker from transcribing the same audio file twice. Once a turker completes the HIT, the requester can download and review the work (in our case, the speech transcription) and provide payment via Amazon if the work is satisfactory.

3. EXPERIMENTS WITH MECHANICAL TURK

This study had turkers perform a simple transcription task. We examined the influence of payment amount, and of data characteristics, including speaker gender and speaker background (native or non-native English speaker) on the accuracy of the transcription.

3.1. Procedure

The same transcription task was offered to different groups of turkers by posting, at any given time, one "batch" (a preset group of transcriptions that need to be transcribed at a specified payment amount). There were four batches, each paying \$0.005, \$0.01, \$0.03, or \$0.05 per transcription, pending approval by a requester. For any batch, five different transcribers were allowed to transcribe each utterance. While MTurk automatically prevented a single turker from transcribing the same utterance twice within a single payment batch, we also ensured that no turker transcribed the same utterance from two or more different payment batches. We did so in order to avoid any learning effects that turkers may have from transcribing the utterance the first time. We implemented this restriction by checking the identities of the turkers, and reposting the utterances that were transcribed more than once by the same turkers.

Occasionally, turkers submitted incomplete or incoherent work, and this work was rejected via the MTurk web interface. We rejected 16 assignments out of the total 910 assignments (260 in three payment conditions, 130 in one payment condition). Once all transcriptions were done, we computed word-level accuracy against an in-house gold standard transcription, as described in section 5. We recruited 94 unique MTurk workers for this study.

3.2. Transcription Instructions

Turkers were told to listen to utterances by using an audio player embedded in the task web page. Turkers were asked to transcribe every audible word, and were told to follow conventions for marking speaker mispronunciations and restarts; they were not required to mark fillers (e.g., "uh" and "eh"). The instructions stressed the importance of accuracy. Turkers were allowed to replay the audio as many times as necessary to produce a satisfactory transcript.

4. AUDIO MATERIALS

The audio clips were taken from data collected in a separate experiment investigating the formulation of route instructions intended for robots (e.g., "The purple goal area is on the right hand side of Aki [a robot] it is in front of you and diagonally to your left."). Speakers were instructed to inspect a drawing of a scene, consider their instructions then record these using a computer interface; they had the opportunity to re-record utterances that they considered unsatisfactory. All speech was recorded in a quiet meeting room with a headset microphone. In addition, the gold standard transcriptions had one indication of a mispronunciation and no indications of restarts, though the transcribers were instructed to transcribe any occurrences of them. Thus the speech materials used for the present investigation were of good quality. This should be kept in mind when interpreting the results described below.

The material used came from five male and five female speakers and ranged in duration from 6 seconds to 10 seconds (mean 8.0 sec), and from 5 words to 28 words (mean 17.4 words). Among these speakers, three were fluent non-native English speakers; the remaining speakers were native English speakers. The speakers were divided into four categories: (1) female native English speakers, (2) a female non-native English speaker, (3) male native English speakers, and (4) male non-native English speakers (see Table 1). From a set of 896 audio clips, 52 clips were selected for this study, 13 from each of the four categories above. Each set of 13 utterances had the lowest possible length variance within that set, that is, all selected audio clips in a given category were approximately the same duration.

Speaker Category	Mean Duration	Mean Word Count	Utterance Count
Female Native	6.0 sec	14.8 words	13 utts
Female Non-native	9.4 sec	22.6 words	13 utts
Male Native	9.6 sec	20.4 words	13 utts
Male Non-native	6.9 sec	11.7 words	13 utts

Table 1. Utterance information across speaker categories (female or male, native or non-native English).

5. MTURK TRANSCRIPTION ANALYSIS

Transcription quality was determined by computing word error rate (WER) using a standard procedure [6], with a gold standard that was transcribed using in-house guidelines [7] as reference. For purposes of the current comparison, both the reference transcription and those obtained from MTurk

[2] http://www.esnips.com

Payment (per trans)	WER
$0.005	4.79%
$0.01	6.53%
$0.03	4.33%
$0.05	4.19%
Aggregate	4.96%

Speaker Gender	WER
Male	3.74%
Female	6.01%

Speaker Native Language	WER
English	3.55%
Other	6.40%

Table 2. WER across payment, speaker gender, and speaker native language.

were normalized by using a spell checker to correct common spelling errors and remove annotations (mispronunciations, restarts, fillers, punctuation, etc.) that were not relevant for the current analysis.

