Creative Data Collection: Crowdsourcing Translation

April 3, 2012
More data is better
Few languages have sufficient amounts of data.
Goals for this lecture

• Introduce you to crowdsourcing via Amazon’s Mechanical Turk

• Can we create training data for SMT?
  – How good are non-professional translators?
  – How much would it cost to create a parallel corpus?
  – Can we find speakers for low resource languages?
  – Can we train an SMT system from this data?

• What would low cost, high quality translations enable us to do?
Mechanical Turk is a marketplace for work. We give businesses and developers access to an on-demand, scalable workforce. Workers select from thousands of tasks and work whenever it’s convenient. 37,649 HITs available. View them now.

Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:
- Can work from home
- Choose your own work hours
- Get paid for doing good work

Find an interesting task → Work → Earn money

Find HITs Now

or learn more about being a Worker

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. Register Now

As a Mechanical Turk Requester you:
- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you’re satisfied with the results

Fund your account → Load your tasks → Get results

Get Started

or learn more about being a Requester
Enter Postmark & Stamp Information for this card

Postmark City:   Postmark State:   Postmark Date:   Postmark Year:   Stamp:
(or Country)    (Ex: Nov-09)    (Ex: 1909)    (Ex: 1c, 2c, half penny)
Postmark City: Barre
Postmark State: MA
Postmark Date: Oct-11
Postmark Year: 1886
Stamp: 1c

Reward: $0.01 per HIT
HITs Available: 2
Duration: 5 minutes
Who are the Turkers?

• Requesters are given very little information about Turkers - basically just a serial number
• No names, no demographic information (like what languages they speak)
• Cannot assume that they have a particular set of skills
• They should be treated as non-experts
• Quality control is a major challenge
• It important to design tasks to be simple and easy to understand
Other NLP applications

- Workshop on Using Mechanical Turk for Speech and Language Applications
- 35 researchers spent $100, wrote papers, and distributed their data
- Mark Dredze and I wrote an overview paper digesting the results
$100 Shared Task

- Traditional NLP tasks
  - WSD, RTE, NLG, common sense knowledge
- Speech and Vision
  - Transcribed speech, accented speech, handwriting OCR
- Sentiment, Polarity, Bias
  - Cross language, blogs
- Information Retrieval
  - TREC style annotations
- Information Extraction
  - Relation extraction, NER
- Machine Translation
  - Paraphrases, alignments, training and eval sets, rule cleaning
<table>
<thead>
<tr>
<th>Requester ID</th>
<th>Requester</th>
<th>#Task</th>
<th>#HITs</th>
<th>Rewards</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dolores Labs</td>
<td>A2IR7ETVOIULZU (RSS)</td>
<td>470</td>
<td>317857</td>
</tr>
<tr>
<td>2</td>
<td>ContentGalore</td>
<td>A2XL3J4NH6JI12 (RSS)</td>
<td>674</td>
<td>10622</td>
</tr>
<tr>
<td>3</td>
<td>SpeechInk</td>
<td>A1AQ7EJ5P7ME65 (RSS)</td>
<td>9019</td>
<td>13613</td>
</tr>
<tr>
<td>4</td>
<td>CastingWords</td>
<td>A3MI6MIUHWCR7F (RSS)</td>
<td>9454</td>
<td>14036</td>
</tr>
<tr>
<td>5</td>
<td>QuestionSwami</td>
<td>AD7C0BZNKGYGYY (RSS)</td>
<td>629</td>
<td>4116</td>
</tr>
<tr>
<td>6</td>
<td>Chris Callison-Burch</td>
<td>A32TTE4XXN6MQZ (RSS)</td>
<td>11</td>
<td>9961</td>
</tr>
<tr>
<td>7</td>
<td>Smartsheet.com Clients</td>
<td>A1197OGL0WOQ3G (RSS)</td>
<td>434</td>
<td>38212</td>
</tr>
<tr>
<td>8</td>
<td>retaildata</td>
<td>AD14NALRDOSN9 (RSS)</td>
<td>8</td>
<td>50288</td>
</tr>
<tr>
<td>9</td>
<td>Classify This</td>
<td>A1CTI3ZAWTR5AZ (RSS)</td>
<td>25</td>
<td>94590</td>
</tr>
<tr>
<td>10</td>
<td>Andrew Stephen</td>
<td>A1Y25F6MZCMQGY (RSS)</td>
<td>3</td>
<td>22705</td>
</tr>
<tr>
<td>11</td>
<td>Dolores Labs 2</td>
<td>A3JX8WONBL5N9X (RSS)</td>
<td>34</td>
<td>8976</td>
</tr>
<tr>
<td>12</td>
<td>RelevanceQuest</td>
<td>A8RMEN71ICE57 (RSS)</td>
<td>15</td>
<td>47881</td>
</tr>
<tr>
<td>13</td>
<td>Crowd Task</td>
<td>AFAOUHS65HNDS (RSS)</td>
<td>4</td>
<td>2388</td>
</tr>
<tr>
<td>14</td>
<td>nlds.soe.ucsc.edu</td>
<td>A1HI9DWCF794RE (RSS)</td>
<td>4</td>
<td>4702</td>
</tr>
</tbody>
</table>
Translation on Mechanical Turk

