The Efficacy of Human Post-Editing for Language Translation

Spence Green

Jeffrey Heer

Christopher D. Manning

Stanford University

CHI 2013 // 29 April 2013
The man is spearing the kangaroo
The man is spearing the kangaroo.
Ngarrka-ngku ka wawirri panti-rni

man  kangaroo  spear

The man is spearing the kangaroo
Scaling up language translation

**NLP**—fully automatic translation (MT)
   Not yet human quality

**HCI**—collaborative and crowdsourced translation
   Cost-effective but slow
Scaling up language translation

NLP—fully automatic translation (MT)
   Not yet human quality

HCI—collaborative and crowdsourced translation
   Cost-effective but slow

Our work: NLP + HCI = interactive translation
NLP + HCl: Interactive translation

[Bisbey and Kay 1972]
The unforgettable actor of "Butch Cassidy and the Sundance Kid" died as a result of cancer at the age of 83 in his house in Connecticut.
Interactive MT: YouTube captions

Kenji chuting clothes

Translation: English » Chinese (Simplified)

<table>
<thead>
<tr>
<th>Time</th>
<th>English</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:00:03,610 - 00:00:10,610</td>
<td>Kenji likes chuting his own clothes</td>
<td>健二喜欢自己落下自己的衣服</td>
</tr>
<tr>
<td>00:00:00,280 - 00:00:03,280</td>
<td>Just say &quot;Chute clothes&quot;</td>
<td>只是说滑落衣服</td>
</tr>
<tr>
<td>00:00:33,280 - 00:00:34,249</td>
<td>and he'll pick up his clothes</td>
<td>他会拿起他的衣服</td>
</tr>
<tr>
<td>00:00:34,249 - 00:00:41,249</td>
<td>and then run to the laundry chute,</td>
<td>然后跑去洗衣机</td>
</tr>
<tr>
<td>00:00:41,660 - 00:00:47,760</td>
<td>laughing and giggling all the way</td>
<td>一路笑，傻笑</td>
</tr>
<tr>
<td>00:00:47,760 - 00:00:54,760</td>
<td>it's so cute!</td>
<td>它是如此的可爱</td>
</tr>
</tbody>
</table>
Does interactive MT enhance productivity?

Mixed prior results

Faster or slower?

Higher or lower translation quality?
Does interactive MT enhance productivity?

Mixed prior results

Faster or slower?
Higher or lower translation quality?

Expert translator skepticism of MT

Low quality?
You want to pay me less!?
“Advantages” of post-editing machine translation

If Mox quit translation

Life is beautiful!

???

How can you be so happy while cleaning this huge s**t?

This is nothing. In my previous career I had to clean much crappier stuff.

2012 Alejandro Moreno-Ramos

http://moxingenierotraductor.com
Our view: MT improving rapidly
This work: Post-editing user study

Simplest interactive MT: Post-editing
This work: Post-editing user study

Simplest interactive MT: Post-editing

Hypotheses:

1. Post-edit reduces translation time
This work: Post-editing user study

Simplest interactive MT: Post-editing

Hypotheses:

1. Post-edit reduces translation time
2. Post-edit increases quality
This work: Post-editing user study

Simplest interactive MT: Post-editing

Hypotheses:

1. Post-edit reduces translation time
2. Post-edit increases quality
3. Suggestions prime the translator
This work: Post-editing user study

Simplest interactive MT: Post-editing

Hypotheses:

1. Post-edit reduces translation time
2. Post-edit increases quality
3. Suggestions prime the translator
4. Post-edit reduces drafting
This work: Post-editing user study

Simplest interactive MT: Post-editing

Hypotheses:

1. Post-edit reduces translation time
2. Post-edit increases quality
3. Suggestions prime the translator
4. Post-edit reduces drafting

Exploratory and confirmatory analysis
Post-editing experimental design

Task translate an English sentence to ...
### Post-editing experimental design

<table>
<thead>
<tr>
<th><strong>Task</strong></th>
<th>translate an English sentence to ...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target languages</strong></td>
<td>Arabic, French, German</td>
</tr>
</tbody>
</table>
Post-editing experimental design

**Task**
translate an English sentence to ...

**Target languages**
Arabic, French, German

**Conditions**
Unaided and post-edit
Post-editing experimental design

<table>
<thead>
<tr>
<th>Task</th>
<th>translate an English sentence to ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target languages</td>
<td>Arabic, French, German</td>
</tr>
<tr>
<td>Conditions</td>
<td>Unaided and post-edit</td>
</tr>
<tr>
<td>Expert Subjects</td>
<td>16 per target language</td>
</tr>
</tbody>
</table>
Experimental design

Two-way, mixed design

Translation conditions (within subjects)

Source sentences (between subjects)
Experimental design

Two-way, mixed design

Translation conditions (within subjects)
Source sentences (between subjects)

Two timed translation efforts

Untimed break
Total time: about 60 min. per subject
Experimental design

Two-way, mixed design

Translation conditions (within subjects)
Source sentences (between subjects)

Two timed translation efforts

Untimed break
Total time: about 60 min. per subject

MT from Google

[March 2012]
The physicist Arthur Eddington drew on Borel's image further in The Nature of the Physical World (1928), writing: If I let my fingers wander idly over the keys of a typewriter it might happen that my screed made an intelligible sentence.
The physicist Arthur Eddington drew on Borel's image further in The Nature of the Physical World (1928), writing: If I let my fingers wander idly over the keys of a typewriter it might happen that my screed made an intelligible sentence.

