A Large-Scale Evaluation of Pre-editing Strategies for Improving User-Generated Content Translation

Violeta Seretan, Pierrette Bouillon, Johanna Gerlach, *Université de Genève*

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 288769

In the ACCEPT project…

Consortium

UNIVERSITÉ DE GENÈVE (coordinator)
acrolinx speak with one voice
THE UNIVERSITY OF EDINBURGH
Symantec.
Lexcelera innovating translation

Project Description on Website

The ACCEPT project is part of the seventh European framework programme and works on enabling machine translation for the emerging community content paradigm, allowing citizens across the EU better access to communities in both commercial and non-profit environments. There are five countries participating in the project: Switzerland, United Kingdom, Germany, Ireland, and France.

ACCEPT: Automated Community Content Editing PorTal

Seventh Framework Programme: THEME ICT-2011.4.2(a) Language Technologies

Facts

Duration: 36 months
Start date: 01/01/2012
Finish date: 31/12/2014
EU funding: €1,825,000
EU project officer: Pierre-Paul Sondag
Website: www.accept-project.eu

We want to bring machine translation closer to user communities.

Mission — *Help communities share knowledge more effectively across the language barrier.*

Use Cases

Norton™ by Symantec
support forum community (commercial use case)

Aim: Help companies engage with their customers across the language barrier.

TRANSLATORS WITHOUT BORDERS
volunteer translator community (NGO use case)

Aim: Help NGOs deliver critical information in the right language at the right time.

R&D Axes

Develop minimally-intrusive strategies for pre-editing for statistical machine translation (SMT).

Improve SMT for community content:
- domain adaptation
- linguistic back-off
- text analytics.

Develop strategies for post-editing to learn corrections and automate them (in loops).

We developed technology to pre-edit and post-edit community content:

In particular, rules for content checking based on shallow parsing.

Case	*norton/Norton*
Punctuation	*When .../When ...,*
Hyphenation	*system-based*
Spelling	*instalation*
Spaces	*environ.Hier*
Redundant words	*the the*
Homophones	*sur 'on'/ sûr 'sure'*
Word choice	*participate to/in*
Word form	*aimera/aimerait bien*
Agreement	*problem occur*
Ungram. sequences	*data the to*
Style	*C'est nickel/parfait*
SMT-specific	*have to/must*

For details: [1]

English ~50 rules
French ~90 rules

Rule Example acrolinx

```
//example: a dogs
TRIGGER(80) == @det_sg^1 [{@mod|@noun}]*! @noun_pl^2
                    -> ($det_sg, $noun_pl)
                    -> { mark : $det_sg, $noun_pl;}

//example: a dogs -> a dog
SUGGEST(10)  ==   $det_sg []* $noun_pl
       -> {   suggest: $det_sg -> $det_sg, $noun_pl ->
  $noun_pl/generateInflections([number="singular"]);
                }
```

http://attempto.ifi.uzh.ch/site/cnl2012/slides/lehmannetal_forumdata.pdf

We evaluated the rules:

We conducted a large-scale evaluation campaign with Master's students in Translation to assess the impact of pre-editing on translation quality.

1 Pre-editing

Re : Erreur 3048,3
Tu as le dernier correctif.
Attends 48h. Fais de temps en temps un LiveUpdate manuel et on verra à chaque redémarrage du PC. (…) N'aie crainte, tu es toujours protégé.
Merci de tenir au courant si le problème persiste.

Re : Erreur 3048,3
Vous avez le dernier correctif.
Attendez 48h. Faites de temps en temps un LiveUpdate manuel et on verra à chaque redémarrage du PC. (…) N'ayez crainte, vous êtes toujours protégé. Veuillez me tenir au courant si le problème persiste.

