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:

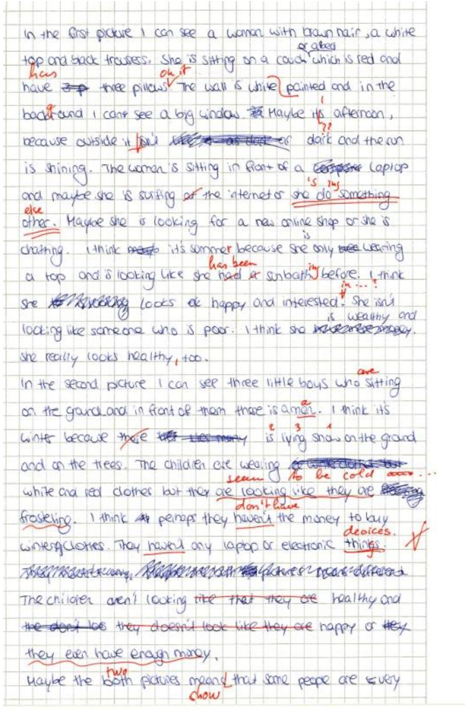
Find these errors and provide a correction.

Problem:

Existing automatic solutions are insufficient.

Friends, family and co-workers have limited time.

Professional proofreaders are expensive.



In the first picture I can see a woman with brown hair, a white
or a bed
top and black trousers. She is sitting on a couch which is red and
has on it
have ~~3 p~~ three pillows. The wall is white painted and in the
g '
backround I can't see a big window. Maybe its afternoon,
??
because outside it isn't dark and the sun
is shining. The woman is sitting in front of a ~~computer~~ laptop
's ing
and maybe she is surfing ~~at~~ the internet or she do something
else
other. Maybe she is looking for a new online shop or she is
is
chatting. I think it's sommer because she only wearing
has been ing
a top and is looking like she had ~~a~~ sunbath before. I think
in ... ?
she looks ~~a~~ happy and interested. She isn't
is wealthy and
looking like someone who is poor. I think she
she really looks healthy, too.
are
In the second picture I can see three little boys who sitting
a
on the ground and in front of them there is a men. I think it's
2 3 1
winter because ~~there~~ is lying snow on the ground
and on the trees. The children are wearing
seem to be cold ...
white and red clothes but they are looking like they are
don't have
frosteling. I think perhaps they haven't the money to buy
devices.
wintersclothes. They haven't any laptop or electronic things.
The children aren't looking ~~like that they are~~ healthy and
~~the don't loo~~ ~~they doesn't look like they are~~ happy or ~~they~~
they even have enough money.
two
Maybe the both pictures means that some peope are very
show

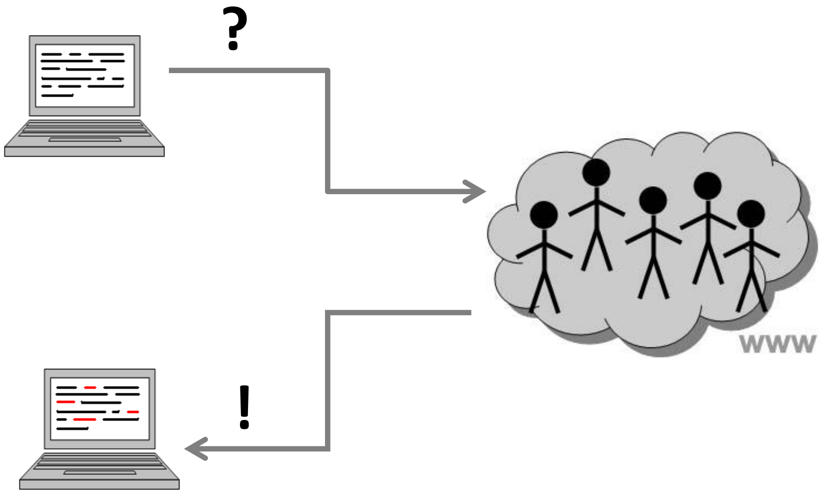
Our approach:

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.

