Editors' foreword to the special issue on human factors in neural machine translation

Sheila Castilho¹ · Federico Gaspari¹ · Joss Moorkens² · Maja Popović¹ · Antonio Toral³

Published online: 2 May 2019
© Springer Nature B.V. 2019

Over the past 5 years the machine translation (MT) community has become aware of the potential of neural machine translation (NMT) to sustain the increases in output quality that had appeared to plateau when using statistical MT (SMT) (Kenny 2018). This has led an increasing number of MT providers and research groups to focus their energies and resources on developing NMT systems.

Early studies on NMT quality demonstrated that, in general, this MT paradigm yields higher automatic evaluation metric scores than its predecessor, SMT (Bahdanau et al. 2014; Jean et al. 2015; Bojar et al. 2016; Koehn and Knowles 2017). NMT has also been shown to provide a jump in fluency when compared with SMT (Bentivogli et al. 2016; Toral and Sánchez-Cartagena 2017). This increased fluency has quickly made NMT the preferred MT paradigm for assimilation, as is evident from the move to NMT by many major online MT providers. Where MT for *dissemination* is concerned, when text is “machine translated as an intermediate step in production” (Forcada 2010), we might reasonably assume that the reported increase in quality would result in a concomitant productivity boost. However, studies such as Castilho et al. (2018) reported that NMT delivers only minor improvements in productivity and technical effort, relative to the improved scores using automatic metrics and human fluency evaluation, when comparing with phrase-based SMT (PBSMT) systems.

The rule of thumb for MT deployment suggested by Way (2018) is that “the degree of human involvement required—or warranted—in a particular translation scenario will depend on the purpose, value and shelf-life of the content.” However, positive evaluations of NMT for assimilation alongside occasionally hyperbolic reports in the media (as reported in Castilho et al. 2017; Toral et al. 2018) have

✉ Joss Moorkens
joss.moorkens@dcu.ie

¹ ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland

² ADAPT Centre, School of Applied Language and Intercultural Studies, Dublin City University, Dublin, Ireland

³ University of Groningen, Groningen, The Netherlands

pushed raw and post-edited MT into action in use-cases for which MT would previously have been considered inappropriate (Schmidtke 2016; Guerberoef 2018). The rise of NMT as the state-of-the-art has been accompanied by growing awareness in the community of the need to improve methodologies and procedures for translation quality assessment on an ongoing basis, with a view to overcoming the limitations of both automatic metrics and human approaches, limiting the overhyping of NMT, and explaining the somewhat paradoxical results for NMT for dissemination (Läubli et al. 2018; Moorkens et al. 2018).

The current special issue attempts to address the latter point. Due to the novelty of NMT, little is known as yet about how humans—especially translation professionals, translation students, and end-users—engage with NMT output. Will the same types of errors that occur in SMT and rule-based MT (RBMT) systems recur in NMT outputs? Will translators take longer or become faster when post-editing (or otherwise processing) NMT output to improve productivity? Is cognitive effort higher or lower when processing NMT output? How is the end-user experience with NMT systems? How does post-editing (PE) NMT output compare with using translation memories (TMs) and adapting their fuzzy matches? This special issue aims to address these and similar questions around human factors in NMT by bringing together a collection of novel articles offering state-of-the-art research on a wide range of topics related to translation quality in terms of PE, error analysis, as well as the application of controlled languages in pre-processing. The articles adopt multiple complementary perspectives to tackle the issues at hand and cover a variety of language pairs and domains, showing the wide applicability of NMT to real-life tasks.

In this special issue, while most of the papers focus specifically on several aspects of PE, contributions that more closely consider the role of interactive MT, error analysis and controlled language in the human factors of NMT are also included.

1 Post-editing

PE effort (temporal, technical and cognitive, as per Krings 2001) with NMT output is usually reported in comparison with different translation approaches, i.e. human translation (HT) with or without TM matches, or with PE of other MT systems. While NMT PE shows large differences on the cognitive, temporal and technical levels when compared to HT, when it is compared to SMT output and TM matches, research does not yet seem to indicate that it is a significantly faster task in all scenarios. In this special issue, a good number of articles aim to investigate the differences between other translation approaches and translating with the aid of NMT.

Jia et al. compare fluency, accuracy and PE effort of Google's PBSMT and NMT engines for English-to-Chinese translation of two news texts. Their findings suggest that post-editing NMT reduces temporal, technical, and cognitive effort for this language pair and text type. Interestingly, they also find a strong correlation between pause-based metrics that have been independently proposed very recently for cognitive effort, and that translation from scratch is more prone to speed variability based on source text complexity.

