Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6018–6023
Marseille, 20-25 June 2022

On "Human Parity" and "Super Human Performance" in Machine Translation Evaluation

Thierry Poibeau
Laboratoire LATTICE
CNRS & École normale supérieure/PSL & Université Sorbonne nouvelle
1, rue Maurice Arnoux – 92120 Montrouge, France
thierry.poibeau@ens.psl.eu

Abstract

In this paper, we reassess claims of human parity and super human performance in machine translation. These terms have already been discussed, as well as the evaluation protocols used to achieved these conclusions (human-parity is achieved *i*) only for a very reduced number of languages, *ii*) on very specific types of documents and *iii*) with very literal translations). However, we think it is necessary to consider these questions again. We show that the terms used are themselves problematic, and that human translation involves much more than what is embedded in automatic systems. We also discuss ethical issues related to the way results are presented and advertised. Finally, we claim that a better assessment of human capacities should be put forward and that the goal of replacing humans by machines is not a desirable one.

Keywords: Neural machine translation, evaluation, human parity, human factors, ethics,

1. Introduction

Machine translation (MT) has recently achieved impressive results. Thanks to the success of deep learning approaches, state of the art systems have been made available online, which translate quite accurately between a large variety of languages (Bojar et al., 2018; Barrault et al., 2019), although differences from one language to the other remain important. This success is clear, it has been widely acclaimed and end-users can feel this progress very directly in their everyday life. Whereas previous systems were providing translations that needed to be heavily revised and corrected, recent systems provide decent and readable translations most of the time. Of course, the quality depends on a lot of parameters, among others the language pair considered, the type of text and the required quality of the final result.

Following this success, several teams have claimed to have reached super human performance for different pairs of languages. The term "human parity" (Hassan et al., 2018) and even "super-human performance" (Barrault et al., 2019) have been used for example in the context of the Workshop on Machine Translation (WMT) evaluation campaigns[1]. This means that MT is supposed to be equal or even to have surpassed the quality of human translation, at least in this context. Practically, during the evaluation, the translation produced by a machine was regularly preferred to the one obtained from a professional translator.

This sounds impressive, but should be discussed further and put into perspective. In this paper, we claim that the term ”human parity” and, moreover, ”super human performance” are misleading, in that they concern only very specific types of texts, for a very limited number of languages, in very specific conditions. Authors of course mention this when they discuss their results, see for example (Hassan et al., 2018), but paper titles, and thereafter headlines in the press, leave apart these as details, making the general public confused about the real state of the art [2].

Although our concerns are not new and have already been detailed by different authors, especially (Toral et al., 2018; Toral, 2019; Toral, 2020; Läubli et al., 2018; Läubli et al., 2020), we think it is important to discuss them again, as these terms are themselves problematic. They are reproduced in the media for a general audience without much care, as if MT was a solved task, at least between some languages, whereas it is clear that performance varies a lot from one domain to the other, or from one type of text to the other. MT is not a solved task and has not reached human parity between any language pair (although some results obtained in evaluation conferences on specific data from specific domains may suggest the opposite).

We first describe previous work (Section 2), before detailing why we think it is not really appropriate to claim that MT has achieved "human parity" for some language pairs (Section 3). Finally we show that the notion of usability, especially the interaction between MT and translators (through the post-editing process) is a more fundamental concept that should be better taken into account (Section 4).

[1] See for example: "English to German: Facebook-FAIR achieves super-human translation performance; several systems are tied with human performance." (Barrault et al., 2019)

[2] For example, in the tech magazine TechRadar: "Microsoft's new AI translates Chinese-to-English as well as a human translator" `https://tinyurl.com/2mctxzvc`

2. Previous Work

The terms ”human parity” and ”super human performance” have been used in the Machine Translation community, especially in the framework of the WMT conferences.

