Using the Amazon Mechanical Turk to Transcribe and Annotate Meeting Speech for Extractive Summarization

Matthew Marge Satanjeev Banerjee Alexander I. Rudnicky
School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213, USA
{mrmarge,banerjee,air}@cs.cmu.edu

Abstract

Due to its complexity, meeting speech provides a challenge for both transcription and annotation. While Amazon's Mechanical Turk (MTurk) has been shown to produce good results for some types of speech, its suitability for transcription and annotation of spontaneous speech has not been established. We find that MTurk can be used to produce high-quality transcription and describe two techniques for doing so (voting and corrective). We also show that using a similar approach, high quality annotations useful for summarization systems can also be produced. In both cases, accuracy is comparable to that obtained using trained personnel.

1 Introduction

Recently, Amazon's Mechanical Turk (MTurk) has been shown to produce useful transcriptions of speech data; Gruenstein et al. (2009) have successfully used MTurk to correct the transcription output from a speech recognizer, while Novotney and Callison-Burch (2010) used MTurk for transcribing a corpus of conversational speech. These studies suggest that transcription, formerly considered to be an exacting task requiring at least some training, could be carried out by casual workers. However, only fairly simple transcription tasks were studied.

We propose to assess the suitability of MTurk for processing more challenging material, specifically recordings of meeting speech. Spontaneous speech can be difficult to transcribe because it may contain false starts, disfluencies, mispronunciations and other defects. Similarly for annotation, meeting content may be difficult to follow and conventions difficult to apply consistently.

Our first goal is to ascertain whether MTurk transcribers can accurately transcribe spontaneous speech, containing speech errors and of variable utterance length.

Our second goal is to use MTurk for creating annotations suitable for extractive summarization research, specifically labeling each utterance as either "in-summary" or "not in-summary". Among other challenges, this task cannot be decomposed into small independent sub-tasks—for example, annotators cannot be asked to annotate a single utterance independent of other utterances. To our knowledge, MTurk has not been previously explored for the purpose of summarization annotation.

2 Meeting Speech Transcription Task

We recently explored the use of MTurk for transcription of short-duration clean speech (Marge et al., 2010) and found that combining independent transcripts using ROVER yields very close agreement with a gold standard (2.14%, comparable to expert agreement). But simply collecting independent transcriptions seemed inefficient: the "easy" parts of each utterance are all transcribed the same. In the current study our goal is determine whether a smaller number of initial transcriptions can be used to identify easy- and difficult-to-transcribe regions, so that the attention of subsequent transcribers can be focused on the more difficult regions.

2.1 Procedure

In this *corrective* strategy for transcription, we have two turkers to independently produce transcripts. A word-level minimum edit distance metric is then used to align the two transcripts and locate disagreements. These regions are replaced with underscores, and new turkers are asked to transcribe those regions.

Utterances were balanced for transcription difficulty (measured by the native English back-

Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 99–107,
Los Angeles, California, June 2010.

ground of the speaker and utterance length). For the first pass transcription task, four sets of jobs were posted for turkers to perform, with each paying $0.01, $0.02, $0.04, or $0.07 per approved transcription. Payment was linearly scaled with the length of the utterance to be transcribed at a rate of $0.01 per 10 seconds of speech, with an additional payment of $0.01 for providing feedback. In each job set, there were 12 utterances to be transcribed (yielding a total of 24 jobs available given two transcribers per utterance). Turkers were free to transcribe as many utterances as they could across all payment amounts.

After acquiring two transcriptions, we aligned them, identified points of disagreement and reposted the transcripts and the audio as part of a next round of job sets. Payment amounts were kept the same based on utterance length. In this second pass of transcriptions, three turkers were recruited to correct and amend each transcription. Thus, a total of five workers worked on every transcription after both iterations of the corrective task. In our experiment 23 turkers performed the first phase of the task, and 28 turkers the corrective task (4 workers did both passes).

2.2 First and Second Pass Instructions

First-pass instructions asked turkers to listen to utterances with an embedded audio player provided with the HIT. Turkers were instructed to transcribe every word heard in the audio and to follow guidelines for marking speaker mispronunciations and false starts. Filled pauses ('uh', 'um', etc.) were not to be transcribed in the first pass. Turkers could replay the audio as many times as necessary.