Sánchez-Gijón et al. investigate the differences between PE of a generic NMT system and translation using TM matches in English-to-Spanish technical translation, in terms of edit time and edit distance, as well as translators' perceptions of NMT for productivity, considering in particular how these dimensions vary in relation to segment length. Their findings show that while NMT PE necessitates less editing than TM segments, it takes longer on average. The authors note that translators who perceived MT as boosting their productivity actually performed better when post-editing MT segments than those translators who perceived MT to be a poor resource.

Koponen et al. combine a product-based and a process-based approach to verify whether different editing patterns exist when post-editing NMT, SMT and RBMT outputs. They find that whereas NMT has the greatest numbers of word-form changes and word-substitution edit types, RBMT shows more deletion edits, and SMT more insertions. The effort indicators show a slight increase in keystrokes per word for NMT output, and a slight decrease in average pause length for NMT compared to the other systems. The authors argue that studies in PE quality and effort should identify preferential edits, participant errors, and individual differences in process metrics.

Herbig et al. explore how multiple modalities to measure cognitive load, including eye-, skin- and heart-based indicators, might be combined to predict the level of perceived cognitive load during NMT PE. Their results show that PE time strongly correlates with perceived cognitive load and, moreover, that a combined multimodal approach is able to estimate cognitive load during PE without the actual process being interrupted through manual ratings.

2 Interactive MT

Interactive and adaptive MT is one possible alternative method of employing MT for dissemination outside of PE, which Green (2016) called a “broken usability model” wherein MT suggestions “prime translators” (Green et al. 2013). Daems and Macken compare the differences between interactive adaptive SMT and NMT regarding quality, translation process, perceived usability, and translators' attitude towards an interactive translation tool. The authors find that even though SMT suggestions contain more errors than NMT suggestions, neither translation time nor effort are significantly affected by the difference in quality. The authors argue that the differences found may be due to individual differences between translators, and that, while fewer errors were found in NMT output, these “could be harder to detect and to solve”. Despite this, users prefer to work with NMT output. Improved usability, even without increased productivity, may still be considered to make a move from interactive SMT to interactive NMT worthwhile.

Knowles et al. also explore interactive NMT. However, there are two important differences between the two articles: first of all, while this paper compares interactive NMT to PE NMT, Daems and Macken compare this paradigm against interactive SMT; in addition, the computer-assisted translation (CAT) tool employed in this paper is a research product (CASMAT), while that used by Daems and Macken

is a commercial offering (Lilt). Specifically, Knowles et al. investigate whether human translators' productivity increases in a setting that makes use of interactive translation prediction (ITP) with an NMT system. They find that over half of the eight participant translators are faster when using neural ITP, which is preferred over PE by most of the translators. The authors argue then that ITP would be a viable alternative to PE.

3 Error analysis

Error analysis of NMT systems has also been on the radar of the MT field. Several papers have carried out automatic (Bentivogli et al. 2016; Toral and Sánchez-Cartagena 2017) or human error annotation (Burchardt et al. 2017; Klubička et al. 2017; Popović 2017; Castilho et al. 2018) in order to compare phrase-based and neural approaches for different language pairs and domains. In this issue, Calixto and Liu present an extensive error analysis of several MT systems, including two text-only systems that fall into the PBSMT and NMT paradigms, and a set of multi-modal NMT models which use not only text but also visual information extracted from images. The error taxonomy is based on that of Vilar et al. (2006), with a few adjustments. Their goal is to verify whether the multi-modal engine makes fewer errors when translating Flickr image descriptions in comparison to the other systems. Their findings suggest that adding global and local visual features into NMT significantly improves the output, and, moreover, that the *mistranslation* and *wrong sense* error types—which are arguably the most damaging the most damaging for the translation of image descriptions—were drastically reduced in the multi-modal systems. Finally, they find that not only the translation of terms with a strong visual connotation was improved, but also the translation of error types without a visual interpretation.

4 Controlled language

Controlled languages (CLs) for MT have been widely investigated for SMT and RBMT systems (O'Brien 2006; Aikawa et al. 2007; Temnikova and Orasan 2009; Temnikova 2012). However, the effect of CL for NMT has, to the best of our knowledge, not yet been investigated. In this issue, Marzouk and Hansen-Schirra examine the impact of CL rules on the output quality of NMT for the German-to-English language pair when compared to that of four other MT systems that fall under RBMT, SMT, and hybrid paradigms. Their findings suggest that CL does not have a positive impact on Google's NMT system. GNMT's output was the one with the lowest amount of errors both before and after CL application, with a marginal increase in the number of errors after applying some CL rules. In addition, GNMT had the highest quality levels both with and without applying CL rules, with a quality decrease after its application.