Toral et al. (Toral et al., 2018) have demonstrated several shortcomings in the WMT evaluation. The main issues are that the translations to be evaluated were sometimes based on problematic source texts. For example, in some cases, the source text was itself a translation from another language, which entails biases, like issues with phrasing and idiomaticity, and is thus not considered appropriate for evaluation, Another issue is that evaluators were not always professional translators (evaluators include participants themselves, and often remote crowd-workers or “turkers” who tend to prefer calque or word-to-word translation, even if it is not fluid, compared to a professional translation where the translator has appropriately transposed an idea or a concept through a less direct translation). Regarding the latter, it should be noted that the perception of translation quality varies according to the evaluator (i.e. end-users vs MT developers vs professional translators), although professional translators tend to achieve higher inter-annotator agreement (Toral et al., 2018). It is, however, possible that the non-experts used for the task were actually closer to the target audience for the type of texts being evaluated (news translation) than professional translators.

The problems observed during the 2018 WMT evaluation campaign were partially corrected in 2019 (Toral, 2019). Other potential problems in the type of evaluation performed at the WMT event remain: the evaluation proceeded by text segment, and thus cannot really take into account the text as a coherent whole, see for instance (Läubli et al., 2018). (Läubli et al., 2020) come to similar conclusions when reassessing the findings of (Hassan et al., 2018): they show that “perceived quality in human evaluation depends on the choice of raters, the availability of linguistic context, and the creation of reference translations.”

Some efforts are now made to better take into account the translation context (i.e. the larger text) and achieve better evaluations. (Läubli et al., 2020) in particular propose a set of 5 recommendations to avoid the issue of over achievement claims (these 5 recommendations are: (R1) Choose professional translators as raters; (R2) Evaluate documents, not sentences; (R3) Evaluate fluency in addition to adequacy; (R4) Do not heavily edit reference translations for fluency; (R5) Use original source texts).

Whereas these recommendations are a progress in the field (and claims of human of super-human performance have been less frequent these last months), we think that even if the comparison between humans and machine is legitimate to a certain extent, the obtained result is not the same and this comparison may just be misleading when it is just based on purely informative texts (like news). It should also be noted that translation quality assessment is performed very differently depending on the context (research oriented vs industrial oriented for example) and this should be taken into account in comparative experiments.

Lastly, one should note that we do not address machine translation for under-resourced languages in this paper, a domain in which human expertise is still highly valued. It is however quite important to keep this in mind when addressing the place of humans in the current research paradigm (NMT works well for 20-30 languages, and the most popular online systems are available for 100+ languages but bilingual training corpora are then rare. This should be compared to the circa 7000 languages existing in the world).

3. Does it Make Sense to Speak of ”Human Parity”?

Fundamentally, the whole idea of “super-human performance” is in our opinion problematic, since translations are difficult to evaluate and to compare, especially when they do not contain clear mistakes. It also gives the impression that MT is a problem solved, at least for some languages (and, if the results are said to be “super-human” for some languages, one can assume that automatic systems will soon be better than humans for a large number of other languages). A quick look at current systems clearly shows that we are still far from any “human parity”.

3.1. MT Produces Very Literal Translations

Neural machine translation generally works at the sentence level, although more and more systems try to go beyond the sentence boundary to take a larger context into account, see for example (Popel et al., 2019; Popel et al., 2020). It was already the case with statistical MT (and segment-based MT) (Koehn, 2009), but this previous approach suffered a lot from a more fundamental issue, since it had to assemble different fragments of texts that were not always fully compatible, hence the frequent problem with ill-formed sentences within this research paradigm. NMT, by directly manipulating representations at the sentence level, thanks to transformers (Vaswani et al., 2017), does not suffer so much from this kind of problems.

However, it should be noted that, despite these recent advances in the field, machine translations remain quite literal, see for example (Fonteyne et al., 2020). This is due to the approach itself: NMT is still mainly based on knowledge inferred from large collections of parallel data, where one sentence in the source language corresponds to one sentence in the target language. From this point of view, NMT is a direct continuation of the previous segment-based approaches. Equivalences between languages are found inside the sentence at a more or less local level, despite recent attempts to integrate a wider context in the process. The statistical nature of NMT also favours standard translations over more original ones. Moreover, NMT does not include reformulation or paraphrase modules that would allow the system to produce more varied or more global decisions (for example, human translators choose some translation equivalencies depending on the context and the situation, cf. "(human) translators can adapt word choice according to different communicative demands and circumstances of language use" (Frankenberg-Garcia1, 2021). All this explains why, despite powerful architectures based on transformers and accessing the whole sentence in one go, translations remain literal and based on local equivalencies.