• Can Turkers do tasks as complex as translation?
• How do we design the task to thwart cheating?
• What mechanisms can we use to ensure high quality translations?
<script language="javascript">

function getUserInfo() {
    var userDisplayLanguage = navigator.language ? navigator.language : navigator.userDisplayLanguage;
    var country = sGeobytesCountry;
    var city = sGeobytesCity;
    var region = sGeobytesRegion;

    document.mturk_form.userDisplayLanguage.value = userDisplayLanguage;
    document.mturk_form.country.value = country;
    document.mturk_form.city.value = city;
    document.mturk_form.region.value = region;
}

getUserInfo();
// -->

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Location</th>
<th>Language</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>two of the other respectable career.</td>
<td>Lahore, Pakistan</td>
<td>All</td>
<td>Two were connected with miscellaneous honorable professions. Florida became the center of the initial European settlements, which included Pensacola, Fort Caroline and Saint Augustine. But clear goal in front of revolutionary leaders.</td>
</tr>
<tr>
<td>you people of business for the trade with the detainees.</td>
<td>Lahore, Pakistan</td>
<td>All</td>
<td>The prophet would also carry the merchantise of other people for purposes of trading.</td>
</tr>
<tr>
<td>the increase in employment and construction on a large scale.</td>
<td>Woodbury, United States</td>
<td>All</td>
<td>The rate of employment increased rapidly and construction work commenced on a huge scale.</td>
</tr>
<tr>
<td>palestinian muslim.</td>
<td>Riyadh, Saudi Arabia</td>
<td>All</td>
<td>Palestinian Muslims, Britain has the second world war from the works... Palestinians Muslims are dying but Israel did not get a special benefit. Moshe OOV Dean OOV terrorists in spite of the b...</td>
</tr>
<tr>
<td>the german citizenship.</td>
<td>Riyadh, Saudi Arabia</td>
<td>All</td>
<td>Gained German Nationality once again. 1906a OOV in zurek OOV univisti OOV has...</td>
</tr>
<tr>
<td>Worker ID</td>
<td>Years Speaking Hindi</td>
<td>Years Speaking English</td>
<td>City</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------------</td>
<td>------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>A3561XGP5Z6D75</td>
<td>16</td>
<td>16</td>
<td>Chennai</td>
</tr>
<tr>
<td>A6SJI1FPZTT7I</td>
<td>26</td>
<td>20</td>
<td>Taipei</td>
</tr>
<tr>
<td>A3KBG0N9VNAK3</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>A6SJI1FPZTT7I</td>
<td>26</td>
<td>20</td>
<td>Taipei</td>
</tr>
<tr>
<td>A2IZPICYIBNNXL</td>
<td>7</td>
<td>7</td>
<td>Cochin</td>
</tr>
<tr>
<td>A6SJI1FPZTT7I</td>
<td>26</td>
<td>20</td>
<td>Taipei</td>
</tr>
<tr>
<td>A6SJI1FPZTT7I</td>
<td>26</td>
<td>20</td>
<td>Hong Kong (SAR)</td>
</tr>
<tr>
<td>AK32OHKIJHRLC</td>
<td></td>
<td></td>
<td>Hyderabad</td>
</tr>
<tr>
<td>A6SJI1FPZTT7I</td>
<td>26</td>
<td>20</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>A6SJI1FPZTT7I</td>
<td>26</td>
<td>20</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>Country</td>
<td>Place</td>
<td>Language</td>
<td>Cultural Significance</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------</td>
<td>----------------</td>
<td>---------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>India</td>
<td>Cochin</td>
<td>Hindi</td>
<td>Was the last Mughal King Bahadur Shah Zafar.</td>
</tr>
<tr>
<td>India</td>
<td>Bangalore</td>
<td>Tamil</td>
<td>State legislature</td>
</tr>
<tr>
<td>India</td>
<td>Bangalore</td>
<td>Tamil</td>
<td>Akṣṭj Ejan (search engine) to give Sraṇth kalad...</td>
</tr>
<tr>
<td>Taiwan</td>
<td>Taipei</td>
<td>Mandarin</td>
<td>India (T? Involved) word means or Avideka inner...</td>
</tr>
<tr>
<td>Hong Kong (SAR)</td>
<td>Hong Kong</td>
<td>Mandarin</td>
<td>2. Supreme Court Nirhyanusar ST 324 [1] in the...</td>
</tr>
<tr>
<td>Taiwan</td>
<td>Taipei</td>
<td>Mandarin</td>
<td>Category: Internet</td>
</tr>
<tr>
<td>India</td>
<td>Mumbai</td>
<td>Hindi</td>
<td>Divine Miracle (Maoajja)</td>
</tr>
<tr>
<td>India</td>
<td>Cochin</td>
<td>Hindi</td>
<td>Padoz are rows of other places in the garden an...</td>
</tr>
</tbody>
</table>
In 1994, the remains of pre-historic man, which are believed to be eight hundred thousand years old were discovered and they were named ‘Home Antecessor’ meaning ‘The Founding Man’. Prior to that 6 lac years old humans, named as Homogenises in scientific terms, were believed to be the oldest dwellers of this area.
Measuring Translation Quality