In sum, the findings of the articles collected in this special issue demonstrate that there is still a large amount of research to be done on human factors for NMT

systems. As in many research areas and applications that involve professional translators, the experiments with PE, especially with the recent NMT paradigm, have limitations such as small sample sizes, time constraints, and ecological validity (e.g. tools used in the research may not be the same as those used by translators in production). Further efforts are therefore required to be able to generalize the results that this special issue brings to the community, so that the evidence provided by research filters through to practising translators and to translator training programmes that need to keep abreast of technological progress. This does not mean, however, that the current results are not to be trusted, but rather reinforces the need for further investigation with bigger sample sizes, more professional translators, larger groups of translation students, end-users, considering different levels of experience (e.g. in PE), further language pairs and application domains, etc.

The articles herein are presented in the context that it is still early days in the development of NMT. From the outset, the development of MT was proposed as an interdisciplinary pursuit. Weaver's choice of Norbert Wiener, a proponent of interdisciplinary research, as interlocutor in 1947 suggests that he foresaw MT development as requiring a broad combination of skills. Linguists were deeply involved in RBMT development and, far later, in the ecosystem of pre- and post-processing tools that eventually grew around SMT.¹ The early development of NMT has not involved a great deal of linguistic input, perhaps due to the complex nature of systems and the high barriers to entry (in cost and expertise). In that short time, there have been changes to architecture (Vaswani et al. 2017) and training data (Sennrich et al. 2016) that have been motivated by an engineering rather than linguistic focus. Trying to integrate input (that may be vaguely-defined) from non-engineers will be difficult, but our hope is that the articles in this special issue will provide feedback for interesting avenues of future development while also showcasing contemporary research in the area of NMT and human factors.

As co-editors, we hope that this publication will contribute to instigate and inspire further work to expand our knowledge and understanding of the phenomena involved in NMT for dissemination. At the same time, given the obvious applicability of these studies to real-world scenarios, this special issue also has the ambition to be relevant to interested professional translators, post-editors, project managers in language service providers, translation students, trainers and scholars, with a view to promoting the wider uptake of translation technologies informed by research-based good practice. This inclusive approach reflects the combined interests of the co-editors of the special issue, who are all, to different extents, not only involved in MT, PE and human factors research, but also actively engaged in translator training, e.g. as part of academic programmes, industry-facing initiatives, and lifelong professional development activities. In a similar vein, we see this special issue as a timely and forward-looking attempt to bring academic research, teaching and professional practice closer together, to the mutual benefit of these neighbouring communities.

¹ SMT pioneer Peter Brown states that his "goal was to establish the mathematical framework for MT so that the linguistically-minded could proceed with the research" (Way 2009).

Acknowledgements We would like to thank all of the authors who submitted their work in response to our call for papers. 14 articles were received by the deadline in July of 2018, of which 8 were accepted for publication in this special issue. We would particularly like to thank all 26 colleagues who volunteered their time and effort to review articles, journal editor Andy Way for lots of help, and the production staff at Springer for their responsiveness and assistance. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

