A Framework for Conducting Non-Expert Translations and Summarizations

Christopher Harris
Department of Computer Science
SUNY Oswego

Photo by CEphoto, Uwe Aranas, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=34012183
About me

- Assistant Professor at SUNY Oswego
- Focus on Human Computation
- Research evaluates tradeoffs in using humans and computers for a variety of tasks
  - decision-making
  - knowledge creation
  - games and incentives
Other current research

- Using human computation (HC) in police car identification
- HC trains machine learning (ML) algorithms
- ML algorithms power augmented reality
- Real time decisions

The crowd identifies these features and a probability is assigned. This trains an ML algorithm.
Focus on Crowdsourcing methods

• Applying them to NLP  
  - Rare texts  
  - Low-resource languages

• Text Summarizations  
  - Children  
  - Elderly

• Transcriptions (?)
Use of the Crowd for NLP tasks

‘+summarizing +text +crowdsourcing

‘+translating +text +crowdsourcing

Google scholar
Conducted on 4/7/17
Linguistic groups of China
How can it be done?

• Crowdsourcing
  • MTurk
  • TaskCN
  • Others

• Freelancers
  • Upwork
  • Others

• Translators
How else can it be done?

- What about Edu-sourcing?
- Using students (high school and above) to perform translations and text summarizations
Objectives of this talk...

<table>
<thead>
<tr>
<th>1. Describe</th>
<th>2. Examine</th>
<th>3. Evaluate</th>
</tr>
</thead>
</table>
| a framework for crowdsourcing both translations and text summarizations | some recent empirical experiments conducted using this framework. | some design elements, including  
  • the number (depth) of crowdworkers needed for different tasks in the framework  
  • how this depth affects output quality and task completion time. |
Framework
Translations using the Crowd

A well-trodden path

- **Snow et. al. (2008)**
  - One of the first to use Mturk for translations
  - Used Majority voting
- **Callison-Burch (2009)**
  - Used crowd output to score MT translations
- **Zaidan & Callison-Burch (2011)**
  - Split up document into snippets
  - Redundancy (parallel tasks) built in
Translations using the Crowd

- **Ambati et. al. (2012)**
  - Annotations from multiple tukers
  - Examined quality vs. cost

- **Yan (2014)**
  - Two-step approach introduced
    - Translator
    - Editor
  - Relationship between the two improves reliability
Text Summarizations using the Crowd

Fewer Examples of Empirical Work
• Hourcade and Gehrt, (2015)
• used crowdworkers in a two-step process:
  • first to summarize ACOVE medication warnings
  • vote for the best summarization
Text Summarizations using the Crowd

El-Haj et al. (2010)

• used AMT to collect a corpus of single-document summaries from Wikipedia and newspaper articles in Arabic.

• Produced by extracting the most relevant sentences of the source document.
Text Summarizations using the Crowd

Buzek et al. (2010)
Mturk used to create paraphrase lattices as MT inputs.

- create the paraphrase lattices
- verify the generated paraphrases
Missing a bigger picture?

• Hard to say a technique works without considering the entire model!

• For example, consider a model:
  1. Divide document into snippets
  2. Translate
  3. Recombine snippets into document
 ...But did the recombined document lose context and flow?

• Using one translator vs. many translators
We seek a framework with the following qualities

1. **Robust:**
   - Our framework should be impervious to low-quality inputs from a malicious crowdworker.

2. **Verifiable:**
   - Should be able to perform an evaluation of outputs after each crowdworker-dependent step in our framework.

3. **Consistent:**
   - The same inputs should produce approximately the same outputs, even with different crowdworkers.

4. **Flexible:**
   - As few components as possible should rely exclusively on multi- and bilingual crowdworkers.
Benefits of a Framework in CS/NLP

• Reproducible and repeatable
• Permits critical evaluation of assumptions
• Focus on the components can be done iteratively
• Constant improvement through refinements
Need a framework
Crowdsourcing-dependent components to include in the framework....

- **Ranking:**
  - Also called voting
  - Asks crowdworkers to place text in order of relative preference.
  - Helpful in situations where users have few choices and can clearly discriminate between the choices.
  - Can use a single-winner technique (e.g., Borda counting) or a multi-round technique

- **Scoring:**
  - Also called rating
  - Asks crowdworkers to provide a score to each text on a Likert scale.
  - Preferable when there are too many choices available to the worker to determine a clear relative preference.
Other components to include in the framework....

