The Promise of Crowdsourcing

Fast  Cheap  Scalable  Good enough

Chris Callison-Burch
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

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

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

<table>
<thead>
<tr>
<th>Postmark City:</th>
<th>Postmark State:</th>
<th>Postmark Date:</th>
<th>Postmark Year:</th>
<th>Stamp:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(or Country)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Postmark State: Massachusetts

Postmark Date: Oct 12, 1909

Stamp: (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
Snow, O'Connor, Jurafsky and Ng’s EMNLP 2008 paper pioneered the use of Mechanical Turk for NLP

- **Affect Recognition**
  - fear(“Tropical storm forms in Atlantic”) > fear(“Goal delight for Sheva”)

- **Word Similarity**
  - sim(lad, boy) > sim(rooster, noon)

- **Textual Entailment**
  - if “Microsoft was established in Italy in 1985” then “Microsoft was established in 1985”?

- **Word Sense**
  - “a bass on the line” v. “a funky bass line”

- **Temporal Annotation**
  - ran happens before fell in: “The horse raced past the barn fell.”
Combine non-expert judgments for high correlation with experts

Cost to gather 21,000 labels for five NLP tasks

Weight w/ small amount of gold standard data for better results

$25,820
In 1994, the remains of early human beings who were believed to be eight hundred thousand years old were discovered who were given the name homo antecessor meaning the 'founder man'. Prior to this, the six hundred thousand years old man, called homo heidelebergenisis in scientific terms, was believed to be the earliest resident of this area.
Professionals versus Turkers*

*Without Quality Control

- LDC professional translation: 42
- Machine Translation: 27
- Turker average: 28
<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>It 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>Experimentations 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>
<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>Experimentations 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>Avoid</td>
<td>dieting</td>
<td>to</td>
<td>prevent from flu</td>
</tr>
<tr>
<td>Abstention from dieting</td>
<td>dieting in order to avoid Flu</td>
<td>to</td>
<td>avoid Flu</td>
</tr>
<tr>
<td>Abstain decrease eating</td>
<td>dieting in order to escape flue</td>
<td>to</td>
<td>escape flue</td>
</tr>
<tr>
<td>quit</td>
<td>dieting</td>
<td>to be in order to be safer from flu</td>
<td></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><strong>Avoid</strong></td>
<td>dieting</td>
<td>to</td>
<td>prevent</td>
</tr>
<tr>
<td>Abstention</td>
<td>from dieting</td>
<td>in order</td>
<td>avoid</td>
</tr>
<tr>
<td>Abstain</td>
<td>decrease eating</td>
<td>in order</td>
<td>escape</td>
</tr>
<tr>
<td>quit</td>
<td>dieting</td>
<td>to</td>
<td>be</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>safer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>from</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>flu</td>
</tr>
</tbody>
</table>
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

• **Turker features**
  – Aggregate of sentence feature scores
  – Self-reported language abilities
    (Is native speaker? How long speaking?)
  – Worker location (Pakistan? India?)

• **Ranking features** (based on second pass vote)

• **Calibration feature** (Bleu against professionals)
Professional Quality from Non-Professionals

LDC professional translation: 42
Turker translation w/o quality control: 28
Turker translation w/ QC model: 39
best sentence Oracle: 44
best Turker Oracle: 41

Full details in Zaidan and Callison-Burch (ACL 2011a)
Does it scale? A friendly wager.

Surely it can’t scale. I bet you can’t spend my money fast enough.

• Ken Church bet me $50k that I couldn’t spend $50k on MTurk in two months.
Fast, Scalable and Cheap

Translation Volume: total English words after k days

MTurk translation is 10-60 times less expensive than professional translation
On-Demand Worldwide Workforce

Census of bilinguals on Mechanical Turk by translating 10,000 words in 100 languages into English

- Hindi (352), Tamil (298), Malayalam (282), Spanish (215), Italian (157), Telugu (139), French (136), Malay (128), Kannada (126), Romanian (110), Portuguese (109), German (109), Arabic (106), Marathi (95), Tagalog (92), Bengali (92), Nepali (91), Urdu (89), Polish (89), Russian (86), Cebuano (83), Serbo-Croatian (78), Bulgarian (77), Bosnian (75), Gujarati (73), Croatian (72), Turkish (70), Serbian (69), Punjabi (66), Irish (68), Chinese (64), Macedonian (63), Czech (63), Dutch (60), Swedish (59), Danish (56), Swahili (51), Slovenian (49), Hungarian (48), Catalan (45), Slovak (40), Finnish (39), Indonesian (38), Hebrew (36), Vietnamese (35), Albanian (35), Norwegian (29), Ukrainian (27), Korean (26), Japanese
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 التلة من المجرمين تخضع للمحاكمة؟

mtY snrY h*h Alvlp mn Almjrmyn txDE IlmHAKmp ?

MT: When will we see this group of offenders subject to a trial?

Levantine:

ايمتى رح نشوف هالشلة من المجرمين بنتحاكم؟

AvmtnY rH n$wf hAl$lp mn Almirmyn bttHAKm ?

MT: Aimit suggested Ncov Halclp Btaathakm of criminals?