- Aikawa T, Schwartz L, King R, Corston-Oliver M, Lozano C (2007) Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In: Proceedings of the MT Summit XI. Copenhagen, Denmark, 10–14 September 2007, pp 1–7
- Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. [arXiv:1409.0473](https://arxiv.org/abs/1409.0473)
- Bentivogli L, Bisazza A, Cettolo M, Federico M (2016) Neural versus phrase-based machine translation quality: a case study. In: Proceedings of the 2016 conference on empirical methods in natural language processing, Austin, Texas, pp 257–267
- Bojar O, Chatterjee R, Federmann C, Graham Y, Haddow B, Huck M, Jimeno Yepes A, Koehn P, Logacheva V, Monz C, Negri M, Neveol A, Neves M, Popel M, Post M, Rubino R, Scarton C, Specia L, Turchi M, Verspoor K, Zampieri M (2016) Findings of the 2016 conference on machine translation. In: Proceedings of the 1st conference on machine translation, Berlin, Germany, pp 131–198
- Burchardt A, Macketanz V, Dehdari J, Heigold G, Peter JT, Williams P (2017) A linguistic evaluation of rule-based, phrase-based, and neural MT engines. *Prague Bull Math Linguist* 108(1):159–170
- Castilho S, Moorkens J, Gaspari F, Calixto I, Tinsley J, Way A (2017) Is neural machine translation the new state of the art? *Prague Bull Math Linguist* 108(1):109–120
- Castilho S, Moorkens J, Gaspari F, Sennrich R, Way A, Georgakopoulou P (2018) Evaluating MT for massive open online courses. *Mach Transl* 32(3):255–278. <https://doi.org/10.1007/s10590-018-9221-y>
- Forcada ML (2010) Machine translation today. In: Gambier Y, van Doorslaer L (eds) *Handbook of Translation Studies*, vol 1. John Benjamins, Amsterdam, pp 215–223
- Green S (2016) Interactive machine translation: from research to practice. In: Paper presented at the Twelfth Conference of Association for Machine Translation in the Americas (AMTA), Austin TX, October 28–November 1
- Green S, Heer J, Manning CD (2013) The efficacy of human post-editing for language translation. In: Proceedings of the SIGCHI conference on human factors in computing systems, 27 Apr–2 May 2013, Paris, pp 439–448
- Guerberof A (2018) Usability and Data: Correlations between quality and usability on HT and MT interfaces: a study using eye-tracking and telemetry. Paper presented at the 12th annual Irish Human Computer Interaction conference (iHCI 2018), Limerick, Ireland, November 2
- Jean S, Firat O, Cho K, Memisevic R, Bengio Y (2015) Montreal neural machine translation systems for WMT'15. In: Proceedings of the 10th workshop on statistical machine translation, Lisbon, Portugal, pp 134–140
- Kenny D (2018) Sustaining disruption? The transition from statistical to neural machine translation. *Revista Tradumàtica* 16:59–70
- Klubička F, Toral A, Sánchez-Cartagena VM (2017) Fine-grained human evaluation of neural versus phrase-based machine translation. *Prague Bull Math Linguist* 108(1):121–132
- Koehn P, Knowles R (2017) Six Challenges for neural machine translation. In: Proceedings of the 1st workshop on neural machine translation, Vancouver, BC, Canada, pp 28–39
- Krings HP (2001) *Repairing texts*. Kent State University Press, Kent
- Läubli S, Sennrich R, Volk M (2018) Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp 4791–4796, October 31–November 4
- Moorkens J, Castilho S, Gaspari F, Doherty S (eds) (2018) *Translation quality assessment: from principles to practice*. Springer, Berlin

- O'Brien S (2006) Controlled Language and Post Editing. In *Multilingual*, Issue 83, pp 17–19
- Popović M (2017) Comparing language related issues for NMT and PBMT between German and English. *Prague Bull Math Linguist* 108(1):209–220
- Schmidtke, D (2016) MT Thresholding: Achieving a defined quality bar with a mix of human and machine translation. In: Paper presented at the AMTA 2016 Workshop on Interacting with Machine Translation (iMT 2016), Austin TX, October 28
- Sennrich R, Haddow B, Birch A (2016) Neural machine translation of rare words with subword units. In: *Proceedings of the 54th annual meeting of the association for computational linguistics (ACL 2016)*, Berlin, Germany, pp 1715–1725
- Temnikova I (2012) Text Complexity and Text Simplification in the Crisis Management Domain. Ph.D. Thesis, University of Wolverhampton
- Temnikova I, Orasan C (2009) Post-editing Experiments with MT for a Controlled Language. In: *Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL)*, Besancon, France, July 1–3
- Toral A, Sánchez-Cartagena VM (2017) A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In: *Proceedings of the 15th conference of the European chapter of the association for computational linguistics*, Valencia, Spain, pp 1063–1073
- Toral A, Castilho S, Hu K, Way A (2018) Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In: *Proceedings of the Third Conference on Machine Translation (WMT)*, Volume 1: Research Papers, Belgium, Brussels, pp 113–123, October 31–November 1
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In *proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA
- Vilar D, Xu J, D'Haro L, Ney H (2006) Error analysis of statistical machine translation output. In: *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC)*, Pisa, pp 697–702
- Way A (2009) A Critique of Statistical Machine Translation. In: Daelemans W, Hoste V (eds) *Journal of translation and interpreting studies: special issue on evaluation of translation technology*, *Linguistica Antverpiensia*. Academic and Scientific Publishers, Antwerp, pp 17–24
- Way A (2018) Traditional and emerging use-cases for machine translation. In: Moorkens J, Castilho S, Gaspari F, Doherty S (eds) *Translation Quality Assessment: From principles to practice*. Springer, Berlin, pp 159–178

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.