• We’d like to quantify how good (or bad) the translations are

• Automatic metrics (like Bleu) compare a translation against a set of reference translations

• Generally used to measure machine translation quality but we apply them to humans

• Measure one professional translator’s translations against three other professionals
  – Gives us a range of what to expect in good translations
Bleu Scores (professional v Turk w/o QC)

- LDC professional translation: 0.42
- Machine Translation\textsubscript{19}: 0.27
- Turker average: 0.28
- Turker average (images only): 0.3
Improving quality through redundancy

• We have collected multiple translations for each source segment
  – 4 translations, a total of 10 edits

• Selection strategies
  – Pick best on a sentence-by-sentence basis
  – Try to pick the best Turker

• Goal: use machine learning to predict translation goodness
<table>
<thead>
<tr>
<th>Avoiding dieting to prevent from flu</th>
<th>abstention from dieting in order to avoid Flu</th>
<th>Abstain from decrease eating in order to escape from flue</th>
<th>In order to be safer from flu quit dieting</th>
</tr>
</thead>
<tbody>
<tr>
<td>This research of American scientists came in front after experimenting on mice.</td>
<td>This research from the American Scientists have come up after the experiments on rats.</td>
<td>This research of American scientists was shown after many experiments on mouses.</td>
<td>According to the American Scientist this research has come out after much experimentations on rats.</td>
</tr>
<tr>
<td>Experiments proved that mice on a lower calorie diet had comparatively less ability to fight the flu virus.</td>
<td>in has been proven from experiments that rats put on diet with less calories had less ability to resist the Flu virus.</td>
<td>It was proved by experiments the low calories eaters mouses had low defending power for flu in ratio.</td>
<td>Experiments have proved that those rats on less calories diet have developed a tendency of not overcoming the flu virus.</td>
</tr>
<tr>
<td>research has proven this old myth wrong that its better to fast during fever.</td>
<td>Research disproved the old axiom that &quot;It is better to fast during fever&quot;</td>
<td>The research proved this old talk that decrease eating is useful in fever.</td>
<td>This Research has proved the very old saying wrong that it is good to starve while in fever.</td>
</tr>
<tr>
<td>Avoiding dieting to prevent from flu</td>
<td>abstention from dieting in order to avoid Flu</td>
<td>Abstain from decrease eating in order to escape from flue</td>
<td>In order to be safer from flu quit dieting</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>---------------------------------------------</td>
<td>-------------------------------------------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>This research of American scientists came in front after experimenting on mice.</td>
<td>This research from the American Scientists have come up after the experiments on rats.</td>
<td>This research of American scientists was shown after many experiments on mouses.</td>
<td>According to the American Scientist this research has come out after much experimentations on rats.</td>
</tr>
<tr>
<td>Experiments proved that mice on a lower calorie diet had comparatively less ability to fight the flu virus.</td>
<td>in has been proven from experiments that rats put on diet with less calories had less ability to resist the Flu virus.</td>
<td>It was proved by experiments the low calories eaters mouses had low defending power for flue in ratio.</td>
<td>Experiments have proved that those rats on less calories diet have developed a tendency of not overcoming the flu virus.</td>
</tr>
<tr>
<td>Research has proven this old myth wrong that its better to fast during fever.</td>
<td>Research disproved the old axiom that &quot;It is better to fast during fever&quot;</td>
<td>The research proved this old talk that decrease eating is useful in fever.</td>
<td>This Research has proved the very old saying wrong that it is good to starve while in fever.</td>
</tr>
<tr>
<td>Avoiding dieting to prevent from flu</td>
<td>Abstention from dieting in order to avoid Flu</td>
<td>Abstain from decrease eating in order to escape from flue</td>
<td>In order to be safer from flu quit dieting</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>---------------------------------------------</td>
<td>--------------------------------------------------------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>This research of American scientists came in front after experimenting on mice.</td>
<td>This research from the American Scientists have come up after the experiments on rats.</td>
<td>This research of American scientists was shown after many experiments on mouses.</td>
<td>According to the American Scientist this research has come out after much experimentations on rats.</td>
</tr>
<tr>
<td>Experiments proved that mice on a lower calorie diet had comparatively less ability to fight the flu virus.</td>
<td>in has been proven from experiments that rats put on diet with less calories had less ability to resist the Flu virus.</td>
<td>It was proved by experiments the low calories eaters mouses had low defending power for flue in ratio.</td>
<td>Experimentaions have proved that those rats on less calories diet have developed a tendency of not overcoming the flu virus.</td>
</tr>
<tr>
<td>research has proven this old myth wrong that its better to fast during fever.</td>
<td>Research disproved the old axiom that &quot;It is better to fast during fever&quot;</td>
<td>The research proved this old talk that decrease eating is useful in fever.</td>
<td>This Research has proved the very old saying wrong that it is good to starve while in fever.</td>
</tr>
</tbody>
</table>
### Avoiding dieting to prevent from flu

