For a common European framework for evaluating AI-based translation technologies

Philippe Langlais, François Yvon

For a common European framework for evaluating AI-based translation technologies

Philippe Langlais
RALI, DIRO, Università di Montréal

François Yvon
Sorbonne Université e CNRS

Deep neural networks, and especially the Transformer architecture (Vaswani *et al.*, 2017), have brought tremendous progress in machine translation (Sutskever *et al.*, 2014; Bahdanau *et al.*, 2016). Many services based on this technology can produce good quality translations, though they are still often literal (Bhardwaj *et al.* 2020), contain contradictions or omissions, and are less pertinent in certain specific areas such as the financial and automotive industries.

To improve machine translation for professional purposes, it is essential to develop a better European framework based on practice-oriented metrics and pertinent, absolutely multisectorial data. This will entail involving the many actors in the world of translation: scholars in NLP and translation studies, companies providing translation devices and software, as well as translators and translation services. Doing so at the European level is essential in order to pursue a successful strategy for reducing the technological inequalities between European languages[1] and, above all, enabling Europe to take the lead in integrating technologies for professional use.

A few words are thus in order concerning current efforts to evaluate machine translation, such as the WMT campaigns[2], which concentrate on evaluating translation technologies and output quality, or their IWSLT[3] counterparts for spoken language translation (interpreting). In the WMT work, data are annotated according to the MQM taxonomy by translation professionals whose tasks include evaluating post-editing effort and predicting whether a translation contains so-called 'catastrophic' errors.

Despite this work, it is still difficult to gauge the results of quality evaluation efforts. Though some system have gone beyond simply detecting errors, few or none analyse them. It thus comes as no surprise that a review of the principal research papers dealing with machine translation published between 2010 and 2020 (Marie *et al.* 2021) found that BLEU scores (Papineni *et al.* 2002) continue to be used to measure how close a machine translation is to a reference human translation by counting the words and phrases they share.

Document-level translation quality evaluation is still uncommon (Specia *et al.* 2020; Zerva *et al.* 2022), though it is extremely useful from a professional standpoint.

[1] See the European Language Equality project at https://european-language-equality.eu/wp-content/uploads/2022/11/ELE___Deliverable_D3_4__SRIIA_and_Roadmap___final_version_-1.pdf

[2] http://www2.statmt.org/wmt23/

[3] https://iwslt.org

Lastly, there are too few studies addressing the management of Translation Memories (TMs) and their use in the translation process, though they are essential professional tools.

As for how deep learning systems are trained and tested, data used for this purpose have been collected from the published proceedings of the European Parliament (Koehn 2005), United Nations documents (Ziemski *et al.* 2016) and from parallel corpora harvested from the Internet (Esplà *et al.* 2019). In addition, specific data have been evaluated in the past for particular sectors, for the media, and so forth.

Despite the abundance of this data (at least for some language pairs), however, no attention has been devoted to splitting training and test corpora in any functionally targeted way. Test corpora are often packed with stereotype-laden and extremely repetitive phrases which, moreover, are already present in the training corpora. This risks 'contaminating' the tests, with repercussions that include overoptimistic evaluations. Few studies have addressed the evaluation of the examples used to train models, where quantity trumps quality. And yet, selecting data according to specific criteria would make it possible to train more robust models and build more consistent datasets (in this connection, see the exemplary case described by Varshney *et al.* 2022).

It should also be borne in mind that the document is a secondary element in organising data and that corpora are usually segmented in equivalent sentences for language pairs (the so-called aligned corpora). This makes it difficult to produce a cohesive translated text, given that the basic unit is the sentence.

More generally, developing a single system capable of dealing with multiple domains, though increasingly fundamental, is still an underinvestigated—and hence unsolved—problem (Pham *et al.* 2021). Most of the studies in this area have addressed a small number of highly diverse sectors (biomedicine, finance, technology) and thus do not encompass the broad array of domains that translation services must consider. For example, it has been found (Frenette 2021) that a generic neural translation system had difficulty in translating texts in several of the sectors handled by the Canadian Government's Translation Bureau, and that technical attempts to provide the system with further information in these sectors proved useless.