In the second pass, turkers were instructed to focus on the portions of the transcript marked with underscores, but also to correct any other words they thought were incorrect. The instructions also asked turkers to identify three types of filler words: "uh", "um", and "lg" (laughter). We selected this set since they were the most frequent in the gold standard transcripts. Again, turkers could replay the audio.

2.3 Speech Corpus

The data were sampled from a previously-collected corpus of natural meetings (Banerjee and Rudnicky, 2007). The material used in this paper comes from four speakers, two native English speakers and two non-Native English speakers (all male). We selected 48 audio clips; 12 from each of the four speakers. Within each speaker's set of clips, we further divided the material into four length categories: ~5, ~10, ~30 and ~60 sec. The speech material is conversational in nature; the gold standard transcriptions of this data included approximately 15 mispronunciations and 125 false starts. Table 1 presents word count information related to the utterances in each length category.

Utterance Length	Word Count (mean)	Standard Deviation	Utterance Count
5 sec	14	5.58	12
10 sec	24.5	7.26	12
30 sec	84	22.09	12
60 sec	146.6	53.17	12

Table 1. Utterance characteristics.

3 Meeting Transcription Analysis

Evaluation of first and second pass corrections was done by calculating word error rate (WER) with a gold standard, obtained using the transcription process described in (Bennett and Rudnicky, 2002). Before doing so, we normalized the candidate MTurk transcriptions as follows: spell-checking (with included domain-specific technical terms), and removal of punctuation (periods, commas, etc.). Apostrophes were retained.

Utterance Length	First-Pass WER	Second-Pass WER	ROVER-3 WER
5 sec.	31.5%	19.8%	15.3%
10 sec.	26.7%	20.3%	13.8%
30 sec.	20.8%	16.9%	15.0%
60 sec.	24.3%	17.1%	15.4%
Aggregate	23.8%	17.5%	15.1%

Table 2. WER across transcription iterations.

3.1 First-Pass Transcription Results

Results from aligning our first-pass transcriptions with a gold standard are shown in the second column of Table 2. Overall error rate was 23.8%, which reveals the inadequacy of individual turker transcriptions, if no further processing is done. (Remember that first-pass transcribers were asked to leave out fillers even though the gold standard contained them, thus increasing WER).

In this first pass, speech from non-native speakers was transcribed more poorly (25.4% WER) than speech from native English speakers (21.7% WER). In their comments sections, 17% of turkers noted the difficulty in transcribing non-native speakers, while 13% found native English speech difficult. More than 80% of turkers thought the amount of work "about right" for the payment received.

3.2 Second-Pass Transcription Results

The corrective process greatly improved agreement with our expert transcriptions. Aggregate WER was reduced from 23.8% to 17.5% (27% relative reduction) when turkers corrected initial transcripts with highlighted disagreements (third column of Table 2). In fact, transcriptions after corrections were significantly more accurate than initial transcriptions ($F(1, 238) = 13.4$, $p < 0.05$). With respect to duration, the WER of the 5-second utterances had the greatest improvement, a relative reduction of WER by 37%. Transcription alignment with the gold standard experienced a 39% improvement to 13.3% for native English speech, and a 19% improvement to 20.6% for non-native English speech (columns 2 and 3 of Table 3).

We found that 30% of turkers indicated that the second-pass correction task was difficult, as compared with 15% for the first-pass transcription task. Work amount was perceived to be about right (85% of the votes) in this phase, similar to the first.

3.3 Combining Corrected Transcriptions

In order to improve the transcriptions further, we combined the three second-pass transcriptions of each utterance using ROVER's word-level voting scheme (Fiscus, 1997). The WER of the resulting transcripts are presented in the fourth column of Table 2. Aggregate WER was further reduced by 14% relative to 15.1%. This result is close to typical disagreement rates of 6-12% reported in the literature (Roy and Roy, 2009). The best improvements using ROVER were found with the transcriptions of the shorter utterances: WER from the second-pass of 5-second utterances transcriptions was reduced by 23% to 15.3%. The 10-second utterance transcriptions experienced the best improvement, 32%, to a WER of 13.8%.