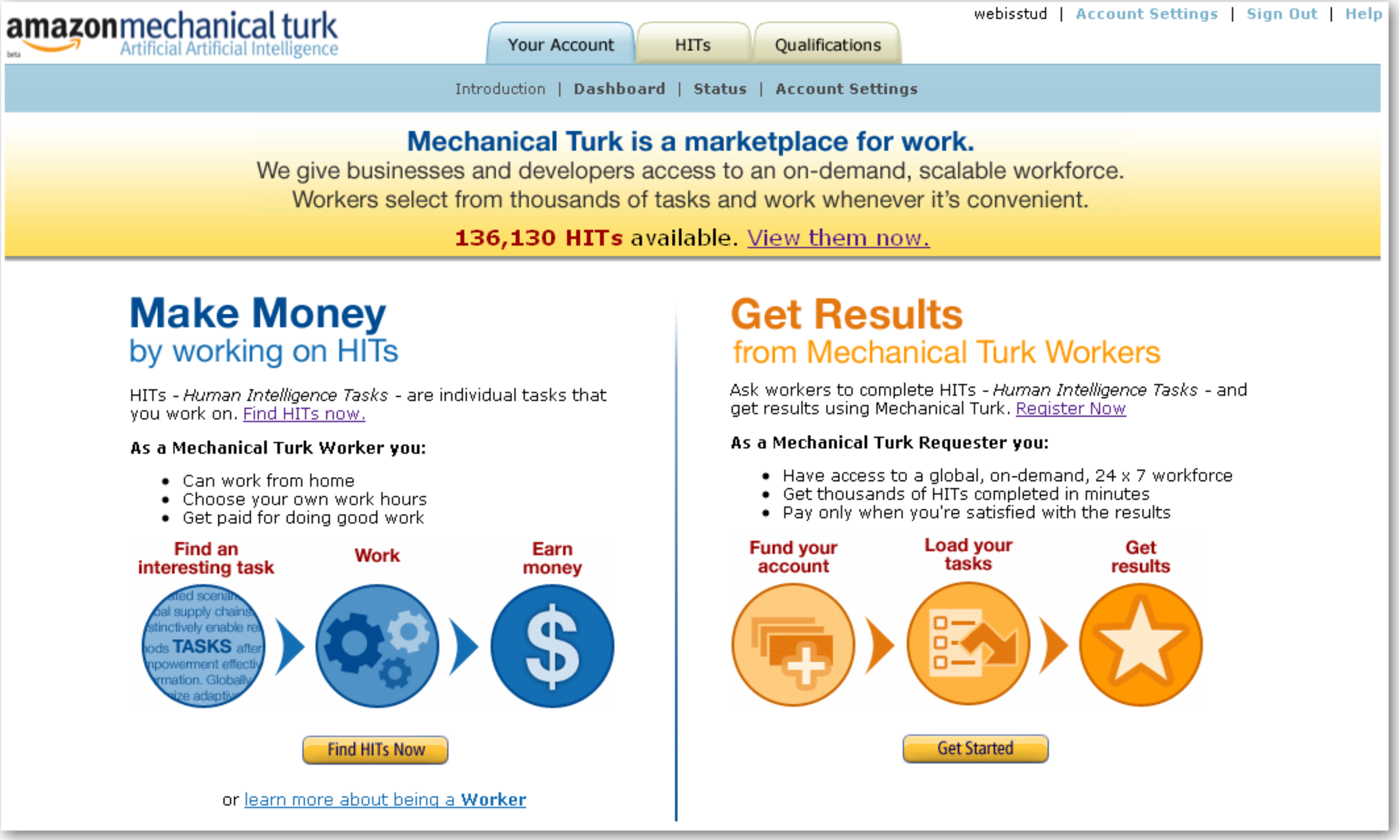


Proofreading task using HBC:

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

Human-based Computation

Amazon Mechanical Turk (MTurk):



amazonmechanical turk
beta
Artificial Artificial Intelligence
webisstud | Account Settings | Sign Out | Help
Your Account
HITs
Qualifications
Introduction | Dashboard | Status | Account Settings

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

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

Evaluation

Task:

Evaluate performance of proofreading using human-based computation.

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:

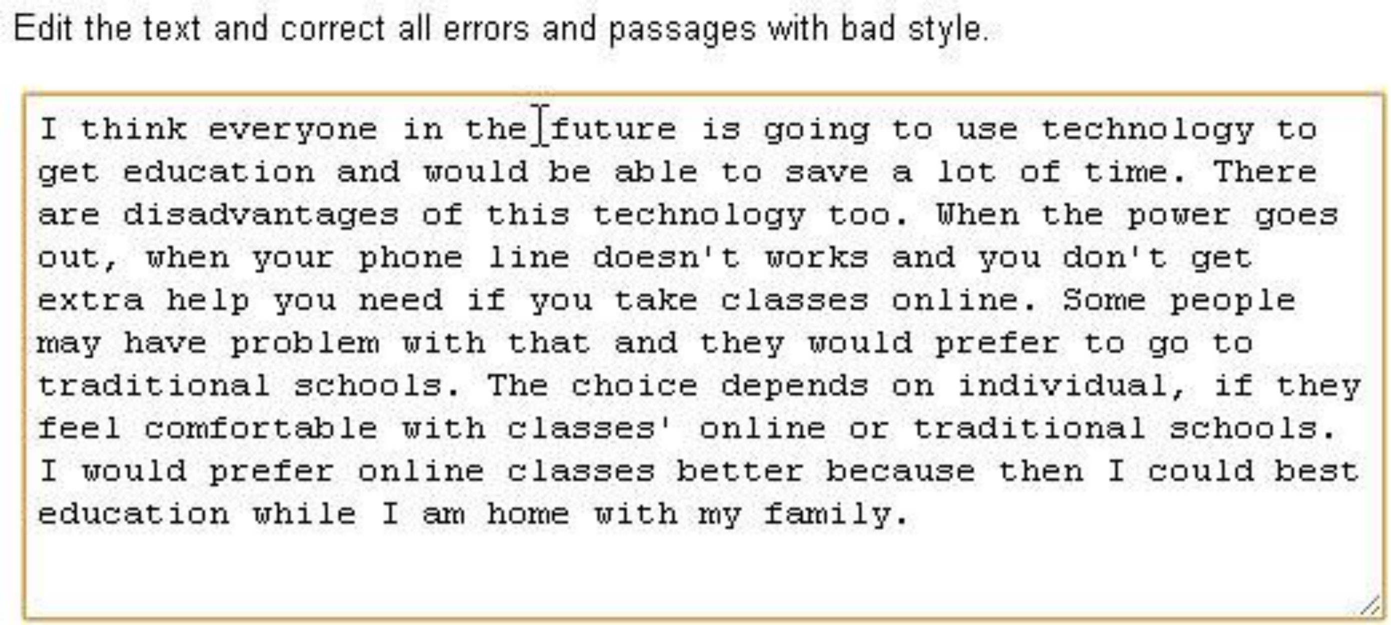
“Editing a paragraph”

“Editing a sentence”

“Annotating a paragraph”

Evaluation: User Interfaces

“Editing a paragraph”:

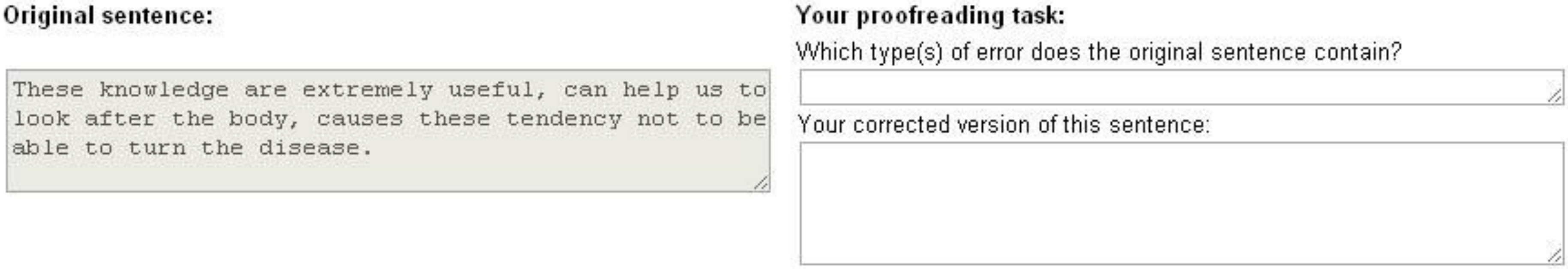


Edit the text and correct all errors and passages with bad style.

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 works 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.