The consequence is an accurate but very literal way of translating, that is not always convenient. This especially explains the large discomfort and disdain expressed by professional translators, especially when the target is a rather elaborate text. The question is not to translate literature, but even for articles in newspapers, NMT is often judged too literal

3.2. MT Works better on Purely Informative Texts

As we have just seen, the nature of the text to be translated plays a major role. Evaluations (and especially the annual WMT evaluations, within the workshop on machine translation) have shown that performance is now comparable with those of humans for some language pairs. However, this evaluation has been traditionally performed on short news, i.e. informative texts written in a direct and simple style so that they can be easily re-used by journalists to write more complete articles. The content of such texts can be translated quite literally and generally does not require much rephrasing in the target language (Reiss, 1981). This is highly specific, and does not cover the full range of needs and types of text one may need to translate.

The WMT conferences organizers are aware of the problem and have recently introduced different tasks and different types of texts appropriately. However, the main evaluation (and claims of human parity performance) are still based on the main track concerning news translation. It is however interesting to have a look at the different test suites and changes of domain to examine how robust the systems are.

When one looks at other tasks, results are mixed and predictable issues appear with terminology and domain specific phrasing. For example, for WMT 2019, a subtask consisted in translating "audit reports and agreements from English to Czech, without domain adaptation (Vojtěchová et al., 2019)." Although "syntax and overall understandability was scored on par or better than the human reference", "Terminological choices [were] a little worse" (Barrault et al., 2019). On the other hand, "the micro-study on agreements reveals that even very good systems produce practically useless translations of agreements because none of them handles document specific terms and their consistent translations whatsoever." (*idem*) Audit agreements are still technical texts and are not literature. Beyond terminological problems, other issues can be observed in more complex types of documents concerning syntax and semantics.

3.3. Super Human just does not Make Sense

Toral (2020) explains that the term "super-human" actually refers to recent advances in artificial intelligence, especially games like Go, where the machine can play against itself and, in doing so, can develop new and original strategies (these strategies are now studied by humans and are very different from anything proposed so far). Alpha Go, the program developed by DeepMind, beat the world champion of Go, Lee Sedol, in 2016 and has continued improving since. From this point of view, one can say that artificial intelligence-based Go programs are indeed 'super-human'. But the rules of Go are simple compared to the complexity of languages (the complexity of Go is mainly related to the number of possibilities at each stage of the game), and at least in the end, it is easy to see who has won. With language, and especially translation, there is nothing to win and no unique solution.

It should also be emphasized that the term ''human parity" is problematic by itself, since MT works very differently from humans. Human translation involves understanding the text on two levels, i.e. establishing relations between elements of the text themselves (text internal cohesion) and correspondences between elements of text and the real world (discourse external coherence). While the machine can reasonably be expected to handle issues of text cohesion at some point, discursive coherence is out of its reach, since it requires knowledge of the world.

It is also clear that most current systems work at the sentence level and are, for instance, very bad at dealing with pronouns, when the source and the target language differ in this respect ("Most MT systems at the sentence-level do not have access to adequate context that may be required for the translation of pronouns" (Jwalapuram et al., 2020), see also (Sennrich, 2018). Note however some preliminary attempts to address the problem, like in (Luong and Popescu-Belis, 2016) or (Fu et al., 2019)). For example, when the source language only has one gender and the target language different pronouns for masculine and feminine, one gender has to be chosen, often at random, by the translation system, which is thus often wrong. This kind of mistake is frequent for any decision the system has to make that involves a context larger than a single sentence. This is thus understandable (and probably negligible from a statistical point of view in a BLEU score), but should prevent the use of terms like "human parity".