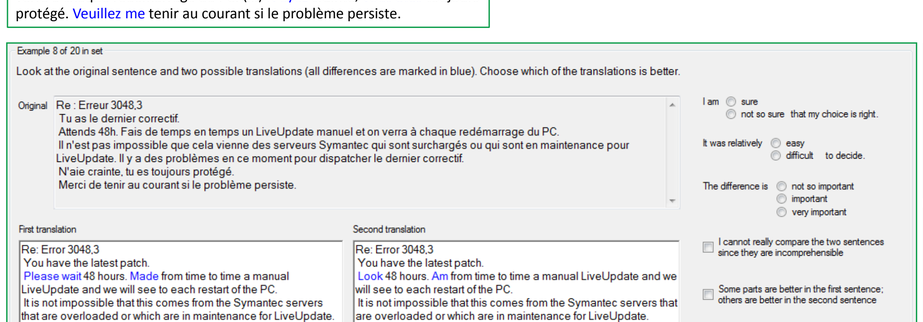
2 Translation

ACCEPT

For details: [2]

3 Evaluation

For details: [3]



Categories for hypothesis testing
better
same
worse

Fleiss *k*
French-English 0.43
English-French 0.20
English-German 0.38
(up to ***moderate***)

1000 posts avg. length (words) English **93.7** French **78.6**

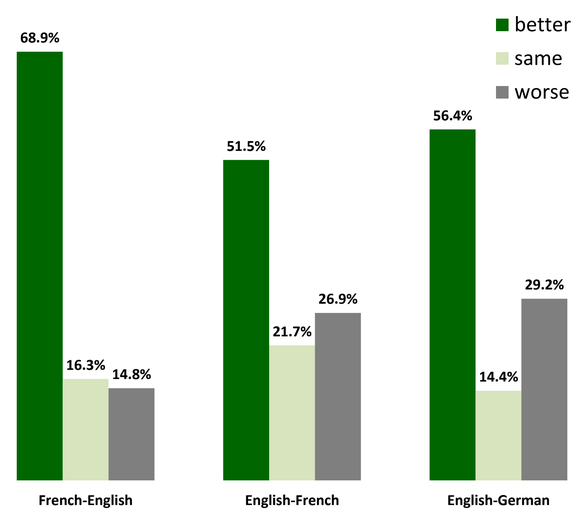
time spent for **judging**/ **feedback** avg. seconds/post – outliers removed

French-English	English-French	English-German
43 / 19	30 / 20	52 / 31

We built reference translations for a subset of the data and computed automatic metric scores.

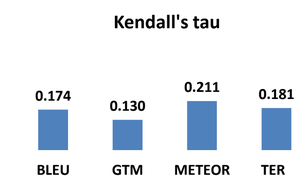
50 posts French-English — BLEU GTM METEOR TER

And this is what we found:



Pre-editing improves translation quality significantly (McNemar test, $p < 0.0001$).

But the gain in quality is **not** reflected by automatic metric scores (non-significant difference).



There is *weak/weak or no* correlation with human judgments.

Error analysis showed we need to improve named entity recognition and avoid auto spell check.

Pre-editing is worth the effort.

Pre-editing is relatively fast

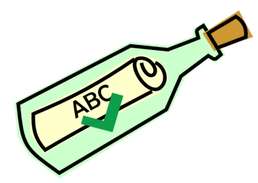
English	French
238	261

Words per minute (wpm)

and has a positive impact on post-editing, according to a related study [4]:

- The post-editing time was roughly reduced by half thanks to pre-editing.
- The processing speed increased from 28 to 52 wpm (or to 37 wpm, if we consider the time spent pre-editing).

So, pre-edit your message to reach a wider audience!



The ACCEPT plug-in & middleware can be freely tested and downloaded: **www.accept-portal.eu**

References

[1] ACCEPT deliverable D 2.2 *Definition of pre-editing rules for English and French*

[2] ACCEPT deliverable D 4.1 *Baseline MT systems*

[3] ACCEPT deliverable D 9.2.2 *Survey of evaluation results*

[4] Gerlach, J., V. Porro, P. Bouillon, S. Lehmann (2013). Combining pre-editing and post-editing to improve SMT of user-generated content. In *Proc. of Workshop on Post-editing Technology and Practice*, Nice, France.