Although segmenting audio into shorter segments may yield fast turnaround times, we found that utterance length is not a significant factor in determining alignment between combined, corrected transcriptions and gold-standard transcriptions ($F(3, 44) = 0.16$, $p = 0.92$). We speculate that longer utterances show good accuracy due to the increased context available to transcribers.

Speaker Background	First-Pass WER	Second-Pass WER	ROVER-3 WER
Native	21.7%	13.3%	10.8%
Non-native	25.4%	20.6%	18.4%

Table 3. WER across transcription iterations based on speaker background.

3.4 Error Analysis

Out of 3,281 words (48 merged transcriptions of 48 utterances), 496 were errors. Among the errors were 37 insertions, 315 deletions, and 144 substitutions. Thus the most common error was to miss a word.

Further analysis revealed that two common cases of errors occurred: the misplacement or exclusion of filler words (even though the second phase explicitly instructed turkers to insert filler words) and failure to transcribe words considered to be out of the range of the transcriber's vocabulary, such as technical terms and foreign names. Filler words accounted for 112 errors (23%). Removing fillers from both the combined transcripts and the gold standard improved WER by 14% relative to 13.0%. Further, WER for native English speech transcriptions was reduced to 8.9%. This difference was however not statistically significant ($F(1,94) = 1.64$, $p = 0.2$).

Turkers had difficulty transcribing uncommon words, technical terms, names, acronyms, etc. (e.g., "Speechalyzer", "CTM", "PQs"). Investigation showed that at least 41 errors (8%) could be attributed to this out-of-vocabulary problem. It is unclear if there is any way to completely eradicate such errors, short of asking the original speakers.

3.5 Comparison to One-Pass Approach

Although the corrective model provides significant gain from individual transcriptions, this approach is logistically more complex. We compared it to our one-pass approach, in which five turkers independently transcribe all utterances (Marge et al., 2010). Five new transcribers per utterance were recruited for this task (yielding 240 transcriptions).

Individual error rate was 24.0%, comparable to the overall error rate for the first step of the corrective approach (Table 2).

After combining all five transcriptions with ROVER, we found similar gains to the corrective approach: an overall improvement to 15.2% error rate. Thus both approaches can effectively produce high-quality transcriptions. We speculate that if higher accuracy is required, the corrective process could be extended to iteratively re-focus effort on the regions of greatest disagreement.

3.6 Latency

Although payment scaled with the duration of utterances, we observed a consistent disparity in turnaround time. All HITs were posted at the same time in both iterations (Thursday afternoon, EST). Turkers were able to transcribe 48 utterances twice in about a day in the first pass for the shorter utterances (5- and 10-second utterances), while it took nearly a week to transcribe the 30- and 60-second utterances. Turkers were likely discouraged by the long duration of the transcriptions compounded with the nature of the speech. To increase turnaround time on lengthy utterances, we speculate that it may be necessary to scale payment non-linearly with length (or another measure of perceived effort).

3.7 Conclusion

Spontaneous speech, even in long segments, can indeed be transcribed on MTurk with a level of accuracy that approaches expert agreement rates for spontaneous speech. However, we expect segmentation of audio materials into smaller segments would yield fast turnaround time, and may keep costs low. In addition, we find that ROVER works more effectively on shorter segments because lengths of candidate transcriptions are less likely to have large disparities. Thus, multiple transcriptions per utterance can be utilized best when their lengths are shorter.

4 Annotating for Summarization

4.1 Motivation

Transcribing audio data into text is the first step towards making information contained in audio easily accessible to humans. A next step is to condense the information in the raw transcription, and produce a short summary that includes the most important information. Good summaries can provide readers with a general sense of the meeting, or help them to drill down into the raw transcript (or the audio itself) for additional information.

4.2 Annotation Challenges

Unfortunately, summary creation is a difficult task because "importance" is inherently subjective and varies from consumer to consumer. For example, the manager of a project, browsing a summary of a meeting, might be interested in all agenda items, whereas a project participant may be interested in only those parts of the meeting that pertain to his portion of the project.

Despite this subjectivity, the usefulness of a summary is clear, and audio summarization is an active area of research. Within this field, two kinds of human annotations are generally created—annotators are either asked to write a short summary of the audio, or they are asked to label each transcribed utterance as either "in summary" or "out of summary". The latter annotation is particularly useful for training and evaluating *extractive* summarization systems—systems that create summaries by selecting a subset of the utterances.