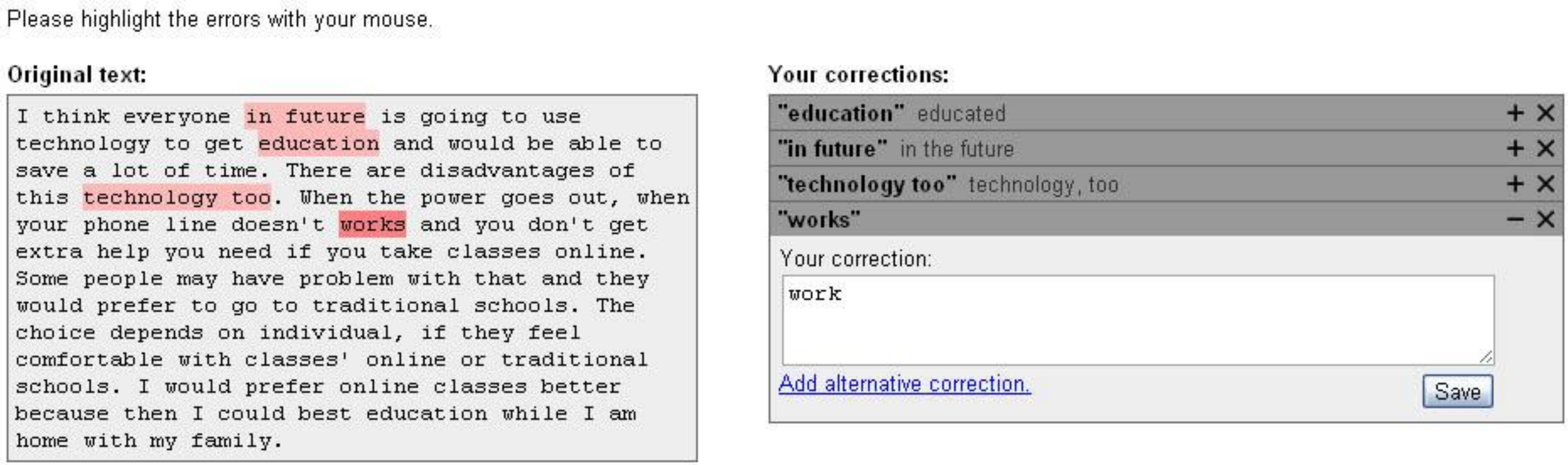
Your proofreading task:

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 works 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:

"education" educated

"in future" in the future

"technology too" technology, too

"works"

Your correction:

work

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

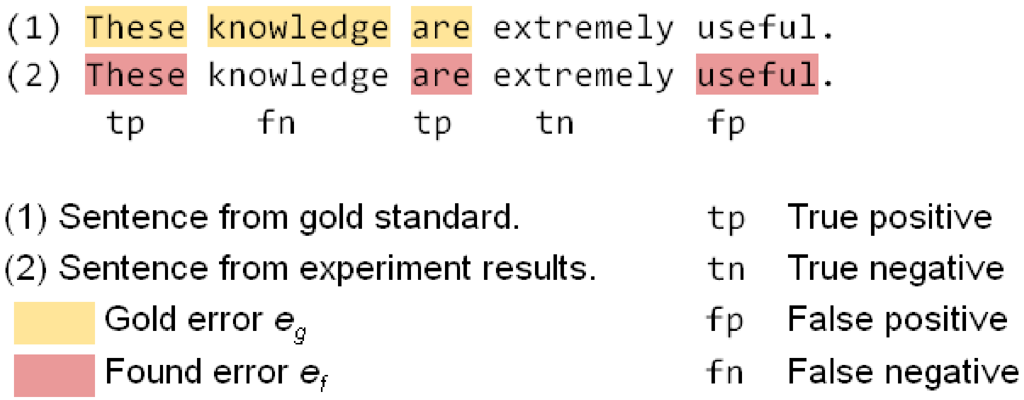
* Assignments per HIT:
Number of workers
proofreading the same text

Evaluation: Experiments

ID	Corpus	User Interface	Qualification	Assignment / HIT
#1	ESL123	Editing a sentence	None	3
#2	MELD	Editing a paragraph	None	5
#3	MELD	Annotating a paragraph	None	5
#4	MELD	Annotating a paragraph	95% approval rate	5
#5	MELD	Annotating a paragraph	U.S. residency	5
#6	MELD	Annotating a paragraph	None	10

Evaluation: Performance Measures

Error **detection:**



```
(1) These knowledge are extremely useful.
(2) These knowledge are extremely useful.
     tp       fn     tp     tn       fp
```

(1) Sentence from gold standard.

(2) Sentence from experiment results.

🟨 Gold error e_g

🟥 Found error e_f

tp	True positive
tn	True negative
fp	False positive
fn	False negative

Precision:

How many found errors were gold errors?

$$\frac{|\mathrm{tp}|}{|\mathrm{tp}+\mathrm{fp}|}$$

Recall:

How many gold errors have been found?

$$\frac{|\mathrm{tp}|}{|\mathrm{tp}+\mathrm{fn}|}$$

F-measure:

Harmonic mean of precision an recall

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):

Measure	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**
Precision	0.26	**0.28**	0.21	0.18	0.20	0.20
Recall	0.90	0.76	0.63	0.83	0.85	**0.91**
F-measure	0.40	0.41	0.32	0.30	0.33	0.33
Mean Lev. dist.	24.99	**69.15**				
Mean BLEU	0.48	**0.67**				

#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:

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

#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.