Le physicien Arthur Eddington a attiré sur l'image de Borel dans le caractère du monde physique (1928), écrit: Si je laisse mes doigts se promener les bras croisés sur les touches de la machine à écrire, il peut arriver que mon chape fait une phrase intelligible.
Experimental setup: Linguistic data

Topic selections from Wikipedia

1. Flag of Japan   easy
2. 1896 Olympic Games   easy
3. Schizophrenia   hard
4. Infinite Monkey Theorem   hard

One easy, one hard per condition
It was the first international Olympic Games held in the Modern era.
The chance of their doing so is decidedly more favourable than the chance of the molecules returning to one half of the vessel.
Experimental setup: Human subjects

Expert freelance translators on oDesk

Ecological validity

Fair payment: subjects bid on job
Experimental setup: Human subjects

Expert freelance translators on oDesk

Ecological validity

Fair payment: subjects bid on job

Lots of subject data

oDesk language skills tests

Hours worked per week

Demographic information
Experimental setup: Quality rating

Same setup as annual Workshop on Machine Translation
Experimental setup: Quality rating

Same setup as annual Workshop on Machine Translation

Crowdsourced, pairwise evaluation on MTurk
Experimental setup: Quality rating

Same setup as annual Workshop on Machine Translation

Crowdsourced, pairwise evaluation on MTurk

Three judgments per translation pair
**Source:**
The Games had the largest international participation of any sporting event to that date.

**Reference:**
Les Jeux recueillirent la plus grande participation parmi tous les évènements sportifs en date.

<table>
<thead>
<tr>
<th>#</th>
<th>Translation</th>
<th>Select Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Les Jeux ont eu la plus grande participation internationale de n'importe quel évènement sportif jusqu'à cette date-là.</td>
<td>1 is better</td>
</tr>
<tr>
<td>2</td>
<td>Les jeux avaient le plus grand taux de participation internationale par rapport à n'importe quel évènement sportif de l'époque.</td>
<td>Translations are equally good</td>
</tr>
</tbody>
</table>
Results
Fixed effects fallacies

**Fixed effect**—Data includes all factor levels

Gender

Machine configuration
Fixed effects fallacies

**Fixed effect**—Data includes all factor levels

  Gender

  Machine configuration

**Random effect**—sampled levels

  Human subjects (RM-ANOVA)
Fixed effects fallacies

**Fixed effect**—Data includes all factor levels

- Gender
- Machine configuration

**Random effect**—sampled levels

- Human subjects (RM-ANOVA)
- English source sentences
- Target languages

“Language as fixed-effect fallacy” [Clark 1973]
Mixed effects models

\[ y = x^T \beta + z^T b + \eta \]

- Linear predictor
- Random effects structure
- Error term
Post-editor variance

Subject

1 2 3 4 5 6 7 8 9 10 11 12 13

Condition
- Post-Edit
- Unaided

Better → Average Rank → Worse
Recap: Experimental hypotheses

1. Post-edit reduces translation time
2. Post-edit increases quality
3. Suggestions prime the translator
4. Post-edit reduces drafting
Hypothesis #1: Reduced time

- Arabic: -43.3s (-17.8%)
- French: -40.0s (-21.2%)
- German: -41.2s (-21.0%)
Hypothesis #1: Reduced time

Post-edit reduces translation time?
Hypothesis #1: Reduced time

Post-edit reduces translation time?

Yes! $p < 0.001$

Significant covariates

Source length

% nouns in sentence
The latter may range from loss of train of thought, to sentences only loosely connected in meaning, to incoherence known as word salad in severe cases.
Source hover patterns predict time?

The latter may range from loss of train of thought, to sentences only loosely connected in meaning, to incoherence known as word salad in severe cases.

“Noun %” significant in time models
Hypothesis #2: Higher quality

Arabic

French

German

Condition
- Unaided
- Post-Edit

Post-edit rank improvement
- +18.3%
- +19.2%
- +9.3%
Hypothesis #2: Higher quality

Post-edit increases quality?
Hypothesis #2: Higher quality

Post-edit increases quality?

Yes! $p < 0.001$

Significant covariates

Source language proficiency test
Hypothesis #3: Priming

Suggestions prime the translator?

Yes! $p < 0.001$ for each language.
Hypothesis #3: Priming

Suggestions prime the translator?

Yes! $p < 0.001$ for each language

Test setup

- Edit distance to MT
- Paired $t$-test
Hypothesis #4: Less drafting

Unaided condition

Post-edit condition
Hypothesis #4: Less Drafting

Post-edit results in less drafting?
Hypothesis #4: Less Drafting

Post-edit results in less drafting?

Yes! $p < 0.01$

Post-edit condition behavior

Fewer, longer pauses

Pauses are larger % of total time
Conclusions

Simple source lexical features predict time
Conclusions

Simple source **lexical features** predict time

Post-edit → **different interaction patterns**
Conclusions

Simple source **lexical features** predict time

Post-edit → **different interaction patterns**

Suggestions **prime the translator**
Conclusions

Simple source **lexical features** predict time

Post-edit → **different interaction patterns**

Suggestions **prime the translator**

Post-edit **improves speed and quality**
The Efficacy of Human Post-Editing for Language Translation

Spence Green, Jeffrey Heer, and Christopher D. Manning

Data and code are available
vis.stanford.edu
nlp.stanford.edu
spencegreen.com