Introduction to special issue on post-editing

Sharon O'Brien · Michel Simard

Published online: 21 November 2014
© Springer Science+Business Media Dordrecht 2014

With the increasing deployment of machine translation (MT) in certain sectors of the translation industry, a spotlight has turned to the task of post-editing, which is still essential when high quality translation is required. In a recent survey of 1,000 Language Service Providers (LSPs) (DePalma et al. 2013), 44 % reported that they were offering MT and Post-Editing as a service. At the same time, LSPs appear to struggle with the introduction of post-editing as a service, anecdotally reporting that there is significant translator resistance to the task. There are many reasons for this resistance and an in-depth discussion is beyond the scope of this introduction (for more detailed discussion see, e.g. O'Brien and Moorkens 2014). The increasing demand for post-editing has led to a propagation of research in the past decade. We have seen the production of several theses (e.g. Tatsumi 2011; Guerberof 2012; De Almeida 2013), journal articles (e.g. García 2010, 2011), edited volumes (O'Brien et al. 2014a), workshop proceedings (O'Brien et al. 2012, 2013, 2014b) as well as many individual conference papers. In addition, EU funded projects such as CasMaCat¹ and MateCat² have contributed to the topic and to technological development (see also Moran et al. 2013). The topics that have received most attention in research to date include productivity, impact on quality, cognitive effort and, to a lesser extent, automatic post-editing and correlations between effort and automatic quality scores. The research has been conducted by scholars

---

¹ <http://www.casmacat.eu/>.

² <http://www.matecat.com/matecat/the-project/>.

S. O'Brien (✉)

School of Applied Language and Translation Studies, Centre for Translation and Textual Studies,
CNGL, Dublin City University, Dublin, Ireland
e-mail: sharon.obrien@dcu.ie
<http://cttsdcu.wordpress.com>

M. Simard

National Research Council Canada, 1200 Montreal Road, Building M-50, Room 380B, Ottawa,
ON K1A 0R6, Canada

working in the domain of MT and in Translation Studies, sometimes in collaboration with each other and in collaboration with translation service providers and translators, both professional and novice.

The papers that have been selected for this special issue reflect both the topics that are currently of most concern and the fact that research is being conducted within and across two related disciplines: natural language processing and translation studies. We believe that the papers collected here are a good representation of the various research questions and methodologies being used. Broadly, the papers can be divided into the following topics: Productivity and Quality (Guerberof, Mitchell et al., Huang et al., Turchi et al.); Cognitive Effort (Vieira); and New Technologies (Sanchis-Trilles et al., Bertoldi et al.).

Ana Guerberof's article reports on a productivity and quality-focused research project that compares post-editing productivity across different types of segments, namely translation memory (TM) fuzzy match segments (in the 85–94 % fuzzy match range), machine-translated segments generated using a domain-trained Moses engine (Koehn et al. 2007) and 'No Matches', i.e. segments for which no proposal was offered to the participants. One of the strong points of this article is that the research was conducted in an industrial localisation scenario, employing professional translators as post-editors with content that is typical of the localisation industry. Her main research question was also of high relevance to the translation industry: what level of pay should companies offer for post-editing? Guerberof reports that productivity was slightly higher for the MT matches, though this is reported to be not statistically significant when compared with the fuzzy matches. This finding suggests that a pay rate similar to that of fuzzy matches in the 85–94 % range would be fair. At the same time, Guerberof rightly highlights the fact that her 'raw' MT output was of high quality to start with, as suggested by the overall BLEU score (Papineni et al. 2002) of 0.60. The engine was trained using clean, domain-specific data, and so her productivity results should be viewed in light of this fact. She draws interesting links with similar studies which all seem to point towards the same trend, despite the fact that those studies were conducted using different MT engines, content, post-editors and language pairs. At the same time, it would be unwise to conclude that the findings on productivity would apply to any MT engine; raw MT quality appears to be key here. Guerberof's article also points towards other key issues: the three professional reviewers did not share a high level of agreement when reviewing TM and MT matches, but seemed to agree more when reviewing 'No Matches', and some translators appear to benefit more from TM matches while others benefit more from MT. This latter point has implications for pricing of course and Guerberof makes the interesting point that potential post-editors would need to decide for themselves whether a pricing model would be to their benefit or disadvantage.

