Pause Metrics and Machine Translation Utility

Isabel Lacruz[1] Michael Carl[2] Masaru Yamada[3] Akiko Aizawa[4]

1. Kent State University, United States
2. National Institute of Informatics, Japan and Copenhagen Business School, Denmark
3. Kansai University, Japan
4. University of Tokyo, National Institute of Informatics, Japan

1 Introduction

Traditionally, attempts to measure Machine Translation (MT) quality have focused on how close output is to a "gold standard" translation. TER (Translation Error Rate) is one standard measure that can be generated automatically. It is the normalized length of the shortest path (smallest number of edits per word) needed to convert MT output to an average of "ideal" translations (Snover et al., 2006).

MT quality has now improved so much that post-edited (or in some cases, raw) MT output is routinely used in many applications in place of from-scratch translations. Despite the translators' continued resistance to post-editing, there is increasing evidence that productivity is greater when translators post-edit rather than translate from scratch (e.g., Green et al., 2013). Machine-assisted alternatives to post-editing, such as Interactive Translation Prediction (see for example Sanchis-Trilles et al., 2014) are also making rapid advances.

Because of these changing paradigms, alternative ways of measuring MT quality are being developed. Under many circumstances, perfect accuracy is not necessary: it is enough for MT output to be "good enough." The end-user of the raw product should be able to use it with little effort, and the post-editor should easily be able to produce a satisfactory product.

MT *utility* is determined by the effect the MT output has on the actual effort expended by the user, while MT *adequacy* is determined by the anticipated demand the MT output places on the user. Adequacy has been measured by human judgments along Likert scales, as well as by automatic metrics such as TER. In the context of post-editing, TER is modified to HTER, to measure the discrepancy between MT output and the final post-edited product. Thus, HTER measures the smallest number of *necessary* edits per word during post-editing.

2 Utility and cognitive effort

A utility measure is a measure of expended effort. Krings (2001) carefully studied effort in translation tasks and identified three separate, but related effort components. The simplest to measure is temporal effort, the time taken to complete the task: more time on task indicates more effort and less productivity. Technical effort is the effort of keyboarding to make insertions or deletions characters, using a mouse to cut and paste or move around the text, and so on. It is usually measured from counts extracted from logging software: more actions imply more effort.

Cognitive effort is the mental effort of reading, planning, making decisions and reflecting on the choices made. An end-user or a post-editor working with low quality MT output makes more cognitive effort, but, unlike temporal or technical effort, this cognitive effort cannot be measured directly. However, understanding cognitive effort is key to gaining insights into the translation process and to managing mental fatigue and so productivity.

Automatic mental processes, which tend to become more prevalent as expertise develops (Göpferich et al., 2011), are essentially effortless. Conscious mental processes, on the other hand, generate cognitive effort as they draw on the limited resources of working memory (Tyler et al., 1979). Cognitive effort increases as the proportion of allocated working memory resources increases, and this manifests itself through behavioral

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

characteristics that can be measured.

The eyes do not move smoothly during reading. Instead, they fixate for periods averaging about a quarter of a second before jumping rapidly to the next fixation. Just and Carpenter's (1980) eye-mind hypothesis is that the eye fixates on what the mind is processing. There is thus a direct relationship between eye fixations and cognitive effort: longer or more numerous fixations indicate greater cognitive effort. In the past few years, eye tracking measures such as these have become important in translation process research (see, for example, O'Brien, 2011.) However, they are still relatively difficult to generate, and they do not provide insight into exactly what mental processes are engaged during eye fixations.

3 Keylogging pauses and cognitive effort

Eye fixations can be interpreted as pauses for mental processing. With this perspective it becomes interesting to lever more information by comparing eye tracking and keystroke log data. Pauses in language production are also indicators of mental processing (e.g., Schilperoord, 1996; O'Brien, 2006; Lacruz et al., 2012; Lacruz & Shreve, 2014). Accordingly, pauses between keystrokes or mouse clicks during translation or post-editing may provide information on cognitive effort.

Translators and post-editors tend to make lengthy orientation pause at specific places, such as the beginning of sentences. They also tend to make longer pauses between production units – coherent sequences of keystrokes separated by pauses shorter than a relatively short threshold – and the density of production units is known to correlate with eye tracking measures of cognitive effort, including average fixation times and average fixation counts (Daems et al., 2015).