- **Translation/Summarization:**
  - Core essential component
  - Translated/summarized versions of the input text are generated

- **Disassembly/Reassembly:**
  - Divide a document (or set of documents) into snippets
  - Recombine the translated/summarized segments into a single document
  - Usually done through automation
Introducing a Framework

Conceptual View
1. Original Document
2. Disassembly to snippets
3. Translation**
4. Ranking/Scoring**
5. Document Reassembly (Pre-smoothing)
6. Smoothing**
7. Ranking/Scoring**
8. Final Document (Post-smoothing)

Componentized View

(A) $t_n$ $t_{n(2)}$ ... $t_{n(m)}$
(B) $t_{n(1),r-1}$ $t_{n(1),r-2}$ ... $t_{n(1),r-p}$
(C) $t_{A(1)}'$ $t_{B(1)}'$ ... $t_{n(1)}'$

** = crowd-assisted
Framework Elements Previously Explored in CS/NLP

1. Original Document
2. Disassembly to snippets
3. Translation**
4. Ranking/Scoring**
5. Document Reassembly (Pre-smoothing)
6. Smoothing**
7. Ranking/Scoring**
8. Final Document (Post-smoothing)

** = crowd-assisted

- Zaidan & Callison-Burch (2011)
- Buzek et al. (2010)
- Snow et. al. (2008)
- Yan (2014)

??
Framework Components

1. Original Document
2. Disassembly to snippets
3. Translation**
4. Ranking/Scoring**
5. Document Reassembly (Pre-smoothing)
6. Smoothing**
7. Ranking/Scoring**
8. Final Document (Post-smoothing)

** = crowd-assisted
Framework: Disassembly

1. Original Document
2. Disassembly to snippets
3. Translation**
4. Ranking/Scoring**
5. Document Reassembly (Pre-smoothing)
6. Smoothing**
7. Ranking/Scoring**
8. Final Document (Post-smoothing)

** = crowd-assisted
Framework: Translation

1. Original Document
2. Disassembly to snippets
3. Translation**
4. Ranking/Scoring**
5. Document Reassembly (Pre-smoothing)
6. Smoothing**
7. Ranking/Scoring**
8. Final Document (Post-smoothing)

** = crowd-assisted
Preliminary studies have shown that once a child confesses to their parents, they are often held in disbelief.
Preliminary studies have shown that once a child confesses to their parents, they are often held in disbelief
Framework: Ranking or Scoring Translated Alternatives

1. Original Document
2. Disassembly to snippets
3. Translation
4. Ranking/Scoring
5. Document Reassembly (Pre-smoothing)
6. Smoothing
7. Ranking/Scoring
8. Final Document (Post-smoothing)

** = crowd-assisted
Preliminary studies have shown that once a child confesses to their parents, they are often held in disbelief

1. Preliminary studies have shown that once a child confesses to their parents, they tend to be skeptical

2. Preliminary studies show that once a child confesses to their parents, they tend to not believe it

3. Initial studies show that once children are confessed to their parents, they often hold doubt

4. Preliminary studies have shown that once children are confessed to their parents, they are often incredible

5. Initial studies have shown that once children are confessed to their parents, they are usually skeptical

6. Preliminary studies show that once a child is confused as a parent, most of them will not be trusted

* = two identical translations were returned
Preliminary studies have shown that once a child confesses to their parents, they are often held in disbelief

1. Preliminary studies have shown that once a child confesses to their parents, they tend to be skeptical*

2. Preliminary studies show that once a child confesses to their parents, they tend to not believe it

3. Initial studies show that once children are confessed to their parents, they often hold doubt

4. Preliminary studies have shown that once children are confessed to their parents, they are often incredible

5. Initial studies have shown that once children are confessed to their parents, they are usually skeptical

6. Preliminary studies show that once a child is confused as a parent, most of them will not be trusted

Normally we set a selection depth of 1 (winner take all), but if we set a selection depth > 1 (above we have n=3), we can actually have multiple translation-ranking cycles for each snippet