Lucas Vieira's paper also focuses on the topic of post-editing effort, but he emphasizes that measuring 'effort' only as a factor of time or edit distance could be misleading. The focus of this paper is on predicting cognitive effort during the post-editing task by factoring in a number of aspects such as source-text complexity, MT output characteristics and individual characteristics such as working memory capacity and source-language proficiency. As highlighted in Guerberof's article, and reflected in many publications on post-editing to date, individual variation is inevitable.

Vieira tries to combat this issue by using mixed-effect models. His research is also unusual because it employs French as the source language, whereas most post-editing and MT research to date uses English as the source language. An additional interesting aspect to this paper is the search for correlations between post-editing cognitive effort and the Meteor automatic evaluation metric, with observations that Meteor (Banerjee and Lavie 2005) is in fact a good predictor of cognitive effort, especially for longer sentences. Vieira's experiment leads to some tentative conclusions about the relationships between source-text characteristics, such as type-token ratio and the number of nouns, and the cognitive effort involved in post-editing, though he emphasizes that effects of source-text characteristics were small and thus need further investigation. His conclusions on source-language competence and perceived cognitive effort as well as working memory capacity and effort lead to the conclusion that further research is required on these aspects. Nonetheless, this is a first step at trying to understand the interplay between complex factors during post-editing and is a welcome move away from more simplistic observations on post-editing effects.

The paper authored by Sanchis-Trilles et al. considers productivity, but from a different perspective again. The CasMaCat Workbench is deployed in a field trial involving professional translators in order to explore whether different modes have an impact on post-editing data. The conventional 'mode' is, of course, where a static machine-translated segment is presented beside the source-language segment. The relatively new 'mode' of interest to Sanchis-Trilles and his co-authors is that of *interactive translation prediction* (ITP), in which the MT segment changes depending on the choices made by the post-editor in real time. The authors do not only compare the standard mode with an interactive mode, but also introduce a variant on the ITP mode, called *advanced interactive mode* (AITP), in which all interactive features are available to the post-editors, but they can select which ones they would like to turn on and off. The main focus of this paper is on whether one or both of the interactive modes can improve post-editing productivity and also on what the attitudes are of the professional translators who use each mode. The authors use some traditional approaches to measure which mode facilitates higher productivity, e.g. pause-time analysis and edit distance. In addition, the novel integration of CasMaCat with an eye tracker enables an analysis of gaze data. Furthermore, they include an essential analysis of the impact of each mode on the final quality of the machine-translated and post-edited text. Although reviewer behaviour is heterogeneous, as also highlighted in Guerberof's article, the general conclusion is that the variance in mode does not affect the final product quality. Sanchis-Trilles et al. make useful observations about the positive impact of the ITP mode, which appears to require fewer insertions and deletions compared to other modes, but slightly increases post-editing time. They also observe that the interactive modes result in longer gaze times on the target-language window, when compared with the conventional post-editing mode. This is not surprising; as text changes on screen we are probably more likely to focus on that evolving text. A question that emerges here is what impact the increasing focus on the target text and the decreasing focus on the source text might have on accuracy. That is a question that may be worthy of further analysis into the future.

The relatively novel topic of user-generated content (UGC) in online technical forums, and the feasibility of having such content post-edited to a satisfactory level

by the forum members themselves is tackled in Mitchell et al's paper. An experiment in which members of Symantec's German-speaking Norton Antivirus online community are asked to participate in post-editing forms the basis for analysis in this paper. Companies like Symantec rely somewhat on their online communities to provide knowledge and support, but they face the problem that much of the content is user-generated and is in English. How then can the non-English speaking community members benefit from this English content? The obvious solution is MT. However, given the expected quality of MT for UGC, and the potentially large audience to which this content may be exposed, it is often advisable to submit it to post-editing. As the online community is willing to provide content, are they also willing to post-edit? If so, can they post-edit to a quality level that is accepted by the community itself? Will the community also engage in quality measurement and, if they do so, how does their evaluation compare with more traditional forms of quality evaluation, such as rating for adequacy and fluency, or annotating errors? These questions around "community post-editing" are tackled in the paper by Mitchell and her co-authors. What differentiates this research from previous publications on community or volunteer translation is that the community members did not join the community to participate in a volunteer translation effort, but are there presumably to learn about the company's products and services and to supply support to fellow users.