O'Brien (2006) proposed a pause metric, the pause ratio, as a measure of cognitive effort in post-editing. For each post-editing segment, she defined pause ratio to be the total pause time in the segment divided by the total post-editing time for the segment. She predicted that, since key logging pauses iindicate cognitive effort, source text segments with linguistic characteristics known to be challenging for machine translation would yield higher pause ratios in post-editing. Surprisingly, she did not find this effect.

However, a case study (Lacruz et al., 2012) gave evidence that post-editors make clusters of relatively short pauses (between 500 ms and 2,000ms) as they work on cognitively challenging production units. While these relatively short pauses did not contribute much to O'Brien's pause ratio, they did appear to be markers of high cognitive effort, perhaps playing a monitoring role. To capture the effect of the numerous shorter pauses, Lacruz et al. proposed a modification of the pause ratio. They defined the average pause ratio (APR) for a post-edited segment to be the average pause time divided by the average post-editing time per word. They predicted that an increase in the number of production units in a post-edited segment would signal increased cognitive effort, which would result in more short pauses and so smaller APR values. Their prediction was confirmed and later replicated on a slightly larger scale in Lacruz & Shreve (2014).

Lacruz & Shreve (2014) also introduced a slightly simpler pause metric, reminiscent of HTER. They defined the pause to word ratio (PWR) for a post-edited segment to be the number of pauses in the segment divided by the number of words in the segment. PWR correlated strongly with APR, HTER, and with the density of production units in a segment. APR and PWR were promising metrics for cognitive effort in post-editing.

Later, Daems et al. (2015) found correlations between APR and eye-tracking metrics. Liu & Du (2014) demonstrated that APR was lower for Chinese-to-English translations of poems than for more routine texts. Finally, Schwartz et al. (2015) applied a known strategy for increasing post-editing accuracy by providing word-by-word alignments between the source text and the MT output. They found increased quality gains, as rated by human judgments, for the same amount of cognitive effort expended, when measured by PWR.

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

4 Comparisons across language pairs

Previous studies of the relationship between pause metrics and cognitive effort have, with the exception of Liu & Du (2014), used languages that are very similar to each other. We now describe preliminary recent work using CRITT TPR-DB (Carl et al., 2016), an extensive database that allows comparisons of translation activities from English to Danish, German, Spanish, Hindi, Chinese, and, most recently, Japanese. For the first time, this allows comparative studies of cognitive effort, as measured by PWR, for post-edited MT in languages that have very different structure from the source text language. As expected, preliminary computations show moderate correlation between PWR and average fixation count and between PWR and average fixation time.

4.1 PWR comparisons – En ➔De, Es, Hi, Ja

Among other data, the CRITT TPR-DB has data from 21 English-Hindi, 24 English-German, 32 English-Spanish, and 38 English-Japanese translators performing a variety of translation tasks using selections from 6 general English source documents. We computed the average PWRs for each participant at a 300 ms pause threshold for two conditions: post-editing (using Google Translate) and translation from scratch. The results are displayed by participant and grouped by target language in Figures 1 (post-editing) and 2 (translation.)

Since the higher the PWR value, the more cognitive effort is expended, the target languages can be ranked from least effortful to most effortful in the order: Spanish, German, Japanese, Hindi. The same order applies to both conditions, but as in other studies (e.g., Green et al., 2013) translation from scratch is always more effortful than post-editing.

It is not surprising that the two European languages are the least effortful for translators. For an English speaker, German has more structural complexity than Spanish, so the finding that English-German translating/post-editing are more effortful than English-Spanish translating/post-editing is to be expected.

The other two languages, Hindi and Japanese, are written with different scripts than the European languages, so the increased effort for these languages is also to be expected. However, the finding that English-Hindi is more effortful than English-Japanese merits more study: on the surface, the Indo-European Hindi might be expected to have more in common with English than Japanese.

