Is Machine Translation Ready Yet?

Researcher: Dr Ignacio García (i.garcia@uws.edu.au)

Abstract

New translation memory tools and new versions of established ones offer the translator the option to post-edit machine generated text for segments for which no match is found in the memories. This research considers whether translators should take up the offer by discussing the results of tests performed by translator trainees using the Google Translator Toolkit to translate passages both from the machine-translated version and from the source text, then measuring the time taken and the quality of the translation.

Project details

Each test consisted of two passages of about 250 words, each passage to be completed within one hour. Participants were asked to translate them from the source text (the control group) and from the MT baseline (the experimental group). For the second passage, those in the control group will translate from MT and vice versa – a repeated measures design. The dependant variables to observe were time on task and quality of output. The independent variable, whether the MT option for no matches was enabled. Quality was measured according to NAATI guidelines. Departing hypotheses were, in line with current advice given on post-editing, that translating from MT seeded passages (i.e. by post-editing) will be faster, and that the quality of passages translated the traditional way (i.e. from the source text) would be higher.

Detailed results

The first set of experiments, carried out early in the second semester 2009, gathered data from 14 participants, all with Chinese as the mother tongue translating into English, doing two tests each. Tests were then subjected to blind marking by two independent markers. Results indicated that in regards to time, translating by post-editing was not significantly faster than translating from the source text - faster in 15/28 tests, i.e. in 55% of the cases ($t(27)=1.05$, two tailed $p = .304$). In regards to quality, translating by post-editing actually produced better results in 34/56 test marked – 61% (for marker 1 $t(27) = 2.84$, two-tailed $p = .008$; for marker 2 $t(27) = .60$, two-tailed $p = .552$; putting all data together this gives us a two tailed $p = .026$ thus statistically significant).

To ensure reliability, the experiment was repeated later in the semester, with another set of 14 participants and two different markers. For this second set, Time results were similar for post-editing and for working from the source text, 14/28 tests in both cases – 50% ($t(27)=.70$, two-tailed $p = .492$). This result was consistent with the previous set of data. If we put together the data for the 28 participants, 56 tests, we have PE being faster in 29 tests – 52% (two-tailed $p = .234$). Passages post-edited received a higher mark 40/56 passages (71% ($t(55) = 3.55$,two-tailed $p < .001$).

For the first semester of 2010 the experiment was repeated again, but this time for the Chinese into English direction. Another significant difference was introduced: all participants were allowed when translating to search for information on the web as they saw fit. In the English into Chinese sets and to neutralise other variables, i.e. web searching skills, this was not allowed. The number of participants that sat through the tests was 21. Tests were marked by just one examiner (the reason for this being lack of available funding).
The post-edited passages finished faster in 26/42 tests – 62% \((t(41) = 3.033, \text{two-tailed } p = .004)\). This time the difference could be considered as statistically significant. This result could have been in part helped by the fact that participants in these tests English into Chinese had the chance of doing word (and other type of) searches, which means it could have taken them comparatively longer to complete the passages from the source text. Results from post-editing were better than those from the source text in 28/42 tests – 67% \((t(41) = 2.2, \text{two-tailed } p= .002)\).

These results in regards to quality hold for best performers as well as for the weaker ones. Adding the figures from both markers, the four participants with higher marks in the first set of data English into Chinese did better by post-editing in 4/8 tests, and the participants with the lower marks, in 5/8 tests; in the second set, the four with higher marks do better in 7/8 tests, and the four with the lower in 6/8. If we were to put them together will find no difference in how best achievers do by post-editing compared with how participants with the lowest marks did. In Chinese into English data, the four best marks did better with post-editing in 6/8 tests, while the four lowest in 5/8. We could thus say that translating from MT may serve good performers as well as it does weak ones.

The findings thus disprove the two hypotheses put forward: translating from the MT baseline (i.e. post-editing) was not significantly faster than translating the traditional way; the quality of the passages translated from the MT baseline was not worse, as expected, but better, and in a statistically significant way.