Summarising, we can say that despite the undeniable advances in machine translation, current frameworks for evaluating MT are inadequate, and that a thorough rethinking is required in order to develop more useful technologies meeting professional needs. This calls for more work in data preparation and annotation, formulating representative metrics and developing new technologies (interactive translation and/or TM pre-translation, devices for managing translation flows, and so forth.

Undoubtedly, developing a common European evaluation framework is an ambitious project requiring synergistic efforts on the part of all the actors in the world of translation. But is also a challenge that Europe, with its multilingual strengths, is certain to overcome.

References

Bahdanau Dzmitry, Cho KyungHyun, Bengi Yoshua (2014). "Neural Machine Translation by Jointly Learning to Align and Translate". *ICLR*. https://arxiv.org/pdf/1409.0473.pdf

Bhardwaj Shivendra, Hermelo David Alfonso, Langlais Phillippe, Bernier-Colborne Gabriel, Goutte Cyril, Simard Michel (2021). "Human or Neural Translation?". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona: International Committee on Computational Linguistics, 6553–6564. DOI: 10.18653/v1/2020.coling-main.576

Devlin Jacob, Chang Ming-Wei, Lee Kenton, Toutanova Kristina (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Volume 1 (Long and Short Papers)* di *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis: Association for Computational Linguistics, 4171–4186. DOI: 10.18653/v1/N19-1423

Espla Miquel *et al*. (2019). "ParaCrawl : Web-scale parallel corpora for the languages of the EU". In: *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*. Dublino: European Association for Machine Translation, 118–119. https://aclanthology.org/W19-6721.pdf

Frenette Xavier (2021). *Utilisation du plongement du domaine pour l'adaptation non supervisée en traduction automatique*. Tesi di laurea magistrale, Università di Montréal. DOI: https://doi.org/1866/26528

Koehn Philipp (2005). "Europarl: A parallel corpus for statistical machine translation". In: *Proceedings of Machine Translation Summit X: Papers*. https://aclanthology.org/2005.mtsummit-papers.11.pdf

Marie Benjamin, Fujita Atsushi, Rubino Raphael (2021). "Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers". In: *Vol. 1: Long Papers*. *Proceedings of the joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language*. Association for Computational Linguistics, 7297–7306. https://aclanthology.org/2021.acl-long.566.pdf

Papineni Kishore, Roukos Salim, Ward Todd, Zhu Wei-Jing (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311-318. https://doi.org/10.3115/1073083.1073135

Pham MinhQuang *et al.* (2021). "Revisiting Multi-Domain Machine Translation". *Transactions of the Association for Computational Linguistics*, 9:17–35. DOI: 10.1162/tacl_a_00351

Specia Lucia *et al.* (2020). "Findings of the WMT 2020 Shared Task on Quality Estimation". In: *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics,743-764.https://aclanthology.org/2020.wmt-1.79.pdf

Sutskever Ilya, Vinyals Orion, Le Quoc V. (2014). "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems*. https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf

Varshney Neeraj *et al.* (2022). "ILDAE : Instance-Level Difficulty Analysis of Evaluation Data". In: *Volume 1: Long Papers : Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Dublin: Association for Computational Linguistics, 3412-3425. DOI: 10.18653/v1/2022.acl-long.240

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Lukasz, Polosukhin Illia (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. DOI: https://doi.org/10.48550/arXiv.1706.03762

Zerva Chrysoula *et al.* (2022). "Findings of the WMT 2022 Shared Task on Quality Estimation". In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Abu Dhabi: Association for Computational Linguistics, 69-99. https://aclanthology.org/2022.wmt-1.3.pdf

Ziemski Michal *et al.* (2016). "The United Nations Parallel Corpus v1.0". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portoroz: European Language Resources Association (ELRA), 3530-3534. https://aclanthology.org/L16-1561.pdf