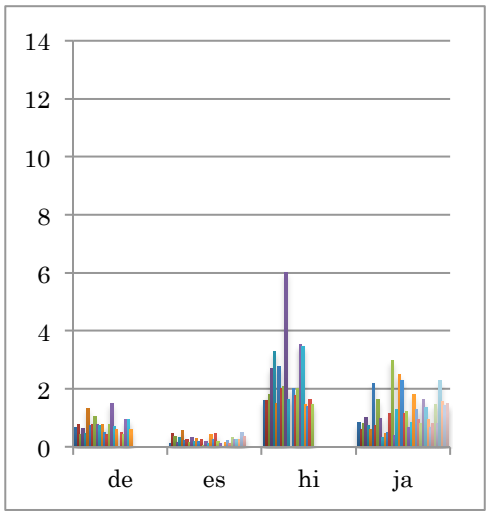

Figure 1. Average post-editing PWR values by participant

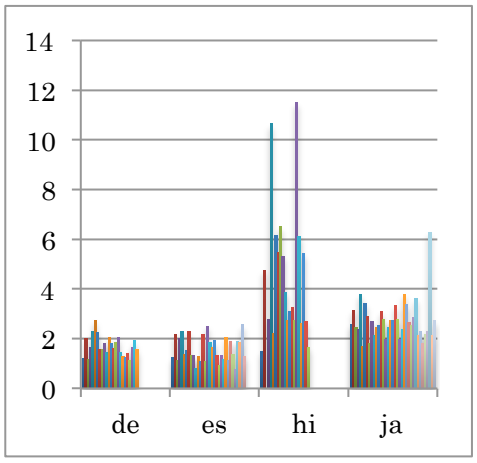

Figure 2. Average translation PWR values by participant

5 Conclusions

We have seen that pause data generated from keystroke loggers provides good information about cognitive effort in various translation tasks. It complements information from eye

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

tracking. Taken together the two modalities for measuring cognitive effort promise to yield deeper insights into the translation process and translation utility, and so productivity.

Bibliography

Daems, J., Vandepitte, S., Hartsuiker, R., & Macken, L. (2015). The impact of machine translation error types on post-editing effort indicators. In 4th AMTA Workshop on Post-Editing Technology and Practice (WPTP4) (pp. 31-45).

Carl, M., Schaeffer, M., and Bangalore, S. 2016. The CRITT Translation Process Research Database. In: New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB. Carl, M., Bangalore, S., and Schaeffer, M. (Eds.) Cham: Springer (pp. 13-54).

Göpferich, S., Bayer-Hohenwarter, G., Prassl, F., & Stadlober, J. 2011. Exploring translation competence acquisition: Criteria of analysis put to the test. O'Brien, S. (Ed.), 57-85.

Green, S., Heer J., Manning, C. D. 2013. The Efficacy of Human Post-Editing for Language Translation. Human Factors in Computing Systems (CHI'13), Paris, France.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. Psychological Review, 87(4), 329.

Krings, H. P. 2001. Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes. Geoffrey S. Koby (Ed.). Kent, Ohio: Kent State University Press.

Lacruz, I., Shreve, G. M., and Angelone, E. 2012. Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing: A Case Study. Proceedings of AMTA Workshop on Post-Editing Technology and Practice (pp. 29-38), San Diego, California.

Lacruz, I., and Shreve, G. M. 2014. Pauses and Cognitive Effort in Post-Editing. In O'Brien, S., Balling, L. W., Carl, M., Simard, M., and Specia, L. (Eds.). Post-editing of Machine Translation: Processes and Applications. Cambridge Scholars Publishing.

Liu, Y. and Du, M. 2014. A Multiple Case Study of Chinese-English Translation Studies in Literature and Language, 9(3)

O'Brien, S. 2006. Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output. Across Languages and Cultures, 7(1), 1-21.

O'Brien, S. 2011. Towards predicting post-editing productivity. Machine Translation 25, 197-215.

Sanchis-Trilles, G., Alabau, V., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., Germann, U., González-Rubio, J., Hill, R. L., Koehn, P., Leiva, L. A., Mesa-Lao, B., Ortiz-Martínez, D., Saint-Amand, H., Tsoukala, C., & Vidal, E. 2014. Interactive translation prediction versus conventional post-editing in practice: a study with the CasMaCat workbench. Machine Translation 28, 217–235.

Schilperoord, J. 1996. It's About Time: Temporal Aspects of Cognitive Processes in Text Production. Amsterdam: Rodopi.

Schwartz, L., Lacruz, I., and Bystrova, T. 2015. Effects of word alignment visualization on post-editing quality and speed. Proceedings of MT Summit XV, Vol. 1: MT Researchers' Track (pp. 186-199).

Snover, M., Dorr, B., Schwartz, R., Miccuilla, L., and Makhoul, J. 2006. A Study of Translation Edit Rate with Targeted Human Annotations. Proceedings of AMTA (pp. 223-231). Cambridge, Massachusetts, USA.

Tyler, S.W., Hertel, P.T., McCallum, M. C., & Ellis, H. C. 1979. Cognitive Effort and Memory. Journal of Experimental Psychology: Human Learning and Memory, (5), 607-617.

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.