5.1. Transcription Results

Results for the payment conditions are summarized in Table 2. Individual transcription WER is low, ranging from 4.19% to 6.53%. Note that transcribers had no knowledge of the domain other than that suggested by the speech content. Accuracy approaches the 95% transcriber agreement criterion that is expected when experts transcribe an utterance [8]. As reported by NIST, disagreement among expert transcribers is typically at a WER of 2-4% [6], and our results approach this level as well. The aggregate WER across the 1,040 transcriptions (4 payment conditions × 5 transcriptions per utterance) was 4.96%. The Sentence Error Rate (SER) – the fraction of transcriptions that had at least one disagreement with the gold standard – was 46.0%. Disagreements were often minor (see Section 5.2).

To investigate whether financial incentive might potentially influence transcription quality, we posted the same transcription tasks at 4 different payment levels, as mentioned in Section 3.1. Each batch of tasks had a set payment amount, and at any given time, only one batch was available for turkers to perform. Based on the work batch available, a turker was paid $0.005, $0.01, $0.03, or $0.05 per satisfactory transcription. Transcriptions were rejected if they were empty or had none of the words in the audio. Batches were posted in decreasing order of cost. We found similar error levels across conditions, except for the $0.01 condition. There was no statistically significant difference in WER across the $0.005, $0.03, and $0.05 payment groups ($F(2, 153) = 0.28$, $p = 0.76$). We believe that the outcome of the $0.01 condition is random, but the occurrence of such fluctuations needs to be taken into account in assessing the usefulness of the process.

We found that male speakers were transcribed more accurately than female speakers across all payment conditions ($F(1, 206) = 8.40$, $p < 0.05$). The native language of the speaker also influenced transcription quality, not surprisingly, with a statistically significant difference in WER ($F(1, 206) = 22.68$, $p < 0.05$).

Payment	Latency	Indiv.	ROVER-3	ROVER-4	ROVER-5
$0.005	62 hrs	4.79%	2.91%	3.37%	2.44%
$0.01	21.5 hrs	6.53%	4.00%	3.64%	2.27%
$0.03	1.25 hrs	4.33%	2.62%	2.72%	1.55%
$0.05	13.3 hrs	4.19%	2.68%	3.37%	2.33%
Aggregate		4.96%	3.05%	3.27%	2.14%

Table 3. Latency (turnaround time) and WER across payment with and without ROVER.

5.2 Latency & Error Analysis

The latency column in Table 3 shows that cost appears to affect the amount of time it takes to obtain transcription results (i.e., latency) although there is a great deal of variability. Each work batch required 52 utterances to be transcribed 5 times. As expected, our more lucrative HITs ($0.03/transcription and $0.05/transcription) were completed much faster than the cheaper HITs (82.7% less turnaround time). Note however, that the fact that the $0.05/transcription HIT was posted in the evening, during off-peak work hours, may have influenced turnaround time.

Out of a total MTurk transcription pool of 20,116 words transcribed (across all the payment conditions), 997 were errors (4.96% WER). Overall, the most common errors were substitutions (682) followed by insertions (210) and deletions (105). An analysis of errors revealed variable use of contractions, e.g. “you will” versus “you’ll”, “till” versus “until”, “you are” versus “you’re”, etc. Removing these differences further reduced aggregate WER to 3.61%.

6. USING MULTIPLE TRANSCRIPTIONS TO IMPROVE ACCURACY

One of the benefits of a service such as MTurk is that multiple transcriptions of an utterance can be obtained at a reasonable cost. Thus, one way to increase accuracy is to combine transcriptions from multiple turkers. NIST’s ROVER algorithm [9] is a well-known procedure for combining alternative recognition hypotheses through word-level voting. We can apply this technique to improve the accuracy of our transcriptions, although we need to establish the minimum number of transcriptions that need to be obtained for an improvement to be observed. In general, an

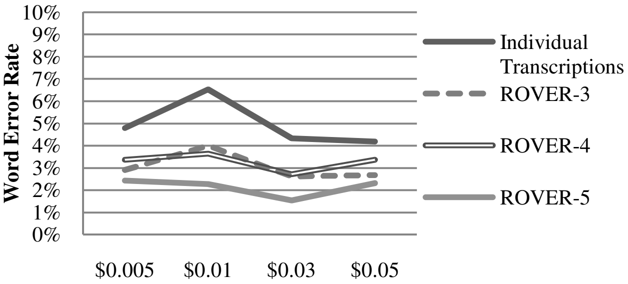

Figure 1. WER across payment amount with and without ROVER.

odd number of transcriptions is desirable because it will likely break ties as candidates are merged (ROVER randomly selects from the candidates if words are tied at a given position). We varied the number of transcriptions per utterance from three ("ROVER-3") to five ("ROVER-5").