For a number of years now, post-editing has been part of IBM's internal translation workflow. In their contribution to this issue, Huang et al. describe the company's effort to integrate reliable MT confidence estimation (CE) to this process. For translators working with MT, an important part of the task is determining rapidly whether an individual MT proposal will be useful, or if it would be more productive to ignore it and translate from scratch. CE is a mechanism by which the MT system itself communicates how confident it is about the quality of its own output. In principle, a well-designed CE mechanism can save the translator time and effort, by drawing his or her attention to those MT proposals that are more likely to be useful. In practice, however, CE is a difficult task; in fact, it can be argued to be as difficult as translation itself. In the case of IBM's internal procedure, this is complicated by their reliance on a document-adapted MT strategy: for each document, a new, customised MT system is generated, using only those segments from their translation archive that are most relevant to the document. This raises coherence issues for standard, static CE solutions, because the same CE component may produce very different estimates for the same translation, when produced by two different systems. The authors propose an elegant solution to this problem, building document-specific CE models in tandem with the MT systems. In experiments on English to Italian and English to Chinese translation, they demonstrate that their approach leads to coherent CE models, whose outputs are both reliable and comparable across systems and documents. Further experiments with English to Japanese suggest that providing CE information allows post-editors to make better decisions, increasing their productivity by approximately 10 %.

Turchi et al. examine a related problem with CE. In general, CE is cast as a standard machine-learning regression task, in which the system learns to predict confidence estimates from labelled examples. Typically, these labels are numerical values: they can be estimates of common MT evaluation metrics, such as BLEU, TER (Snover et al. 2006) or Meteor, of Likert scale quality assessments or even of post-editing

time. Such estimates are often difficult to interpret and use in practice, especially if the goal is to make a quick decision on whether or not to use a given MT proposal to translate a given segment. Typically, this problem is addressed by implicitly or explicitly setting arbitrary thresholds on the CE component's output, that define the border between useful and useless MT proposals. But this task is complicated by the fact that often, the confidence estimate in itself is not sufficient to make this decision. For example, while a five-word segment with a high estimated TER may be useless as a starting point for a translation, a 20-word segment with the same TER may turn out to contain a long segment that can be productively reused. Turchi and his co-authors suggest that a binary labelling (useful/useless) would be more directly usable; unfortunately, learning data with such annotations does not normally exist. Therefore, they propose an automatic relabelling approach for the learning data, based on a three-way comparison between the MT output, a post-edited version of this translation and an independently produced reference translation. Intuitively, useful MT should be very similar to its post-edited version, while the post-edited version of useless MT should be as different from the MT as any independently produced translation. The authors find that their method produces more balanced datasets, which in turn result in more reliable classifiers. More importantly, their experiments with post-editors suggest that this approach can capture individual post-editor preferences, while at the same time encouraging more productive usage of MT proposals.