Due to the subjectivity involved, we find very low inter-annotator agreement for this labeling task. Liu and Liu (2008) reported Kappa agreement scores of between 0.11 and 0.35 across 6 annotators, Penn and Zhu (2008) reported 0.38 on telephone conversation and 0.37 on lecture speech, using 3 annotators, and Galley (2006) reported 0.32 on meeting data. Such low levels of agreement imply that the resulting training data is likely to contain a great deal of "noise"—utterances labeled "in summary" or "out of summary", when in fact they are not good examples of those classes.

Disagreements arise due to the fact that utterance importance is a spectrum. While some utterances are clearly important or unimportant, there are many utterances that lie between these extremes. In order to label utterances as either "in-summary" or not, annotators must choose an arbitrary threshold at which to make this decision. Simply asking annotators to provide a continuous "importance value" between 0 and 1 is also likely to be infeasible as the exact value for a given utterance is difficult to ascertain.

4.3 3-Class Formulation

One way to alleviate this problem is to redefine the task as a 3-class labeling problem. Annotators can be asked to label utterances as either "important", "unimportant" or "in-between". Although this formulation creates two decision boundaries, instead of the single one in the 2-class formulation, the expectation is that a large number of utterances with middling importance will simply be assigned to the "in between" class, thus reducing the amount of noise in the data. Indeed we have shown (Banerjee and Rudnicky, 2009) that in-house annotators achieve high inter-annotator agreement when provided with the 3-class formulation.

Another way to alleviate the problem of low agreement is to obtain annotations from many annotators, and identify the utterances that a majority of the annotators appear to agree on; such utterances may be considered as good examples of their class. Using multiple annotators is typically not feasible due to cost. In this paper we investigate using MTurk to create 3-class-based summarization annotations from multiple annotators per meeting, and to combine and filter these annotations to create high quality labels.

5 Using Mechanical Turk for Annotations

5.1 Challenges of Using Mechanical Turk

Unlike some other tasks that require little or no context in order to perform the annotation, summarization annotation requires a great deal of context. It is unlikely that an annotator can determine the importance of an utterance without being aware of neighboring utterances. Moreover, the appropriate length of context for a given utterance is likely to vary. Presenting all contiguous utterances that discuss the same topic might be appropriate, but would require manual segmentation of the meeting into topics. In this paper we experiment with showing *all* utterances of a meeting. This is a challenge however, because MTurk is typically applied to quick low-cost tasks that need little context. It is unclear whether turkers would be willing to perform such a time-consuming task, even for higher payment.

Another challenge for turkers is being able to understand the discussion well enough to perform the annotation. We experiment here with meetings that include significant technical content. While in-house annotators can be trained over time to understand the material well enough to perform the task, it is impractical to provide turkers with such training. We investigate the degree to which turkers can provide summarization annotation with minimal training.

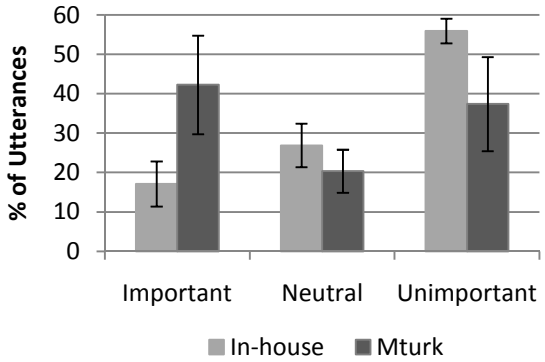


Figure 1. Label distribution of in-house and MTurk annotators.

5.2 Data Used

We selected 5 recorded meetings for our study. These meetings were not scripted—and would have taken place even if they weren't being recorded. They were project meetings containing discussions about software deliverables, problems, resolution plans, etc. The contents included technical jargon and concepts that non-experts are unlikely to grasp by reading the meeting transcript alone.

The 5 meetings had 2 to 4 participants each (mean: 3.5). For all meetings, the speech from each participant was recorded separately using head-mounted close-talking microphones. We manually split these audio streams into utterances—ensuring that utterances did not have more than a 0.5 second pause in them, and then transcribed them using an established process (Bennett and Rudnicky, 2002). The meetings varied widely in length from 15 minutes and 282 utterances to 40 minutes and 948 utterances (means: 30 minutes, 610 utterances). There were 3,052 utterances across the 5 meetings, each containing an mean of 7 words. The utterances in the meetings were annotated using the 3-class formulation by two in-house annotators. Their inter-annotator agreement is presented along with the rest of the evaluation results in Section 6.