<table>
<thead>
<tr>
<th>Avoid</th>
<th>dieting</th>
<th>to</th>
<th>prevent</th>
<th>from</th>
<th>flu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstention from</td>
<td>dieting</td>
<td>in order</td>
<td>to</td>
<td>avoid</td>
<td>Flu</td>
</tr>
<tr>
<td>Abstain</td>
<td>decrease eating</td>
<td>in order</td>
<td>to</td>
<td>escape</td>
<td>flue</td>
</tr>
<tr>
<td>quit</td>
<td>dieting</td>
<td>to</td>
<td>be</td>
<td>safer</td>
<td>from</td>
</tr>
</tbody>
</table>

In order to be safer from flu quit dieting.
<table>
<thead>
<tr>
<th>Avoiding dieting to prevent from flu</th>
<th>abstention from dieting in order to avoid Flu</th>
<th>Abstain from decrease eating in order to escape from flue</th>
<th>In order to be safer from flu quit dieting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avoid</td>
<td>dieting</td>
<td>to prevent from flu</td>
<td></td>
</tr>
<tr>
<td>Abstention from dieting in order to avoid Flu</td>
<td>in order to avoid Flu</td>
<td>Abstain from decrease eating in order to escape from flue</td>
<td>In order to be safer from flu quit dieting</td>
</tr>
<tr>
<td>Abstain decrease eating in order to escape from flue</td>
<td>to escape from flue</td>
<td>Abstain from decrease eating in order to escape from flue</td>
<td>In order to be safer from flu quit dieting</td>
</tr>
<tr>
<td>quit dieting</td>
<td>to be safer from flu</td>
<td>Abstain from decrease eating in order to escape from flue</td>
<td>In order to be safer from flu quit dieting</td>
</tr>
</tbody>
</table>
Bleu Scores (ORACLE selection)

- LDC professional translation: 0.42
- Turker average: 0.28
- best sentence Oracle: 0.44
- best Turker Oracle: 0.41
Quality Control Model

• Model assigns a score to translations based on weighted features

• 4 groups of features
  – Sentence features
  – Turker features
  – Ranking featured (based on second pass vote)
  – Calibration feature (Bleu against professionals)

• Weights are set on a dev set of professional translations
Quality Control Model

• Sentence Features
  – Language model probability
  – Ratio of source / target sentence lengths
  – Web n-gram match percentage
  – Translation edit rate to other translators
Quality Control Model

• Worker Features
  – Aggregate of sentence feature scores
  – Self-reported language abilities (Is native speaker? How long speaking?)
  – Worker location (Pakistan? India?)
Vote for the best translation

Please read the sentences and vote on the one that you think is the best in each group. The sentences are translations that were produced by people who are not native English speakers. Their translations are often ungrammatical, misspelled, disfluent, or bad in other ways. Your goal is to try to pick the best translation among the set. The one that you choose as the best will be forwarded on for editing, and it will undergo a variety of quality control mechanisms before it is published.

You should consider the following factors when selecting one translation as the best:

- Does it make more sense than the others?
- Is the English reasonably good?
- Do the grammar and spelling require only minimal correction?

