Towards Proofreading Using Human-based Computation

Bachelor thesis defense
by Teresa L. Holfeld
Bauhaus-Universität Weimar | May 10th, 2011
Overview

1. Motivation
   The problem

2. Human-based computation
   Our approach

3. Evaluation
   Reference data
   User interfaces
   Experiments
   Performance measures
   Results

4. Discussion
Motivation

Situation:
When writing texts, authors may commit errors.

Proofreading task:
Find these errors and provide a correction.

Problem:
Existing automatic solutions are insufficient.
Friends, family and co-workers have limited time.
Professional proofreaders are expensive.

Our approach:
Use human-based computation for error detection and error correction.
Human-based Computation

Definition:

Human-based computation (HBC) is the act of using the working power of humans and embed it in a computational environment.

Proofreading task using HBC:

Given a text, let workers on Amazon Mechanical Turk detect and correct the contained errors.
Human-based Computation

Amazon Mechanical Turk (MTurk):

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.

136,130 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

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 TurkRequester 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
Evaluation

Task:

Requirements:
- Texts containing reference errors and corrections
- User interfaces for MTurk
- Experiments (let MTurk proofread erroneous texts)
- Performance measures
Evaluation: Reference data

We need samples of erroneous English writing. Obtained error positions and corrections: gold standard.

**English learner corpora:**

“ESL 123 Mass Noun Examples” (ESL123)

123 sentences; 1,813 words

“Montclair Electronic Language Database” (MELD)

54 paragraphs; 6,659 words

**Example:**

Error: “How do you study the knowledges about computer?”

Correction: “How do you learn about computers?”
Evaluation: User Interfaces

Evaluation:
Let erroneous texts be corrected by MTurk workers.
Compare results to our reference errors and corrections.
Evaluate, which user interface works best (amongst others).

Proofreading user interfaces:
“Editing a paragraph”
“Editing a sentence”
“Annotating a paragraph”
“Editing a paragraph”:

Edit the text and correct all errors and passages with bad style.

I think everyone in the future is going to use technology to get education and would be able to save a lot of time. There are disadvantages of this technology too. When the power goes out, when your phone line doesn't work and you don't get extra help you need if you take classes online. Some people may have problem with that and they would prefer to go to traditional schools. The choice depends on individual, if they feel comfortable with classes' online or traditional schools. I would prefer online classes better because then I could best education while I am home with my family.
Evaluation: User Interfaces

“Editing a sentence”:

Original sentence:

These knowledge are extremely useful, can help us to look after the body, causes these tendency not to be able to turn the disease.

Your proofreading task:

Which type(s) of error does the original sentence contain?

Your corrected version of this sentence:
Evaluation: User Interfaces

“Annotating a paragraph”:

Please highlight the errors with your mouse.

Original text:

I think everyone in future is going to use technology to get education and would be able to save a lot of time. There are disadvantages of this technology too. When the power goes out, when your phone line doesn't work and you don't get extra help you need if you take classes online. Some people may have problem with that and they would prefer to go to traditional schools. The choice depends on individual, if they feel comfortable with classes' online or traditional schools. I would prefer online classes better because then I could best education while I am home with my family.

Your corrections:

<table>
<thead>
<tr>
<th>Corrected Text</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;education&quot;</td>
<td>X</td>
</tr>
<tr>
<td>&quot;in future&quot;</td>
<td>X</td>
</tr>
<tr>
<td>&quot;technology too&quot;</td>
<td>X</td>
</tr>
<tr>
<td>&quot;works&quot;</td>
<td>-</td>
</tr>
</tbody>
</table>

Your correction:

work

Add alternative correction.

Save
Evaluation: Experiments

Input parameters:
- User interface
  - “Editing a paragraph”
  - “Editing a sentence”
  - “Annotating a paragraph”
- Qualification requirements for workers
  - minimum approval rate
  - U.S. residency
    (none)
- Assignments per HIT*

Output parameters:
- Detected error positions
- Correction proposals

* Assignments per HIT: Number of workers proofreading the same text
# Evaluation: Experiments