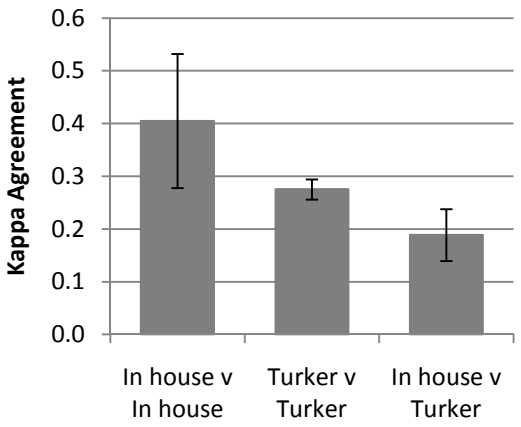


Figure 2. Average kappa agreement between in-house annotators, turkers, and in-house annotators and turkers.

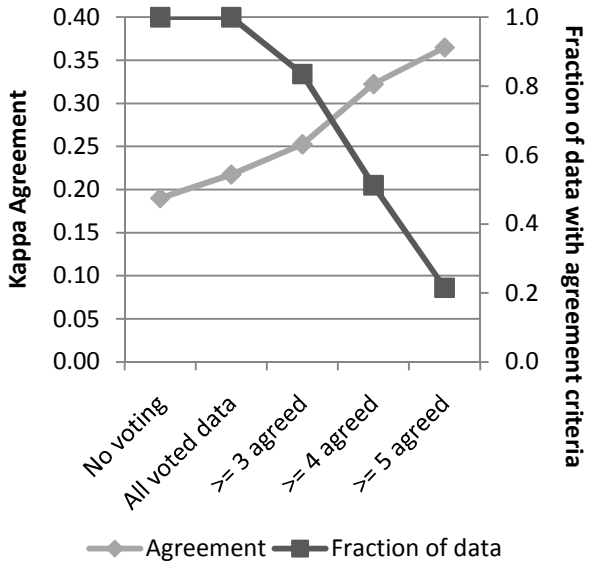


Figure 3. Agreement with in-house annotators when turker annotations are merged through voting.

5.3 HIT Design and Instructions

We instructed turkers to imagine that someone else (not them) was going to eventually write a report about the meeting, and it was their task to identify those utterances that should be included in the report. We asked annotators to label utterances as "important" if they should be included in the report and "unimportant" otherwise. In addition, utterances that they thought were of medium importance and that *may or may not* need to be included in the report were to be labeled as "neutral". We provided examples of utterances in each of these classes. For the "important" class, for instance, we included "talking about a problem" and "discussing future plan of action" as examples. For the "unimportant" class, we included "off topic joking", and for the "neutral" class "minute details of an algorithm" was an example.

In addition to these instructions and examples, we gave turkers a general guideline to the effect that in these meetings typically 1/4th of the utterances are "important", 1/4th "neutral" and the rest "unimportant". As we discuss in section 6, it is unclear whether most turkers followed this guideline.

Following these instructions, examples and tips, we provided the text of the utterances in the form of an HTML table. Each row contained a single utterance, prefixed with the name of the speaker. The row also contained three radio buttons for the three classes into which the annotator was asked to classify the utterance. Although we did not ensure that annotators annotated every utterance before submitting their work, we observed that for 95% of the utterances every annotator did provide a judgment; we ignore the remaining 5% of the utterances in our evaluation below.

5.4 Number of Turkers and Payment

For each meeting, we used 5 turkers and paid each one the same. That is, we did not vary the payment amount as an experimental variable. We calculated the amount to pay for a meeting based on in the length of that meeting. Specifically, we multiplied the number of utterances by 0.13 US cents to arrive at the payment. This resulted in payments ranging from 35 cents to $1.25 per meeting (mean 79 cents). The effective hourly rate (based on how much time turkers took to actually finish each job) was $0.87.