- Experiments have proved that those rats on a less calories diet have developed a tendency of not overcoming the flu virus.
- It has been proven from experiments that rats put on diet with less calories had less ability to resist the Flu virus.
- Experiments proved that mice on a lower calorie diet had comparatively less ability to fight the flu virus.
- It was proved by experiments the low calories eaters mouses had low defending power for flue in ratio.

- The research proved this old talk that decrease eating is useful in fever.
- Research disproved the old axiom that "It is better to fast during fever"
- Research has proven this old myth wrong that its better to fast during fever.
- This Research has proved the very old saying wrong that it is good to starve while in fever.

- According to the scientist a patient should eat more while in fever.
- According to scientists, eat a lot during fever.
- Eat and drink more in fever according to scientists.
- According to the scientists one should eat a lot during fever.

- According to the Researchers of the State of Michigan University diet taking rats have less strength to fight with the infection and also the chances of death increased whereas earlier those rats on common diet had more strength to fight with the flu virus.
Quality Control Model

• Ranking features
  – Average rank: averaged over the five rank labels for this translation
  – Is-Best percentage
  – Is-Better percentage
Baseline selection method

- Turker average (random selection): 0.28
- Consensus Translation (lowest TER): 0.34
- Second pass selection by other Turkers: 0.37
Quality control model

- LDC professional translation: 0.42
- Sentence features alone: 0.35
- Turker features alone: 0.36
- Rank features alone: 0.37
- Calibration features alone (10% of LDC data): 0.38
- All features: 0.39
% of LDC data used for calibration

- Calibration only
- 10% calibration + other features

Graph showing the percentage of LDC data used for calibration with a trend line indicating an increase from 0.37 to 0.40.
Post-Editing

• The translations were produced by non-native English speakers
• Pretty obvious that they have spelling and grammar errors
  – US-based Turkers to the rescue!
• Each translations was post-edited by 3 other Turkers
• How much of a difference did it make?
Sentence features alone
Turker features alone
Rank features alone
Calibration features alone (10% of LDC data)
All features
Why didn’t editing help more?

- Post-editing improves results by ~1 Bleu point
- Because there are many easy-to-correct grammatical, spelling errors, we should get more
- One problem is the design of the editing HIT
  - Free form text box
- We have no good mechanism for quality control
  - Turkers tend to be lazy and make few/no corrects
  - So taking a consensus of the edits doesn’t work
A better editing HIT

• Include controls in the editing HIT to distinguish good/bad editors

• Create a sentence with a known set of corrections:
  – Start with a grammatical English sentence
  – Make several transformations to it
    • Break subject-verb agreement / change prepositions / etc

• Measure how many transformations are fixed

• Easier if we require structured corrections
Sri Lanka's forest region was destroyed by agriculture, wooden works, veterinary feeds, etc.,

several commissions where created to protect the remaining forest region

Sri Lanka is considered as the bird 's sanctuary place.

For further information please see the article on bird sanctuary rights in Indian Subcontinent

There is thousand of animals living in Sri Lanka which includes several Sri Lanka originated animals.

When we compare the area of Sri lanka 's Island, birds are highly found here.
Sri Lanka's forest region was destroyed by agriculture, forestry, animal grazing, etc.

Wooden works, forestry, veterinary feeds, animal grazing

Several commissions were created to protect the remaining forest region

Several were

Sri Lanka is considered to be a bird's sanctuary.

As to be

The a place

For further information please see the article on bird sanctuary rights in the Indian subcontinent.

Indian Subcontinent, the Indian subcontinent

There are thousands of animals living in Sri Lanka which includes several Sri Lanka native species.
Quality wrap up

• Crowdsourced translations can reach high quality after quality control
  – Gather redundant translations
  – Calibrate against small amount of professional translations
  – Do second passes over the data where other Turkers select best translations
  – Post-edit the non-native translations
    • In the future: explicitly deal with ESL
Cost

• Single-pass Mechanical Turk translation is 60 times less expensive than professional translation
• I pay ~$1 per 10 sentences = $0.005 / per word
• Additional costs:
  – 4x redundancy on translation
  – Second-pass selection
  – Editing by native English speakers
• Mechanical Turk is still an order of magnitude cheaper than professionals
Cost

- Paid $1 to translate 10 sentences, $0.25 to edit them, and $0.15 to rank them.

<table>
<thead>
<tr>
<th>Task</th>
<th>Done once</th>
<th>w/ redundancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation</td>
<td>$179.20</td>
<td>$716.80</td>
</tr>
<tr>
<td>Editing</td>
<td>$44.75</td>
<td>$447.50</td>
</tr>
<tr>
<td>Ranking</td>
<td>$26.88</td>
<td>$134.40</td>
</tr>
</tbody>
</table>

- With Amazon’s 10% fee our total <$1,500 for 7k translations, 17k edited sentences, 35k ranks.