6.1 ROVER Results

Figure 1 presents results from combining multiple transcriptions with the ROVER method. All combinations of each utterance's five transcriptions were calculated for each of the ROVER runs (e.g., all combinations of WER were calculated). Results are summarized in columns 3-6 of Table 3 by payment amount. As expected, combining all possible transcriptions of an utterance yielded the best results. The maximum number of possible transcriptions for an utterance was also an odd number (five), which improves results due to the ability to break word-level ties. Using ROVER to combine three transcriptions reduced aggregate WER by 39% to 3.05%, and combining five transcriptions reduced WER by 57% to 2.14%. In the latter case, SER is 24.2%, a 47.4% reduction from that of individual transcriptions. When contractions were resolved, as done in Section 5.2, aggregate WER after combining five transcriptions reduced to 0.84%.

Note that variations between payment conditions are attenuated when candidate utterances are combined together. In fact, there was no statistically significant difference in WER across payment groups when combining all 5 transcriptions of each utterance ($F(3, 203) = 0.59$, $p = 0.62$) The native language speaker characteristic also becomes less impactful (summarized in Figure 2). Based on optimizing both cost and accuracy, good choices here are to use either three or five candidates in the least costly payment condition. This would suggest that collecting multiple transcriptions should be part of the standard procedure in using MTurk.

7. CONCLUSIONS

We have shown that Amazon's Mechanical Turk (MTurk) can be used to accurately transcribe spoken language data. Experiments using MTurk for transcription were presented, and initial disagreement rates (WER) with a gold standard were approximately 5%. ROVER was found to be an effective method for improving accuracy (given sufficient parallel transcriptions). Combining multiple candidate transcriptions improved results to a 1.5-2.5% disagreement rate. This process could arguably make the role of a human transcription checker unnecessary.

Traditional transcription methods for spoken language data cost upwards of $100 per hour of speech, but our methods range in cost from $2.25 to $22.50 (one transcriber, with mean utterance length of 8 sec). Furthermore, accuracy was found to be unrelated to payment. Even at low payments, transcriber disagreement (after ROVER) is similar to traditional inter-transcriber disagreement, although lower payments may lead to longer delays in obtaining transcription results. We are not certain why turkers are willing to do this work, particularly at the very lowest payment rates. It's possible that they perceive these tasks as a diversion rather than as a source of income. We expect that over time turker behavior will evolve but it's difficult to predict in which direction this will happen.

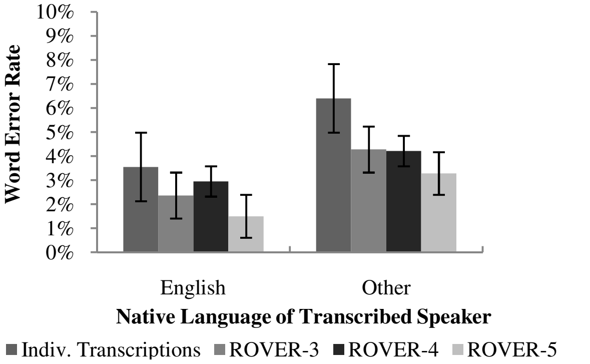

Figure 2. WER by native language of speaker transcribed.

At present we find that MTurk can be effectively and cheaply used for transcription of clean speech data. In the future we will evaluate MTurk for the transcription of more "difficult" speech, such as conversational speech or speech in noisy backgrounds. Although there is evidence that turkers can perform some linguistic annotation tasks, it would be useful to establish the limits of their ability.

8. REFERENCES

[1] A. Gruenstein, I. McGraw, and A. Sutherland, "A self-transcribing speech corpus: collecting continuous speech with an online educational game," in *SLaTE Workshop*, 2009.

[2] I. McGraw, A. Gruenstein, and A. Sutherland, "A self-labeling speech corpus: Collecting spoken words with an online educational game," in *Interspeech*, 2009.

[3] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast - but is it good? Evaluating non-expert annotations for natural langauge tasks," in *EMNLP*, 2008.

[4] C. Callison-Burch, "Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk," in *EMNLP*, 2009.

[5] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with Mechanical Turk," in *CHI '08*, Florence, Italy, 2008.

[6] The NIST Rich Transcription Evaluation Project. [Online]. http://www.itl.nist.gov/iad/mig/tests/rt

[7] C. Bennett and A. I. Rudnicky, "The Carnegie Mellon Communicator corpus," in *ICSLP*, 2002.

[8] K. Zechner, "What did they actually say? Agreement and Disagreement among Transcribers of Non-Native Spontaneous Speech Responses in an English Proficiency Test," in *SLaTE Workshop*, 2009.

[9] J. G. Fiscus, "A post-processing system to yield word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *ASRU Workshop*, 1997.