If there is something that post-editors dislike, it's fixing the same errors over and over again. Bertoldi et al. tackle precisely that problem in their contribution to this issue, through incremental training of MT systems. The idea is straightforward: the MT system that generates the translation proposals for a post-editor should have the ability to learn from its errors; theoretically, if post-edited translations are fed back to the MT system as soon as they are available, then it can incorporate this new knowledge immediately and modify its behaviour accordingly. In an ideal world, an MT system equipped with such online learning capabilities would never make the same mistake twice. In practice, statistical MT systems are unwieldy beasts, and implementing this functionality raises numerous technical challenges. Bertoldi et al. take these challenges one by one, and propose two distinct solutions, that can actually be deployed concurrently. The first solution relies on cache-based mechanisms to store recently observed phrase pairs and sequences of target-language words, which are then provided to the decoder as additional translation hypotheses. The second solution is a reranking scheme, using structured perceptrons, in which the relative order of the decoder's top-scoring translation hypotheses is re-evaluated in light of their similarity to recently observed post-edited translations. In simulated post-editing experiments with various text domains and language pairs, both methods allow substantial gains in BLEU and TER. The best results are obtained when both methods are applied together, on text documents that show high rates of internal repetition. While the actual impact of such mechanisms has yet to be tested with real post-editors, this work constitutes an encouraging first step towards a realistic approach to real-time adaptation of MT systems.

We sincerely thank the reviewers who participated in the review process for this special issue: Fabio Alves, Lynne Bowker, Michael Carl, Maureen Ehrensberger-Dow, Andreas Eisele, Marian Flanagan, Debbie Folaran, George Foster, Miguel Jiménez-

Crespo, Natalie Kübler, Isabel Lacruz, Alon Lavie, Arle Lommel, Elliott Macklovitch, Daniel Marcu, Gary Massey, Alan Melby, Ricardo Muñoz Martín, Kristen Parton, Mirko Plitt, Maja Popović, Anne Scholdager, Serge Sharoff, Gregory Shreve, Lucia Specia, Midori Tatsumi, Guillaume Wisniewski and Ventsislav Zhechev. We also thank Prof. Andy Way for his guidance during the process.

References

- Banerjee S, Lavie A (2005) METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. USA, June, Ann Arbor, pp 65–72
- De Almeida G (2013) Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two romance languages. PhD thesis, Dublin City University, Dublin
- DePalma D, Hegde V, Pielmeier H, Stewart R (2013) The language services market: 2013, Common Sense Advisory Report
- García I (2010) Is machine translation ready yet? *Target* 22(1):7–21
- García I (2011) Translating by post-editing: Is it the way forward? *Machine Translation* 25(3): pp 217–237. Springer, New York
- Guerberof A (2012) Productivity and quality in the post-editing of outputs from translation memories and machine translation. PhD thesis, Universitat Rovira i Virgili, Tarragona
- Koehn P, Huang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin E, Herbst E (2007) Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the association for computational linguistics companion volume: Proceedings of the demo and poster sessions, Prague, June pp 177–180
- Moran J, Lewis D, Saam C (2013) Analysis of post-editing data: A productivity field test using an instrumented CAT tool. In: O'Brien S, Winther Balling L, Carl M, Simard M, Specia L (eds) Post-editing of machine translation: processes and applications. Cambridge Scholars Publishing, Newcastle upon Tyne, pp 126–146
- O'Brien S, and Moorkens J (2014) Towards intelligent post-editing interfaces. In: Baur, W., Eichner, B, Kalina, S, Kessler, N, Mayer F, and Orsted J (eds). Proceedings of the 20th FIT World Congress, August, Berlin, Germany, pp 131–137
- O'Brien S, Simard M, Specia L (2012) Proceedings of the first Workshop on Post-Editing Technologies and Practice (WPTP1), Workshop held at the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2012), October, San Diego
- O'Brien S, Simard M, Specia L (2013) Proceedings of the second Workshop on Post-Editing Technologies and Practice (WPTP2), Workshop held at the Machine Translation Summit XIV, Nice, September
- O'Brien S, Winther Balling L, Carl M, Simard M, Specia L (eds) (2014a) Post-editing of machine translation: Processes and applications. Cambridge Scholars Publishing, Newcastle upon Tyne
- O'Brien S, Simard M, Specia L (2014b) Proceedings of the third Workshop on Post-Editing Technologies and Practice (WPTP3), Workshop held at the Eleventh Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014), October, Vancouver
- Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. Proceedings of 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2006). Cambridge, August pp 223–231
- Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002). Philadelphia, July, pp 311–318
- Tatsumi M (2011) Post-editing machine translated text in a commercial setting: Observation and statistical analysis. PhD thesis, Dublin City University, Dublin