- For 10% calibration data, cost increases by $1k.
Cost Reduction

• Current pipeline is dominated by cost of professional translation used for calibration
  – Can we reduce this by using a small % of that data?
  – Can we further reduce by using 1 ref instead of 4?
  – Can we get away with no calibration at all?
  – Does calibration data have to come from the current test set? Or can we just have one calibration set per language for all time?
Cost Reduction

• In terms of the crowdsourced components, the most costly part is redundancy in translation
  – Do we need redundant translations of every sentence?
  – Can we tell whether one translation is “good enough”?
  – I.e. can we predict if we solicit another translation whether it will be better than the ones we have?
  – Can we identify good Turkers and incentivize them to do more, so that we can reduce redundancy?
Cost Variance

• Some language pairs may be more costly than others
  – Some countries have higher costs of living than others
  – Some languages are not as well represented on Mechanical Turk (classic supply and demand)

• Speed / quality may be affected by payment
Less Commonly Taught / Low Resource Languages

• I would like to use this as a resource for creating data for low resource languages
• Low resource languages are ones where we do not have much parallel data
• Therefore translating statistical machine translation models is problematic
Volume of training data

1000M

French-English
10^9 word webcrawl

200M

DARPA GALE Program

50M

European Parliament

1.5M

Urdu
Feasibility study
Feasibility study
Volume Study

Translation Volume: total English words after k days
Quality v quantity when training statistical models

• Investing more money results in higher quality translation, but we get fewer items
• When training a statistical model, how much should we spend to ensure high quality translations?
• How robust to noise are our models?
• ASR experiments suggested not worth worrying too much about quality
Comparing Cost of Reducing WER

<table>
<thead>
<tr>
<th>System WER</th>
<th>Cost per Hour of Transcription (log scale)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absurdity</td>
<td>10000 vs. 1000 vs. 100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WER</th>
<th>Quality</th>
<th>Quantity</th>
<th>Quality</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>high</td>
<td>professional transcription</td>
<td>mturk transcription, no QC</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>high</td>
<td>professional transcription</td>
<td>mturk transcription, no QC</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>high</td>
<td>professional transcription</td>
<td>mturk transcription, no QC</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>high</td>
<td>professional transcription</td>
<td>mturk transcription, no QC</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>high</td>
<td>professional transcription</td>
<td>mturk transcription, no QC</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>high</td>
<td>professional transcription</td>
<td>mturk transcription, no QC</td>
<td></td>
</tr>
</tbody>
</table>

- mturk transcription, no QC
  - 20hrs = $300
  - 60hrs = $300

- mturk with QC
  - 20hrs = $3000

- high quality professional transcription

- $150/hr - Professional
- $90/hr - CastingWords
- $5/hr - Mechanical Turk
- $15/hr - Mturk w/ Oracle QC
Arabic has **different varieties**. MSA is the standardized form but there are many **distinct** regional dialects.
Translating Dialects with MSA MT

MSA:

متى سنرى هذه الثلثة من المجرمين تخضع للمحاكمة؟

AvmmtY n$wr hAl$lp mn Almjrmytn txDE llmHAKm ?

Levantine:

ايمتى رح نشوف هالشله من المجرمين بنتحاكم؟

AvmmtY rH n$wr hAl$lp mn Almirmvn bttHAKm ?

PT: Quando veremos esse grupo de criminosos serem julgados?

ES-EN: Quando esse group of criminals see Serem julgados?
ابلخاد: ما يجعل مصر والجزائر يجعل التقارب أكثر من ضروري

Article body (MSA)

أكد عبد الخير بلال خادم الممثل الشخصي لرئيس الجزائر والأمين العام لحزب جبهة التحرير الوطني ذات الأغلبية في البرلمان، أن العلاقات بين مصر والجزائر قوية ومتجذرة، على مصالح مشتركة قوية تربط البلدين.

وقال بلال خادم إنه ينانسيت ما حدث من "شبهات الصيف" حقوقاً متبادلاً كرد على عدم قرارات فريق الفيديو في نسخته الفائقة لعام 2009 لا يؤثر على علاقات الأخوة والأعمال Api وسياقات الأحداث. مشدداً على ضرورة العمل على وقف التفاوض بين البلدين على المستوى الحكومي والجماهيري والمؤسساتي، بشكل واضح، عنصرية وسلبية، بين تجارب可用于阿拉伯語。
Identify The Arabic Dialect

Help us classify Arabic text into dialects! This HIT is for Arabic speakers who understand the different local Arabic dialects (اللهجات المحلية, or the الدارجة), and can distinguish them from Fusha Arabic (المحسني).