PT: Quando veremos esse grupo de criminosos serem julgados?
ES-EN: **Quando esse** group of criminals see **Serem julgados**?
بلملاد: ما يجمع مصر والجزائر يجعل التقارب أكثر من ضروري
الجمعة، 16 يونيو 2011 - 09:27

أكد عبد العزيز بلملاد، الممثل الشخصي لرئيس الجزائر وandatory، أن العلاقات بين مصر والجزائر قديمة ومتجذرة علاوة على مصالح متبادلة.

وقال بلملاد إنه بحاجة لمراجعة الأشياء في أحلامه في كل الأوقات لجذب جمهور العالم العربي.

الجزائر إذًا، أزعم لنا وأعلا ومحبباً بالعلاقات المتفوقة.

الجزائر لا تعرف الإغواء، وفي هذا السياق، يتعين أن نجح في توجيه العالم العربي إلى الأمام.

أعتذر عن أخطائي، حيث كانت مطالبة للعديد من مشاكل العرب.

أعتذر عن أخطائي، حيث كانت مطالبة للعديد من مشاكل العرب.

أعتذر عن أخطائي، حيث كانت مطالبة للعديد من مشاكل العرب.

أعتذر عن أخطائي، حيث كانت مطالبة للعديد من مشاكل العرب.

أعتذر عن أخطائي، حيث كانت مطالبة للعديد من مشاكل العرب.
Crowdsourcing Dialect Training Data

- **Labeled 142k comments** gathered from online newspapers from Saudi Arabia, Jordan, Egypt
- 59% MSA, 41% dialect
- Trained classifiers with 85-91% accuracy

- **Translated** dialect-labeled segments
- 1.5M Arabic words / 2.2M English word
- Trained BBN’s machine translation system

Full details in Zaidan and Callison-Burch (ACL 2011b)
Dialect versus MSA

Egyptian Test Set

- MSA: 14.8
- Dialect: 20.7 (+5.9)
- MSA+Dialect: 20.1

Levantine Test Set

- MSA: 12.3
- Dialect: 19.3 (+7)
- MSA+Dialect: 19.1

Full details in Zbib, Malchiodi, Devlin, Stallard, Matsoukas, Schwartz, Makhoul, Zaidan, & Callison-Burch (NAACL 2012)
## 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>انت بتعم لله اعـلان ولا ياـه ؟!!</td>
<td>You are working for a declaration and not?</td>
<td>You are making the advertisement for him or what?</td>
<td>Are you promoting it or what?!</td>
</tr>
<tr>
<td>نفسني اطمئن عليه بـعد 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>
<td>I want to be sure that he is fine after he saw the images.</td>
</tr>
<tr>
<td>لاهيك الجو كنتيير كفوول</td>
<td>God you the atmosphere.</td>
<td>This is why the weather is so cool</td>
<td>This is why the weather is so cool</td>
</tr>
<tr>
<td>طول بالاك عم نمـح</td>
<td>Do you think about a joke long.</td>
<td>Calm down we are kidding</td>
<td>Calm down, we are only kidding</td>
</tr>
</tbody>
</table>
Crowdsourcing Takeaways

For dialect translation

1.5M
Arabic Dialect-English

2 months
$44k
+6 Bleu

> 200M+

Arabic Dialect-English

5 years
$10s of millions
no added Bleu
Crowdsourcing Takeaways

• Provides a fast, cheap, scalable way of collecting data for a huge range of languages

• Annotators are non-experts, but can do surprisingly complex tasks
  – Give clear, simple instructions & create small subtasks

• Quality control is paramount
  – Embed gold standard data
  – Exploit redundancy
Ominous alternate title (suggested by colleague):

“Why the third decade may be the last for the LDC”

Prof. Callison-crowdsourced-doom-Burch
Individual researchers now have their own data production companies

My personal budget for MTurk is likely to be $200k+ over the next 2 years. I expect to be able to create several LDC quality data sets.

- Parallel corpora for six Indian languages
- Arabic dialect ID dataset
- 1.5M word English-Arabic dialect corpus
- Bilingual dictionaries for ~100 languages
- English translation of 180 hrs of spoken Spanish
- Corpus annotating and correcting ESL errors
- Speech collection and transcription for new lang.
If individuals like me can do that...

- What role is left for the LDC?
- Will it start losing its contracts?
- How do you compete with agile outfits that can do the job for pennies on the dollar?

- Lots of re-duplication of effort
- Quality control is extremely difficult to get right
How does the LDC stay relevant?

• Become the CrowdFlower for language annotation
  – Create standardized annotation pipelines / quality controls / data formats for common NLP tasks

- NER
- POS tagging
- WSD
- Transcription
- Translation
- Summarization

- Sentiment analysis
- ESL error annotation
- Coreference resolution
- Image annotation
- MT evaluation
- Reading time experiment
How does the LDC stay relevant?

• Establish a trusted workforce for many languages
  – Verify that workers actually understand the languages that they purport to speak
  – Create long-running tasks to ensure that workers always have tasks in their language
  – These tasks can be used to train annotators

• Create a self-service website
  – Groups can upload their data
  – You charge a markup over labor to cover R&D / hosting
  – Give discounts to groups that allow their results to be distributed through LDC
How does the LDC stay relevant?

• Have professional annotators play different roles than they do now
  – Doing tasks that require considerable linguistic training
  – Creating controls for non-trusted / non-professional annotators
  – Decomposing the tasks and annotation guidelines into simple steps that lay people can complete

• Start using the power of crowdsourcing to your advantage
Thanks!

• Here’s to another 20 great years!