6 Annotation Results

6.1 Label Distribution

We first examine the average distribution of labels across the 3 classes. Figure 1 shows the distributions (expressed as percentages of the number of utterances) for in-house and MTurk annotators, averaged across the 5 meetings. Observe that the distribution for the in-house annotators is far more skewed away from a uniform 33% assignment, whereas the label distribution of turkers is less skewed. The likely reason for this difference is that

turkers have a poorer understanding of the meetings, and are more likely than in-house annotators to make arbitrary judgments about utterances. This poor understanding perhaps also explains the large difference in the percentage of utterances labeled as important—for many utterances that are difficult to understand, turkers probably play it safe by marking it important.

The error bars represent the standard deviations of these averages, and capture the difference in label distribution from meeting to meeting. While different meetings are likely to inherently have different ratios of the 3 classes, observe that the standard deviations for the in-house annotators are much lower than those for the turkers. For example, the percentage of utterances labeled "important" by in-house annotators varies from 9% to 22% across the 5 meetings, whereas it varies from 30% to 57% for turkers, a much wider range. These differences in standard deviation persist for each meeting as well—that is, for any given meeting, the label distribution of the turkers varies much more between each other than the distribution of the in-house annotators.

6.2 Inter-Annotator Agreement

Figure 2 shows the kappa values for pairs of annotators, averaged across the 5 meetings, while the error bars represent the standard deviations. The kappa between the two in-house annotators. (0.4) is well within the range of values reported in the summarization literature (see section 4). The kappa values range from 0.24 to 0.50 across the 5 meetings. The inter-annotator agreement between pairs of turkers, averaged across the 10 possible pairs per meeting (5 choose 2), and across the 5 meetings show that turkers tend to agree less between each other than in-house annotators, although this kappa (0.28) is still within the range of typical agreement (this kappa has lower variance because the sample size is larger). The kappa between in-house annotators and turkers[1] (0.19) is on the lower end of the scale but remains within the range of agreement reported in the literature, suggesting that Mechanical Turk may be a useful tool for summarization.

[1] For each meeting, we measure agreement between every possible pair of annotators such that one of the annotators was an in-house annotator, and the other a turker. Here we present the average agreement across all such pairs, and across all the meetings.

6.3 Agreement after Voting

We consider merging the annotations from multiple turkers using a simple voting scheme as follows. For each utterance, if 3, 4 or 5 annotators labeled the utterance with the same class, we labeled the utterance with that class. For utterances in which 2 annotators voted for one class, 2 for another and 1 for the third, we randomly picked from one of the classes in which 2 annotators voted the same way. We then computed agreement between this "voted turker" and each of the two in-house annotators, and averaged across the 5 meetings. Figure 3 shows these agreement values. The left-most point on the "Kappa Agreement" curve shows the average agreement obtained using individual turkers (0.19) while the second point shows the agreement with the "voted turker" (0.22). This is only a marginal improvement, implying that simply voting and using all the data does not improve much over the average agreement of individual annotators.

The agreement does improve when we consider only those utterances that a clear majority of annotators agreed on. The 3rd, 4th and 5th points on the "Agreement" curve plot the average agreement when considering only those utterances that at least 3, 4 and 5 turkers agreed on. The "Fraction of data" curve plots the fraction of the meeting utterances that fit these agreement criteria. For utterances that at least 3 turkers agreed on, the kappa agreement value with in-house annotators is 0.25, and this represents 84% of the data. For about 50% of the data 4 of 5 turkers agreed, and these utterances had a kappa of 0.32. Finally utterances for which annotators were unanimous had a kappa of 0.37, but represented only 22% of the data. It is particularly encouraging to note that although the amount of data reduces as we focus on utterances that more and more turkers agree on, the utterances so labeled are not dominated by any one class. For example, among utterances that 4 or more turkers agree on, 48% belong to the important class, 48% to unimportant class, and the remaining 4% to the neutral class. These results show that with voting, it is possible to select a subset of utterances that have higher agreement rates, implying that they are annotated with higher confidence. For future work we will investigate whether a summarization system trained on only the highly agreed-upon data outperforms one trained on all the annotation data.