Below, you will see a table that contains several Arabic sentences. For each one:

1. Tell us how much dialect (عامةً) is in the sentence, and then
2. Tell us which Arabic dialect it is.

As you expect, Egyptian is the اللهجة المصرية, and Iraqi is اللهجة العراقية. As for the other dialect names, Levantine (شامي) includes all dialects of بلاد الشام, Gulf (خليجي) includes all dialects of الخليج العربي, and Maghrebi (مغربى) includes all dialects of المغرب العربي.

This following map illustrates the dialect groups:

---

<table>
<thead>
<tr>
<th>Which Dialect?</th>
<th>Dialect Level</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egyptian (مصرى)</td>
<td>A bit of dialect (القليل من العامية)</td>
<td>ولعب فيها فريق هلبي جنبا الى جنب مع المنتخبات الأسيوية على أرض قطر</td>
</tr>
<tr>
<td>Iraqi (عراقى)</td>
<td>Choose level...</td>
<td>وكلف رئيس مفوضية الاتحاد الإفريقي جون بينج والأمين العام للجامعة العربية عمرو موسى بإعداد النظام الأساسي لهذا الصندوق وتحديد اهدافه وإدارته وواجه صرف هذه المساعدات وشروطها</td>
</tr>
</tbody>
</table>
Crowdsourcing Dialect ID

- **Labeled 142k comments** gathered from three online newspapers:
  - Al-Ghad (الغد), a Jordanian newspaper
  - Al-Riyadh (الرياض), a Saudi newspaper
  - Al-Youm Al-Sabe' (اليوم السابع), an Egyptian newspaper

- 59% MSA, 41% dialect

- Trained classifiers with 80-90% accuracy
Label Distribution

(a) All sources

(b) Al-Ghad

(c) Al-Riyadh

(d) Al-Youm
Automatic dialect ID

Baselines
Majority class: 59%
OOV wrt Arabic GigaWord: 66%
1 letter model trained on our data: 68%

<table>
<thead>
<tr>
<th>Training Set Size (K words)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0k</td>
<td>80</td>
</tr>
<tr>
<td>250</td>
<td>82</td>
</tr>
<tr>
<td>500</td>
<td>83</td>
</tr>
<tr>
<td>750</td>
<td>85</td>
</tr>
<tr>
<td>1000</td>
<td>86</td>
</tr>
<tr>
<td>1250</td>
<td></td>
</tr>
<tr>
<td>1500</td>
<td></td>
</tr>
<tr>
<td>1750</td>
<td></td>
</tr>
</tbody>
</table>

Word (1-gram) vs. Letter (5-graph)
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>Gloss</td>
<td>DF (w)</td>
</tr>
<tr>
<td>$w$</td>
<td>what</td>
<td>139.0</td>
</tr>
<tr>
<td>$x/y$</td>
<td>let</td>
<td>132.8</td>
</tr>
<tr>
<td>$x/lS$</td>
<td>enough</td>
<td>117.5</td>
</tr>
<tr>
<td>Al/Hky</td>
<td>the-talk</td>
<td>115.8</td>
</tr>
<tr>
<td>Endw</td>
<td>he has</td>
<td>95.3</td>
</tr>
<tr>
<td>bdy</td>
<td>I will/want</td>
<td>93.6</td>
</tr>
<tr>
<td>Aza</td>
<td>if</td>
<td>93.6</td>
</tr>
<tr>
<td>mnyH</td>
<td>good</td>
<td>93.6</td>
</tr>
<tr>
<td>$S/w$</td>
<td>little</td>
<td>92.8</td>
</tr>
<tr>
<td>$&lt;mw$</td>
<td>that</td>
<td>90.2</td>
</tr>
<tr>
<td>hAy</td>
<td>this (f.)</td>
<td>80.0</td>
</tr>
<tr>
<td>bEdyn</td>
<td>then</td>
<td>70.2</td>
</tr>
<tr>
<td>mw</td>
<td>not</td>
<td>65.5</td>
</tr>
<tr>
<td>AyS</td>
<td>what</td>
<td>63.8</td>
</tr>
<tr>
<td>bdw</td>
<td>he will/wants</td>
<td>60.3</td>
</tr>
<tr>
<td>Ebr</td>
<td>through</td>
<td>0.146</td>
</tr>
<tr>
<td>w&gt;tn</td>
<td>and-that</td>
<td>0.145</td>
</tr>
<tr>
<td>Al/slAm</td>
<td>Islam</td>
<td>0.138</td>
</tr>
<tr>
<td>tEAIlY</td>
<td>almighty</td>
<td>0.138</td>
</tr>
<tr>
<td>SlY</td>
<td>blessed</td>
<td>0.127</td>
</tr>
<tr>
<td>AldymqrATyp</td>
<td>democratic</td>
<td>0.108</td>
</tr>
<tr>
<td>Aljnp</td>
<td>the-committee</td>
<td>0.095</td>
</tr>
<tr>
<td>f&lt;n</td>
<td>(declarative)</td>
<td>0.062</td>
</tr>
<tr>
<td>AlmfAwDAAt</td>
<td>the-negotiations</td>
<td>0.038</td>
</tr>
<tr>
<td>AlmbASrp</td>
<td>the-direct</td>
<td>0.029</td>
</tr>
</tbody>
</table>

```
DF(w) \text{ def } \frac{f(w|D)}{f(w|\text{MSA})}
```

“Dialectness” factor

Extremely dialectal Levantine words

Extremely MSA words (i.e. very unlikely in dialectal context)
What is this useful for?

- **Characterizing communicants**
  - What is this writer’s native dialect?
  - Where are they from?
  - Informal relation with their interlocutor

- **Harvesting written dialect from large web crawls**
  - Useful for training dialect language models for ASR?

- **Identifying dialect sentence to then translate**
  - Training data for a statistical machine translation
Crowdsourcing Translation

- **Translated** dialect-labeled segments
  - Dialect Classification HIT: $10,064
  - Sentence Segmentation HIT: $1,940
  - Translation HIT: $32,061
  - Total Cost: $44,065
  - Num words translated: 1,516,856
  - Cost per word: 2.9 cents/word

- **121 workers** completed 20+ assignments
- **200,000 words** translated **per week**
- **Trained** BBN’s machine translation system
# Examples of Dialect Translation

<table>
<thead>
<tr>
<th>Dialect Input</th>
<th>MSA system</th>
<th>Dialect system</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EGY</strong></td>
<td>أنت بتعمّل لهعلن ولا ايه؟!!</td>
<td>You are working for a declaration and not?</td>
<td>You are making the advertisement for him or what?</td>
</tr>
<tr>
<td><strong>EGY</strong></td>
<td>نفسي اطمئن عليه بعد ما شاف الصوره دي</td>
<td>Myself feel to see this image.</td>
<td>I wish to check on him after he saw this picture.</td>
</tr>
<tr>
<td><strong>LEV</strong></td>
<td>هديك الجو كتيب ييير كيووول</td>
<td>God you the atmosphere.</td>
<td>This is why the weather is so cool</td>
</tr>
<tr>
<td><strong>LEV</strong></td>
<td>طول بالك عم نمزح</td>
<td>Do you think about a joke long.</td>
<td>Calm down we are kidding</td>
</tr>
</tbody>
</table>
Dialect versus MSA

Full details in Zbib, Malchiodi, Devlin, Stallard, Matsoukas, Schwartz, Makhoul, Zaidan, & Callison-Burch (NAACL 2012)
Value of MSA diminishes as dialect data increases

Egyptian Test Set

Value of MSA diminishes as dialect data increases.

Levantine Test Set
Dialect v. MSA takeaway

For dialect translation

1.5M Arabic Dialect-English

2 months
$44k
+6 Bleu

> 200M+

5 years
$10s of millions
no added Bleu
Implications of low cost, high quality translations for research
We want to respond to a “Surprise Language”

The Los Angeles Times reported that at about 5:20 P.M. on Tuesday March 4, 2003, a bomb concealed in a backpack exploded at the airport in Davao City, the second largest city in the Philippines. At least 23 people were reported dead, with more than 140 injured, and President Arroyo of the Philippines characterized the blast as a terrorist act.

With the 13 hour time difference, it was then at 4:20 A.M on the same date in Washington, DC. Twenty-four hours later, at 4:13 A.M. on March 5, participants in the Translingual Information Detection, Extraction and Summarization (TIDES) program were notified that Cebuano had been chosen as the language of interest for a “surprise language” practice exercise that had been planned quite independently to begin on that date. The notification observed that Cebuano is spoken by 24% of the population of the Philippines and that it is the lingua franca in the south Philippines, where the event occurred.
100 languages

• Microsoft translator does 35 languages
• Google does 57 languages
• The DoD’s Center for Applied MT does 64
• There is not enough data to reach acceptably high quality in new languages
• I want us to do 100 languages
Better Representation of Languages

All Languages

- SOV: 40%
- SVO: 36%
- VSO: 14%
- VOS: 7%
- No dominant order: 13%

SMT Languages

- SOV: 13%
- SVO: 61%
- VSO: 22%
- VOS: 4%
- No dominant order: 61%
Questions?