4. Putting Humans back in the Loop

We have shown in the previous sections that it does not really make sense to speak of human parity when evaluating MT. In this section we defend the idea that translation of non purely informative texts involves specific features that are probably not fully attainable using today's technology.

4.1. NMT and Human Invisibilization

We have seen in the previous section that several studies (esp. (Toral et al., 2018; Toral, 2020) reassessed claims of human parity and highlighted that this is only true for certain types of texts, especially news. Although all studies outline that evaluation on other domains and languages that the ones addressed in a specific paper (generally centered around a specific language pair) remain to be done (Hassan et al., 2018), all are optimistic about generalization capabilities.

However practical experiments have shown that generalizing performance beyond the news domain is far from obvious. A famous example is the translation of the book *Deep learning* (Goodfellow et al., 2016) from English to French using NMT techniques, that has been widely advertised in the press as having been done automatically and praised as an astonishing success (which is undoubtedly the case). The process was in fact far from being fully automatic. A large amount of time had to be devoted to the preparation of the data for *i*) marking in the text technical zones that should not be translated (like pseudo-code and algorithms) and *ii*) manually translating technical terms (a dictionary of 200 technical term was manually elaborated before translation). The automatic translation of the rest of the text was also not perfect, as explained in (Escribe, 2019): ”category shifts and paraphrasing seem to be procedures that the NMT system did not implement, which sometimes caused the output to be too literal”. All this was largely ignored by the press, leading to the impression that a whole book can be translated with very high quality in a few hours (“For comparison, a normal ’human’ translation would have taken a few weeks. The computer only took twelve hours”, `https://today.rtl.lu/news/business-and-tech/a/1253287.html`)[3],

Indeed the preparation of the data and the revision steps look like fundamental in this experiment, that was nevertheless interesting as an example of translation of a complex and technical text, whose length was also remarkable. Instead of “Invisibilizing” human work to prepare the data and revise the automatic translation, everybody would have benefited from putting this forward as a successful collaboration between humans and machines in the domain of translation.

4.2. NMT and Human Creativity

If NMT can accurately translate different kinds of texts, why not try with literature? Some experiments have recently been done in this domain (Toral et al., 2020). Taking the same idea as a starting point, (Fonteyne et al., 2020) evaluate a translation obtained using a standard NMT system (their corpus being the translation from English to French of a novel from Agatha Christie). They report issues like ‘mistranslation’, and problems with ‘coherence’ and ‘style & register’, according to a specific taxonomy developed for this purpose.

Moreover, the problem is maybe more fundamental than this kind of evaluation suggests. Machine translation, by providing accurate translations of informational texts, leads to think that even literary translation is within reach. But NMT may just lack fundamental features of literary translation, precisely the ones put forward in (Fonteyne et al., 2020): an explicit encoding of coherence, style and register for example, even if recent systems try to integrate discourse features in the translation process (Sennrich, 2018). Note that the notion of literary translation is in itself very imprecise: the translation of poetry is certainly out of reach, whereas the translation of some modern prose is probably easier than the translation of complex technical books, full of jargon. Translation quality should also always be discussed. For example, (Matusov, 2019) advocates that NMT could be useful for literature, but observing that “the quality is often high enough to understand and even enjoy the story” seems a very low standard for literature, that precisely goes beyond just transmitting information.

As we see, discussing human parity does not make much sense if one does not also discuss style and creativity in a broader context. Here again, everybody would gain in having a deeper analysis of analogies and differences in ways of reasoning between machines and humans (whatever reasoning could mean for a machine).

4.3. NMT and Human translators

As we have seen, terms like “human parity” and “super human parity” are problematic because they suggest that machines can replace humans (this is explicitly stated in (Popel et al., 2020) for example: “deep learning may have the potential to replace humans in applications where conservation of meaning is the primary aim”).

