Edit distances do not describe editing, but they can be useful for translation process research

Félix do Carmo
ADAPT Centre / Dublin City University
CTTS - Centre for Translation and Textual Studies / Dublin City University
felix.docarmo@adaptcentre.ie

Abstract

Translation process research (TPR) aims at describing what translators do, and one of the technical dimensions of translators' work is editing (applying detailed changes to text). In this presentation, we will analyze how different methods for process data collection describe editing. We will review keyloggers used in typical TPR applications, track changes used by word processors, and edit rates based on estimation of edit distances. The purpose of this presentation is to discuss the limitations of these methods when describing editing behavior, and to incentivize researchers in looking for ways to present process data in simplified formats, closer to those that describe product data.

1 Research background

The technical dimension of translation, revision and post-editing is characterized by writing actions. Editing, part of this technical dimension, is a set of actions that is applied to pre-existing text. This implies that editing cannot be analyzed in the same way as translating or writing from scratch. We see editing as being composed of four actions: delete, insert, move and replace (do Carmo 2017). This presentation discusses the implications of this definition of editing and of different methods to describe it.

If we want to know which words were edited and how, we need data that accurately describes the actions performed. After we have that data, we may extract from it features that can be used to train computational models that predict editing patterns and behaviors.

TPR tools, like Translog II (Carl, 2012) and Inputlog (Leijten and Van Waes, 2013) use keylogging to collect process data, in a character and chronological base. However, it has been shown that it is not straightforward to convert TPR data into word-based sequences of edit actions (do Carmo et al., 2018; Leijten et al., 2012). The main reason for this is the fact that process data is not linear: it includes incomplete, repeated, wrong actions, scattered edits, and other process components that cannot be associated with the words that survive in the final edited versions.

Word processors and translation tools often incorporate track change features that record editing, but these too are not straightforwardly converted into editing data.

Product data seems to describe a simpler reality, so simpler methods may be used. Edit distances appeared in the 1960's as methods to identify and correct errors of spelling in text typing (Damerau, 1964), and errors in computer code (Levenshtein, 1966). These edit distances evolved into metrics like WER–Word Error Rate (Popovic and Ney, 2007) and TER–Translation Edit Rate (Snover et al., 2006), both of which have several variants.

Edit rates identify differences between two versions of a text, and they have been extensively used in applications like automatic post-editing (do Carmo et al, 2019) and quality estimation of machine translation (Specia et al., 2018). In these applications, they are seen as good predictors of the editing required by texts or sentences.

2 Experiment

We conducted a brief experiment to assess the capacity of different methods to identify the editing actions actually performed by translators. We created a test set of a few sentences to which we simulated the application of edits in a sequence of growing complexity. This experiment allowed us to describe the structure of different data collection and analysis methods and to show their limitations in identifying the actions that were performed on one version of a text to transform it into another version. Methods like TER are analysed and described in detail.

3 Results and discussion

One of the conclusions of the experiment above is that edit distances should not be used as descriptors of processes. Nevertheless, edit distances are very useful. Their power lies in their intuitiveness and descriptive capacity: everything is a change in a unit, in a position, or in both. And four actions only (delete, insert, replace and move) describe all transformations that can be done to a sentence. But the main contribution of these metrics is the efficiency requirement – the aim is to identify the 'minimum distance' from one string to the other. This has led to an oversimplified view of editing, but it may have a positive use.

For the TPR community, it would be useful to have a description of editing work that benefited from these simplified descriptions. There would be obvious advantages in converting process data into formats inspired by editing rates. One of the advantages would be that machine translation researchers could more easily integrate the knowledge created by the TPR community. Besides, based on simpler data descriptions, more complex research can be done, enabling us to test further dimensions of editing, like the relation between edit rates and technical effort, or to study different rates of intensity of editing in translation, revision and post-editing.

Acknowledgements

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under the EDGE COFUND Marie Skłodowska-Curie Grant Agreement no. 713567. This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number 13/RC/2077.

References

Carl, Michael. 2012. Translog-II: a Program for Recording User Activity Data for Empirical Translation Process Research. *LREC 2012, 8th International Conference on Language Resources and Evaluation.* Istanbul (Vol. 3, pp. 153–162).

Damerau, Fred J. 1964. A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM* 7 (3): 171–76. https://doi.org/10.1145/363958.363994.

do Carmo, Félix. 2017. "Post-Editing: A Theoretical and Practical Challenge for Translation Studies and Machine Learning." Universidade do Porto. https://repositorio-aberto.up.pt/handle/10216/107518.

do Carmo, Félix, Klaus Buchegger, Rossana Cunha, and Michael Carl. 2018. New Ways of Describing Editing in TPR-DB. *5th International Conference on Cognitive Research on Translation and Interpreting.* Beijing, China.

do Carmo, Félix. et al. 2019. 'A Review of the State-of-the-art in Automatic Post-editing', *Machine Translation*, (forthcoming).

Leijten, Mariëlle, & Luuk van Waes. 2013. Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication* 30(3), 358–392 doi: 10.1177/0741088313491692

Leijten, Mariëlle, et al. 2012. From Character to Word Level: Enabling the Linguistic Analyses of Inputlog Process Data. *EACL-Computational Linguistics and Writing (CL&W 2012): Linguistic and Cognitive Aspects of Document Creation and Document Engineering*, 1–8.

Levenshtein, Vladimir I. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10 (8): 707–710.

Popovič, Maja, Hermann Ney. 2007. Word error rates: decomposition over POS classes and applications for error analysis. *Proceedings of the 2nd workshop on Statistical Machine Translation (WMT 2007),* Prague, pp 48–55

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of AMTA 2006,* August: 223–31. https://doi.org/10.1.1.129.4369.

Specia, Lúcia, Scarton, Carolina and Paetzold, Gustavo. 2018. *Quality estimation for machine translation*. Morgan & Claypool. doi: 10.2200/S00854ED1V01Y201805HLT039.