7 Conclusions

In this study, we found that MTurk can be used to create accurate transcriptions of spontaneous meeting speech when using a two-stage corrective process. Our best technique yielded a disagreement rate of 15.1%, which is competitive with reported disagreement in the literature of 6-12%. We found that both fillers and out-of-vocabulary words proved troublesome. We also observed that the length of the utterance being transcribed wasn't a significant factor in determining WER, but that the native language of the speaker was indeed a significant factor.

We also experimented with using MTurk for the purpose of labeling utterances for extractive summarization research. We showed that despite the lack of training, turkers produce labels with better than random agreement with in-house annotators. Further, when combined using voting, and with the low-agreement utterances filtered out, we can identify a set of utterances that agree significantly better with in-house annotations.

In summary, MTurk appears to be a viable resource for producing transcription and annotation of meeting speech. Producing high-quality outputs, however, may require the use of techniques such as ensemble voting and iterative correction or refinement that leverage performance of the same task by multiple workers.

References

S. Banerjee and A. I. Rudnicky. 2007. Segmenting meetings into agenda items by extracting implicit supervision from human note-taking. In *Proceedings of IUI.*

S. Banerjee and A. I. Rudnicky. 2009. Detecting the noteworthiness of utterances in human meetings. In *Proceedings of SIGDial.*

C. Bennett and A. I. Rudnicky. 2002. The Carnegie Mellon Communicator corpus. In *Proceedings of ICSLP.*

J. G. Fiscus. 1997. A post-processing system to yield word error rates: Recognizer Output Voting Error Reduction (ROVER). In *Proceedings of ASRU Workshop.*

M. Galley. (2006). A skip-chain conditional random field for ranking meeting utterances by im portance. In *Proceedings of EMNLP.*

A. Gruenstein, I. McGraw, and A. Sutherland. 2009. A self-transcribing speech corpus: collecting continuous speech with an online educational game. In *Proceedings of SLaTE Workshop.*

F. Liu and Y. Liu. 2008. Correlation between ROUGE and human evaluation of extractive meeting summaries. In *Proceedings of ACL-HLT.*

M. Marge, S. Banerjee, and A. I. Rudnicky. 2010. Using the Amazon Mechanical Turk for transcription of spoken language. In *Proceedings of ICASSP.*

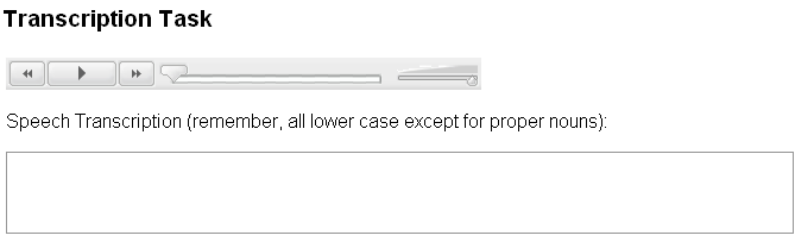
S. Novotney and C. Callison-Burch. 2010. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Proceedings of NAACL.*

G. Penn and X. Zhu. 2008. A critical reassessment of evaluation baselines for speech summarization. In *Proceedings of ACL-HLT.*

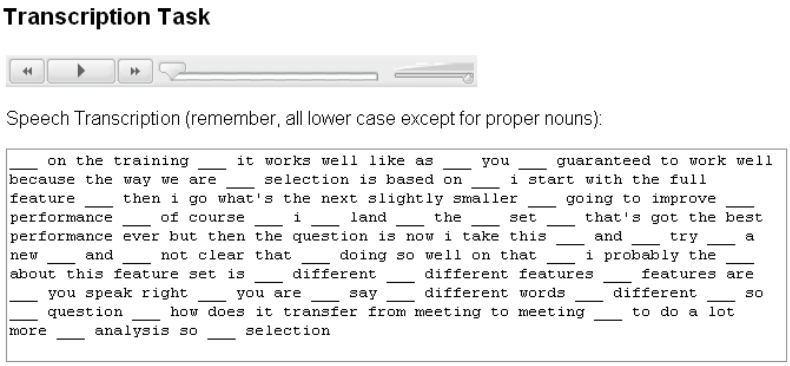
B. Roy and D. Roy. 2009. Fast transcription of unstructured audio recordings. In *Proceedings of In terspeech.*

Appendix

Transcription task HIT type 1:



Transcription task HIT type 2:



Annotation task HIT: