Motivation

• Speech recognition models hunger for data
  – ASR requires thousands of hours of transcribed audio
  – In-domain data needed to overcome mismatches like language, speaking style, acoustic channel, noise, etc…

• Conversational telephone speech transcription is difficult
  – Spontaneous speech between intimates
  – Rapid speech, phonetic reductions and varied speaking style
  – Expensive and time consuming
    • $150 / hour of transcription
    • 50 hours of effort / hour of transcription

• Deploying to new domains is slow and expensive
Amazon’s Mechanical Turk

• Online labor market where *Turkers* complete *HITs* for small change - *artificial* artificial intelligence
  – Pay as low as 1 cent, typically 5 – 25 cents
  – Turkers predominantly in U.S., but heavy presence in India

• New resource for NLP annotations
  – Sentiment Analysis
  – Translation
  – Relevance Feedback
  – Dozens More (see NAACL MTurk workshop this Sunday)

• Used by professional transcription company (CastingWords)
  – Limited to clean speech (podcasts, dictation)
  – Multi-level quality control / transcriber vetting
  – ~$90 / hour of transcription – look this up
Evaluating Mechanical Turk

• Prior work judged quality by comparing Turkers to experts
  – 10 Turkers match expert for many NLP tasks (Snow et al 2008)

• Other Mechanical Turk speech transcription paper this year
  – Robot Instructions ~3% WER (Marge 2010)
  – Street addresses, travel dialogue ~6% WER (McGraw 2010)

• Right metric depends on the data consumer
  – Humans: WER on transcribed data
  – Systems: WER on test data decoded with a trained system

• We explore three new directions
  – Much more difficult transcription task
  – Evaluate Mechanical Turk by system performance
  – Extending transcription beyond English
English Speech Corpus

• English Switchboard corpus
  – Ten minute conversations about an assigned topic
  – Two existing transcriptions for a twenty hour subset:
    • LDC – high quality, ~50xRT transcription time
    • Fisher ‘QuickTrans’ effort – 6xRT transcription time

• For those who care
• Callfriend language-identification corpora
  – Korean, Hindi, Tamil, Farsi, and Vietnamese
  – Conversations from U.S. to home country between friends
  – Mixture of English and native language
  – Only Korean has existing LDC transcriptions
OH WELL I GUESS RETIREMENT THAT KIND OF THING WHICH I DON'T WORRY MUCH ABOUT

UH AND WE HAVE A SOCCER TEAM THAT COMES AND GOES WE DON'T EVEN HAVE THAT PRETTY
Speech Transcription for $5/hour

- Paid $300 to transcribe 20 hours of Switchboard three times
  - $5 per hour of transcription ($0.05 per utterance)
  - 1089 Turkers completed the task in six days
  - 30 utterances transcribed on average (earning 15 cents)
  - 63 Turkers completed more than 100 utterances

- Some people complained about the cost
  - “wow that's a lot of dialogue for $.05”
  - “this stuff is really hard. pay per hit should be higher”

- Many enjoyed the task and found it interesting
  - “Very interesting exercise. would welcome more hits.”
  - “You don't grow pickles they are cucumbers!!!!”
Disagreement with Experts

23% mean disagreement

<table>
<thead>
<tr>
<th>Transcription</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>well ITS been nice talking to you again</td>
<td>12%</td>
</tr>
<tr>
<td>well it's been [DEL] A NICE PARTY JENGA</td>
<td>71%</td>
</tr>
<tr>
<td>well it's been nice talking to you again</td>
<td>0%</td>
</tr>
</tbody>
</table>
Estimated disagreement of 25%
True disagreement of 23%

<table>
<thead>
<tr>
<th>Transcription</th>
<th>WER</th>
<th>Est. WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>well ITS been nice talking to you again</td>
<td>12%</td>
<td>43%</td>
</tr>
<tr>
<td>well it's been [DEL] A NICE PARTY JENGA</td>
<td>71%</td>
<td>78%</td>
</tr>
<tr>
<td>well it's been nice talking to you again</td>
<td>0%</td>
<td>37%</td>
</tr>
</tbody>
</table>
Rating Turkers: Expert vs. Non-Expert

Disagreement Against Other Turkers vs. Disagreement Against Expert
Selecting Turkers by Estimated Skill

Disagreement Against Other Turkers

Disagreement Against Expert

![Graph showing correlation between disagreement against experts and other Turkers.](image)
Selecting Turkers by Estimated Skill

Disagreement Against Expert

- 57%
- 4.5%

Disagreement Against Other Turkers

- 12%
- 25%

Diagram showing the distribution of disagreement against an expert and against other Turkers.
Selecting Turkers by Estimated Skill

Disagreement Against Expert

Disagreement Against Other Turkers
Selecting Turkers by Estimated Skill

Disagreement Against Expert

Disagreement Against Other Turkers
Selecting Turkers by Estimated Skill

Disagreement Against Other Turkers

Disagreement Against Expert

Disagreement Against Other Turkers
Finding the Right Turkers

F-Score vs. WER Selection Threshold

Mean disagreement of 23%
Finding the Right Turkers

Mean Disagreement: 23%

Easy to reject bad workers

Hard to find good workers
Selecting Turkers by Estimated Skill

Disagreement Against Other Turkers

- 92%
- 4%
- 1%
- 2%

Disagreement Against Expert

- Disagreement vs. Expert
- Disagreement vs. Other Turkers

Graph shows a scatter plot with the x-axis representing disagreement against the expert, and the y-axis representing disagreement against other Turkers. The graph includes percentages for different regions of the scatter plot.
<table>
<thead>
<tr>
<th>Selection</th>
<th>LDC Disagreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>23%</td>
</tr>
<tr>
<td>ROVER</td>
<td>21%</td>
</tr>
<tr>
<td>Estimated Best Turker</td>
<td>20%</td>
</tr>
<tr>
<td>True Best Turker</td>
<td>18%</td>
</tr>
<tr>
<td>Oracle Best Utterance</td>
<td>13%</td>
</tr>
</tbody>
</table>
Mechanical Turk for ASR Training

- Ultimate test is system performance
  - Build acoustic and language models
  - Decode test set and compute WER
  - Compare to systems trained on equivalent expert transcription

- 23% professional disagreement might seem worrying
  - How does it effect system performance?
  - Do reductions in disagreement transfer to system gains?
  - What are best practices for improving ASR performance?
Breaking Down The Degradation

- Measured test WER degradation from 1 to 16 hours
  - 3% relative degradation for acoustic model
  - 2% relative degradation for language model
  - 5% relative degradation for both
  - *Despite 23% transcription disagreement with LDC*
Value of Repeated Transcription

• Each utterance was transcribed three times

• What is the value of this duplicate effort?
  – Instead of dreaming up a better combination method, use oracle error rate as upper bound on system combination

<table>
<thead>
<tr>
<th>Transcription</th>
<th>LDC Disagreement</th>
<th>ASR WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>23%</td>
<td>42.0%</td>
</tr>
<tr>
<td>Oracle</td>
<td>13%</td>
<td>40.9%</td>
</tr>
<tr>
<td>LDC</td>
<td>-</td>
<td>39.5%</td>
</tr>
</tbody>
</table>

• Cutting disagreement in half reduced degradation by half

• System combination has at most 2.5% WER to recover
How to Best Spend Resources?

- Given a fixed transcription budget, either:
  - Transcribe as much audio as possible
  - Improve quality by redundantly transcribing

<table>
<thead>
<tr>
<th>Transcription</th>
<th>Hours</th>
<th>Cost</th>
<th>ASR WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mturk</td>
<td>20</td>
<td>$100</td>
<td>42.0%</td>
</tr>
<tr>
<td>Oracle Mturk</td>
<td>20</td>
<td>$300</td>
<td>40.9%</td>
</tr>
<tr>
<td>MTurk</td>
<td>60</td>
<td>$300</td>
<td>37.6%</td>
</tr>
<tr>
<td>LDC</td>
<td>20</td>
<td></td>
<td>39.5%</td>
</tr>
</tbody>
</table>

- Get more data, not better data
  - Compare 37.6% WER versus 40.9% WER

- Even expert data is outperformed by more lower quality data
  - Compare 39.5% WER to 37.6% WER
How to Best Spend Resources?