* = two identical translations were returned
Framework: Reassembly

1. Original Document
2. Disassembly to snippets
3. Translation**
4. Ranking/Scoring**
5. Document Reassembly (Pre-smoothing)
6. Smoothing**
7. Ranking/Scoring**
8. Final Document (Post-smoothing)

** = crowd-assisted
Framework: Smoothing/Editing

1. Original Document
2. Disassembly to snippets
3. Translation**
4. Ranking/Scoring**
5. Document Reassembly (Pre-smoothing)
6. Smoothing**
7. Ranking/Scoring**
8. Final Document (Post-smoothing)

** = crowd-assisted
Preliminary studies have shown that once a child confesses to their parents, they are often held in disbelief

Preliminary studies have shown that once a child confesses to their parents, they tend to be skeptical

Preliminary studies show that once a child confesses to their parents, they tend to not believe it

Initial studies show that once children are confessed to their parents, they often hold doubt

Preliminary studies have shown that once a child confesses to their parents, they tend to be skeptical
Each translation software has its own imperfections, so it is sometimes necessary to use more than one simultaneously.
Framework: Ranking or Scoring the Edited Alternatives

1. Original Document
2. Disassembly to snippets
3. Translation**
4. Ranking/Scoring**
5. Document Reassembly (Pre-smoothing)
6. Smoothing**
7. Ranking/Scoring**
8. Final Document (Post-smoothing)

** = crowd-assisted
Empirical Experiments
Pilot Study: Use Ranking (Voting) or Scoring (Ranking)?

• End result was similar if 7 items were to be ranked/scored
  • More than 7 → scoring
  • Fewer than 7 → ranking

• Stronger preference for ranking when:
  • Size of each snippet contained more text
  • Larger disparity between snippets
Experiment: Collections Used

English-to-Chinese translation
• Used the first 4 paragraphs of 10 randomly-selected OHSUMED articles (Hersh, 1994)

English text summarization
• Used the same 10 randomly-selected OHSUMED articles (Hersh, 1994)
• Limit of 10 words per snippet/40 words for the final smoothed summarization.
Experiment: Baselines

English-to-Chinese translation
• Professional: Paid a professional translator $186.30 for the 10 articles
• Crowd: 10 crowdworkers each translate one full document ($1.25 per document, $0.01/word) for a total cost of $12.50.

English text summarization
• Professional: $6.00 per summarization (an effective rate of $0.15/word), for a total of $60.00
• Crowd: $0.40 per summarization($0.01/word), for a total of $4.00.
Checking Quality: Language Tests

• Honeypots (obvious questions someone paying attention would notice)

• Check against known MT tools and eliminate those that are identical
  • But what if the MT tool happens to be identical to the translator?
    • Short snippets
    • Those with few variations
Known language tests, such as this one in Finnish

Mitkä sanat tarkoittavat samaa tai lähes samaa? Valitse paras vaihtoehto.

- kirjoitettu tai puhuttu sanoma
  - A. □ ? veitsi
  - B. □ ? viesti
  - C. □ ? vitsi

- auttaa esim. työssä
  - A. □ ? tukea
  - B. □ ? pukea
  - C. □ ? lukea

Experiment: Payments to the Crowd

**English-to-Chinese translation**
- Snippet translators
  - 7 translators (@$0.10 per snippet)
  - 3 rankers (@$0.05)
- Smoothers/editors
  - 7 translators (@$0.10 per document)
  - 3 rankers (@$0.05)

**English text summarization**
- Snippet text summarizers
  - 7 summarizers (@$0.10 per snippet)
  - 3 rankers (@$0.05)
- Smoothers/editors
  - 7 summarizers (@$0.10 per document)
  - 3 rankers (@$0.05)
## Experiment: Translation Results

