Panel 22: Post-editing

Authors:

Mercedes García Martínez, Arlene Koglin, Bartolomé Mesa-Lao, Michael Carl

Title:

Automatic evaluation of machine translation: correlating post-editing effort and Translation Edit Rate (TER) scores

Abstract:

The availability of systems capable of producing fairly accurate translations has increased the popularity of machine translation (MT). The translation industry is steadily incorporating MT in their workflows engaging the human translator to post-edit the raw MT output in order to comply with a set of quality criteria in as few edits as possible. The quality of MT systems is generally measured by automatic metrics, producing scores that should correlate with human evaluation.

In this study, we investigate correlations between one of such metrics used by MT developers, i.e. Translation Edit Rate (TER), and actual post-editing effort as it is shown in post-editing process data collected under experimental conditions. Using the CasMaCat workbench as a post-editing tool, process data were collected using keystrokes and eye-tracking data from five professional translators under two different conditions: i) traditional post-editing and ii) interactive post-editing. In the second condition, as the user types, the MT system suggests alternative target translations which the post-editor can interactively accept or overwrite, whereas in the first condition no aids are provided to the user while editing the raw MT output. Each one of the five participants was asked to post-edit 12 different texts using the interactivity options provided by the system and 12 additional texts without interactivity (i.e. traditional post-editing) over a period of 6 weeks.

Process research in post-editing is often grounded on three different but related categories of post-editing effort (Krings 2001), namely i) temporal (time), ii) cognitive (mental processes) and iii) technical (keyboard activity). For the purposes of this research, TER scores were correlated with two different indicators of post-editing effort as computed in the CRITT Translation Process Database (TPR-DB)[1]. On the one hand, post-editing temporal effort was measured using *FDur* values

1 CRITT Translation Process Database: http://bridge.cbs.dk/platform/?q=CRITT_TPR-db

(duration of segment production time excluding keystroke pauses ≥ 200 seconds) and *KDur* values (duration of coherent keyboard activity excluding keystroke pauses ≥ 5 seconds). On the other hand, post-editing technical effort was measured using *Mdel* features (number of manually generated deletions) and *Mins* features (number of manually generated insertions).

Results show that TER scores have a positive correlation with actual post-editing effort as reflected in the form of manual insertions and deletions (Mins/Mdel) as well as time to perform the task (KDur/FDur).

Keywords:

post-editing
technical effort
temporal effort
automatic metrics for MT
Translation Edit Rate (TER)

Bionotes:

– Mercedes García-Martínez is a computer science engineer and a research assistant at the Center for Research and Innovation in Translation and Translation Technology, CBS (Denmark).

– Arlene Koglin is a PhD candidate in Translation Studies at the Federal University of Minas Gerais.

– Bartolomé Mesa-Lao is a freelance translator and a research assistant at the Center for Research and Innovation in Translation and Translation Technology, CBS (Denmark).

– Michael Carl is an associate professor at the Department of International Business Communication, CBS (CBS).