- Given a fixed transcription budget, either:
  - Transcribe as much audio as possible
  - Improve quality by redundantly transcribing

<table>
<thead>
<tr>
<th>Transcription</th>
<th>Hours</th>
<th>Cost</th>
<th>ASR WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mturk</td>
<td>20</td>
<td>$100</td>
<td>42.0%</td>
</tr>
<tr>
<td>Oracle Mturk</td>
<td>20</td>
<td>$300</td>
<td>40.9%</td>
</tr>
<tr>
<td>MTurk</td>
<td>60</td>
<td>$300</td>
<td>37.6%</td>
</tr>
<tr>
<td>LDC</td>
<td>20</td>
<td>~$3000</td>
<td>39.5%</td>
</tr>
</tbody>
</table>

- Get more data, not better data
  - Compare 37.6% WER versus 40.9% WER
- Even expert data is outperformed by more lower quality data
  - Compare 39.5% WER to 37.6% WER
Comparing Cost of Reducing WER

System WER vs. Cost per Hour of Transcription (log scale)

- $150/hr - Professional
- $90/hr - CastingWords
- $5/hr - Mechanical Turk
- $15/hr - Mturk w/ Oracle QC
Non-English Transcription

- VOA
- Can we really do it cheap
- Callhome corpora
• Tiny labor pool (**initially two Turkers versus 1089 for English**)

• Posted separate ‘Pyramid Scheme’ HIT
  – Paid referrer 25% of what referred earns transcribing
  – Transcription costs $25/hour instead of $20/hour
  – 80% of transcriptions came from referrals

• Transcribed three hours in five weeks
  – Paid 8 Turkers $113 at a transcription rate of 10xRT

• Despite 17% CER, test CER only goes down by 1.5% relative
  – from 51.3% CER to 52.1% CER
  – Reinforces English conclusions about the usefulness of noisy data for training an ASR system
Tamil and Hindi

• Collected one hour of transcripts
  – Much larger labor pool – how many?
  – Paid $20/hour, finished in 8 days
  – Difficult to accurately convey instructions
    • Many translated Hindi audio to English

• No clear conclusions
  – A private contractor provided transcriptions
  – Very high disagreement (80%+) for both languages
    • Reference transcripts inaccurate
    • Colloquial speech, poor audio quality
    • English speech irregularly transliterated into Devanagari
    • Lax gender agreement both for speaking and transcribing
  – Hindi ASR might be a hard task
English Conclusions

• Mechanical Turk can quickly and cheaply transcribe difficult audio like English CTS
  – 10 hours a day for $5 / hour

• Can reasonably predict Turker skill w/out gold standard data
  – But this turns out not to be as important as we thought
  – Oracle selection still only cuts disagreement in half

• Trained models show little degradation despite 23% professional disagreement
  – Even perfect expert agreement has small impact on system performance (2.5% reduction in WER)
  – Resources better spent getting more data than better data
Foreign Language Conclusions

• Non-English Turkers are on Mechanical Turk
  – But not a field of dreams
    • “If you post it, they will come”

• Korean results reinforce English conclusions
  – 0.8% system degradation despite 17% disagreement
  – $20/hour (still very cheap)

• Small amounts of errorful data is useful
  – Poor models can still produce useable systems
    • 90% topic classification accuracy possible despite 80%+ WER
  – Semi-supervised methods can bootstrap initial models
    • 51% WER reduced to 27% with a one hour acoustic model

• Noisy data is much more useful than you think
Cheaply Estimating Turkers' Skill

Difference from Professional Estimate

Number of Utterances to Estimate Disagreement
Dealing with Real World Data

• Every word in the transcripts needs a pronunciation
  – Misspellings, new proper name spellings, jeez vs. geez
  – Inconsistent hesitation markings, myriad of ‘uh-huh’ spellings
  – 26% of utterances contained OOVs (10% of the vocabulary)

• Lots of elbow grease to prepare phonetic dictionary

• Turkers found creative ways not to follow instructions
  – Comments like “hard to hear” or “did the best I could :)
  – Enter transcriptions into wrong text box
  – But very few typed in gibberish

• We did not explicitly filter comments, etc…
For this work, we asked the following questions:

- Can/will Turkers perform very difficult tasks? Yes
- How cheaply can we collect training data? $5/hour
- Is redundant transcription cost effective? No
- Can we perform quality control without experts? Yes
- Can we collect non-English transcriptions? Kinda