<table>
<thead>
<tr>
<th>Translation</th>
<th>Pre-smoothing</th>
<th>Post-smoothing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EN to CH</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Google Translate API</td>
<td>--</td>
<td>32.38</td>
</tr>
<tr>
<td></td>
<td>--</td>
<td>0:01</td>
</tr>
<tr>
<td>Baseline - Professional</td>
<td>--</td>
<td>40.54</td>
</tr>
<tr>
<td></td>
<td>--</td>
<td>29:01</td>
</tr>
<tr>
<td>Baseline - CS</td>
<td>--</td>
<td>29.18</td>
</tr>
<tr>
<td></td>
<td>--</td>
<td>6:08</td>
</tr>
<tr>
<td>CS First 1</td>
<td>21.44</td>
<td>28.9</td>
</tr>
<tr>
<td></td>
<td>6:20</td>
<td>5:53</td>
</tr>
<tr>
<td>CS First 3</td>
<td>23.02</td>
<td>35.71</td>
</tr>
<tr>
<td></td>
<td>9:48</td>
<td>8:26</td>
</tr>
<tr>
<td>CS First 5</td>
<td>27.93</td>
<td><strong>38.65</strong></td>
</tr>
<tr>
<td></td>
<td>10:12</td>
<td>10:29</td>
</tr>
<tr>
<td>CS All 7</td>
<td>29.74</td>
<td><strong>39.81</strong></td>
</tr>
<tr>
<td></td>
<td>13:23</td>
<td>12:36</td>
</tr>
</tbody>
</table>
## Experiment: Text Summarization Results

<table>
<thead>
<tr>
<th>Summarization</th>
<th>Pre-smoothing</th>
<th>Post-smoothing</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>Time</td>
<td>BLEU</td>
<td>Time</td>
</tr>
<tr>
<td>TextRank*</td>
<td>--</td>
<td>--</td>
<td>34.46</td>
<td>0:01</td>
</tr>
<tr>
<td>Baseline - Professional</td>
<td>--</td>
<td>--</td>
<td>44.61</td>
<td>6:16</td>
</tr>
<tr>
<td>Baseline - CS</td>
<td>--</td>
<td>--</td>
<td>38.98</td>
<td>2:08</td>
</tr>
<tr>
<td>CS First 1</td>
<td>32.33</td>
<td>1:03</td>
<td>36.42</td>
<td>1:19</td>
</tr>
<tr>
<td>CS First 3</td>
<td>36.02</td>
<td>2:59</td>
<td>42.29</td>
<td>2:13</td>
</tr>
<tr>
<td>CS First 5</td>
<td>37.15</td>
<td>3:54</td>
<td>43.61</td>
<td>3:30</td>
</tr>
<tr>
<td>CS All 7</td>
<td>38.96</td>
<td>4:49</td>
<td>45.95</td>
<td>5:14</td>
</tr>
</tbody>
</table>

* = [http://summanlp.github.io/textrank](http://summanlp.github.io/textrank)
Design Elements
### Evaluation of Cost – Translation Task

<table>
<thead>
<tr>
<th>Number of CS workers Used for each step of translation task</th>
<th>Difference in BLEU score between professional translator and score achieved with this number of CS workers</th>
<th>Total cost of using this number of CS workers for translation</th>
<th>Difference in cost between using a professional translator ($186.30) and the CS workers</th>
<th>Amount paid for each 1 additional BLEU point using a professional over this number of CS workers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.64</td>
<td>$ 7.00</td>
<td>$ 179.30</td>
<td>$ 15.40</td>
</tr>
<tr>
<td>3</td>
<td>4.83</td>
<td>$ 11.00</td>
<td>$ 175.30</td>
<td>$ 36.29</td>
</tr>
<tr>
<td>5</td>
<td>1.89</td>
<td>$ 15.00</td>
<td>$ 171.30</td>
<td>$ 90.63</td>
</tr>
<tr>
<td>7</td>
<td>0.73</td>
<td>$ 19.00</td>
<td>$ 167.30</td>
<td>$ 229.18</td>
</tr>
</tbody>
</table>
## Evaluation of Cost – Text Summarization Task

<table>
<thead>
<tr>
<th>Number of CS workers Used for each step of summarization task</th>
<th>Difference in BLEU score between professional and score achieved with this number of CS workers</th>
<th>Total cost of using this number of CS workers for summarization</th>
<th>Difference in cost between using a professional ($60.00) and the CS workers</th>
<th>Amount paid for each 1 additional BLEU point using a professional over this number of CS workers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.19</td>
<td>$ 7.00</td>
<td>$ 53.00</td>
<td>$ 6.47</td>
</tr>
<tr>
<td>3</td>
<td>2.32</td>
<td>$ 11.00</td>
<td>$ 49.00</td>
<td>$ 21.12</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>$ 15.00</td>
<td>$ 45.00</td>
<td>$ 45.00</td>
</tr>
<tr>
<td>7</td>
<td>-1.34</td>
<td>$ 19.00</td>
<td>$ 41.00</td>
<td>$ (30.60)</td>
</tr>
</tbody>
</table>
## Evaluation of Time – Translation Task