<table>
<thead>
<tr>
<th>ID</th>
<th>Corpus</th>
<th>User Interface</th>
<th>Qualification</th>
<th>Assignment / HIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>ESL123</td>
<td>Editing a sentence</td>
<td>None</td>
<td>3</td>
</tr>
<tr>
<td>#2</td>
<td>MELD</td>
<td>Editing a paragraph</td>
<td>None</td>
<td>5</td>
</tr>
<tr>
<td>#3</td>
<td>MELD</td>
<td>Annotating a paragraph</td>
<td>None</td>
<td>5</td>
</tr>
<tr>
<td>#4</td>
<td>MELD</td>
<td>Annotating a paragraph</td>
<td>95% approval rate</td>
<td>5</td>
</tr>
<tr>
<td>#5</td>
<td>MELD</td>
<td>Annotating a paragraph</td>
<td>U.S. residency</td>
<td>5</td>
</tr>
<tr>
<td>#6</td>
<td>MELD</td>
<td>Annotating a paragraph</td>
<td>None</td>
<td>10</td>
</tr>
</tbody>
</table>
Evaluation: Performance Measures

Error detection:

Precision:
How many found errors were gold errors?
\[
\frac{|tp|}{|tp+fp|}
\]

Recall:
How many gold errors have been found?
\[
\frac{|tp|}{|tp+fn|}
\]

F-measure:
Harmonic mean of precision and recall

\[
\text{F-measure} = \frac{|tp||tp+fp|}{2|tp||tp+fp| + |tp+fn|}
\]
Evaluation: Performance Measures

Error **correction**:  
Gold standard correction: “This knowledge is *extremely useful*.”
Sample correction by MTurk: “This knowledge is *beneficial*.”

Levenshtein distance:  
How much has been changed?

BLEU:  
How similar is the correction to the reference correction?  
Regardless if word-order changed  
Borrowed from statistical machine translation
## Evaluation: Results

**Evaluation Results** (sample):

<table>
<thead>
<tr>
<th>Measure</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.26</td>
<td><strong>0.28</strong></td>
<td>0.21</td>
<td>0.18</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>Recall</td>
<td>0.90</td>
<td>0.76</td>
<td>0.63</td>
<td>0.83</td>
<td>0.85</td>
<td><strong>0.91</strong></td>
</tr>
<tr>
<td>F-measure</td>
<td>0.40</td>
<td>0.41</td>
<td>0.32</td>
<td>0.30</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>Mean Lev. dist.</td>
<td>24.99</td>
<td>69.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean BLEU</td>
<td>0.48</td>
<td><strong>0.67</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#1: “Editing a sentence”  
#2: “Editing a paragraph”  
#3-#6: “Annotating a paragraph”  
#4: Qualification: > 95% approval  
#5: Qualification: U.S. residency  
#6: 10 assignments / HIT
Evaluation: Results

Experiment statistics:

<table>
<thead>
<tr>
<th>Measure</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of words</td>
<td>1,813</td>
<td>2,223</td>
<td>6,659</td>
<td>6,659</td>
<td>2,223</td>
<td>2,223</td>
</tr>
<tr>
<td>Total costs [$]</td>
<td>3.68</td>
<td>3.50</td>
<td>11.00</td>
<td>12.50</td>
<td>4.70</td>
<td>9.85</td>
</tr>
<tr>
<td>Total working time [h]</td>
<td>13.7</td>
<td>8.5</td>
<td>28.1</td>
<td>28.5</td>
<td>9.7</td>
<td>16.8</td>
</tr>
<tr>
<td>Hourly rate [$]</td>
<td>0.27</td>
<td>0.41</td>
<td>0.39</td>
<td>0.44</td>
<td>0.48</td>
<td>0.59</td>
</tr>
</tbody>
</table>

#1: “Editing a sentence”
#2: “Editing a paragraph”
#3-#6: “Annotating a paragraph”

#4: Qualification: > 95% approval
#5: Qualification: U.S. residency
#6: 10 assignments / HIT

Experiment duration: < 24 h
Minimum hourly rate for professional proofreaders: ca. $30
Discussion

Findings:

- Short texts work better than long texts.
- A higher degree of freedom in editing leads to less editing.
- U.S. residency as qualification requirement leads to better results.
- A higher number of assignments per HIT leads to better results.

Added value:

- Proofreading for a small amount of money
- Shortens time for getting multiple proofreading results
- Multiple correction proposals
Discussion

Problems:

- Performance measures: agreement with reference data, not quality
- Requires additional reviewing process

Future work:

- Further performance measures
- Manual evaluation of experiment results
- Embedding into word processor
Thank you.