<table>
<thead>
<tr>
<th>Number of CS workers Used for each step of translation task</th>
<th>Difference in BLEU score between professional translator and score achieved with this number of CS workers</th>
<th>Number of hours taken for translation with this number of CS workers</th>
<th>Difference in time taken, in hours between the professional and CS workers</th>
<th>Number of additional hours needed to increase the BLEU score 1 point with this number of CS workers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.64</td>
<td>12.22</td>
<td>16.80</td>
<td>0.69</td>
</tr>
<tr>
<td>3</td>
<td>4.83</td>
<td>18.23</td>
<td>10.78</td>
<td>0.45</td>
</tr>
<tr>
<td>5</td>
<td>1.89</td>
<td>20.68</td>
<td>8.33</td>
<td>0.23</td>
</tr>
<tr>
<td>7</td>
<td>0.73</td>
<td>25.98</td>
<td>3.03</td>
<td>0.24</td>
</tr>
</tbody>
</table>
## Evaluation of Time – Text Summarization Task

<table>
<thead>
<tr>
<th>Number of CS workers Used for each step of the summarization task</th>
<th>Difference in BLEU score between professional and score achieved with this number of CS workers</th>
<th>Number of hours taken for summarization with this number of CS workers</th>
<th>Difference in time taken, in hours between the professional and CS workers</th>
<th>Number of additional hours needed to increase the BLEU score 1 point with this number of CS workers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.19</td>
<td>2.37</td>
<td>3.90</td>
<td>2.10</td>
</tr>
<tr>
<td>3</td>
<td>2.32</td>
<td>5.20</td>
<td>1.07</td>
<td>2.18</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>7.40</td>
<td>(1.13)</td>
<td>(0.88)</td>
</tr>
<tr>
<td>7</td>
<td>-1.34</td>
<td>10.05</td>
<td>(3.78)</td>
<td>0.35</td>
</tr>
</tbody>
</table>
Conclusions and Future Work
We seek a framework with the following qualities:

1. **Robust:**
   - Our framework should be impervious to low-quality inputs from a malicious crowdworker.

2. **Verifiable:**
   - Should be able to perform an evaluation of outputs after each crowdworker-dependent step in our framework.

3. **Consistent:**
   - The same inputs should produce approximately the same outputs, even with different crowdworkers.

4. **Flexible:**
   - As few components as possible should rely exclusively on multi- and bilingual crowdworkers.
Conclusions

Developed a framework
  • Smoothing step really helps!
  • Found 3-5 crowdworkers can produce very good results
  • Beyond 5 crowdworkers really does not affect our results much
...but this is a small study... more needs to be done
Conclusions

From initial appearances, it is very cost- and time-effective

<table>
<thead>
<tr>
<th>Task</th>
<th>Metric</th>
<th>Relative to Professionals</th>
<th>What others have experienced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translations</td>
<td>Cost</td>
<td>1/20&lt;sup&gt;th&lt;/sup&gt; (using 5 + 3 workers)</td>
<td>1/23&lt;sup&gt;rd&lt;/sup&gt; (Harris &amp; Xu, 2011), 1/30&lt;sup&gt;th&lt;/sup&gt; (Callison-Burch, 2009)</td>
</tr>
<tr>
<td></td>
<td>Time</td>
<td>1/3&lt;sup&gt;rd&lt;/sup&gt;</td>
<td>N/A</td>
</tr>
<tr>
<td>Text Summarizations</td>
<td>Cost</td>
<td>1/4&lt;sup&gt;th&lt;/sup&gt; (using 5 + 3 workers)</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Time</td>
<td>1/6&lt;sup&gt;th&lt;/sup&gt;</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Future Work

• Examine Low Resource Languages
• Evaluate edu-sourcing
• Expand the model to new languages
• Transcriptions
• Motivation/flow/incentives/games
Thank you!

Christopher Harris

christopher.harris@oswego.edu