This may have some advantages, from a certain point of view. First, and even if this is obvious, it needs to be reminded: There are lots of contexts in which texts would

[3] See also, on the MILA website: “A team of experts has managed this masterstroke in just two and a half months, thanks to artificial intelligence (AI). In fact, most of the work was completed within 12 hours by a machine translation tool based on deep learning. The following 10 weeks were devoted to a “human” revision aimed at correcting any inaccuracies, along with providing clarity and nuance.” `https://mila.quebec/livretraduitia/`. See the term “masterstroke”, and the absence of details about the preparation of the data. A quick search shows that at least 3 months have been devoted to prepare the experiment, cf. “Trois mois de développement ont été nécessaires pour préparer l’IA à ce défi” (`https://tinyurl.com/y6tff9bp`). The reusability of parts of this groundwork is possible, but remains uncertain.

not be translated and thus would not be accessible by people who do not know the target language. Machine translation makes these texts accessible, and this is certainly a good thing. On the other hand, machines have the potential to reduce time and costs, and of course, they already replace humans in some contexts, for example MT agencies already widely use NMT (and humans then more and more intervene as post-editors, and not as pure translators) (Koponen, 2016). But should this really be seen as a desirable feature?

This situation leads to an understandable apprehension from translators, as detailed in (Breyel and Grass, 2019): “there is a fear for translators of losing their jobs or seeing the price of their work drop. There is also a reluctance of professional translators towards post-editing”, this reluctance being mainly due to the fact that “post-editing is considered an alienating activity devoid of creativity.” One can advocate that this is unavoidable and just corresponds to the evolution of techniques, like power looms automated the textile industry and led to less weavers during the 19th century, but the implication of this state of affairs should be discussed in detail. The idea that technology can replace humans cannot be put forward as a success without some precautions and some discussion.

The complex relations between AI and the future of work is a question that goes beyond the scope of this paper (Benhamou, 2020). It would be interesting to study more precisely if automatic translation is just a threat for translators, or if it can also open new doors by decreasing the cost of translation (for example for publishers who cannot afford to translate all their books, but could use MT to evaluate the potential of some books in a given language). The minimum is at least to consider these questions seriously and not just put forward the replacement of humans as an unquestionable desirable advantage of technology.

5. Conclusion

In this paper we have shown why terms like “human parity” or “super human performance” are misleading and do not bring much benefit to the evaluation process. We have examined several important features to consider, like the type of texts used for evaluation and the literal (vs idiomatic) nature of the foreseen translations. Finally, we propose to better take into human aspects in the evaluation process, and in particular try to see how human can be integrated, rather than just replaced by machines. Humans play an important role for MT, especially in operational contexts, for example by providing an expertise in the preparation of the data, establishing the terminology of a domain and correcting MT output during the post-edition phase. All this should be valued and the goal of NMT should not be to replace humans in the first place.

6. Bibliographical References

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August. Association for Computational Linguistics.

Benhamou, S. (2020). Artificial intelligence and the future of work. *Revue d’économie industrielle*, 169(1):57–88. https://www.cairn.info/revue-d-economie-industrielle-2020-1-page-57.htm.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Koehn, P., and Monz, C. (2018). Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October. Association for Computational Linguistics.

Breyel, C. and Grass, T. (2019). Machine translation and “biotranslation”: an unlikely marriage. *SKASE Journal of Theoretical Linguistics*, 17(4). http://www.skase.sk/Volumes/JTL45/pdf_doc/02.pdf.

Escribe, M. (2019). Human evaluation of neural machine translation: The case of deep learning. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 36–46, Varna, Bulgaria, September. Incoma Ltd., Shoumen, Bulgaria.

Fonteyne, M., Tezcan, A., and Macken, L. (2020). Literary machine translation under the magnifying glass: Assessing the quality of an NMT-translated detective novel on document level. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3790–3798, Marseille, France, May. European Language Resources Association.

Frankenberg-Garcia1, A. (2021). Can a corpus-driven lexical analysis of human and machine translation unveil discourse features that set them apart? *Target. International Journal of Translation Studies*, https://doi.org/10.1075/target.20065.fra.

Fu, H., Liu, C., and Sun, J. (2019). Reference network for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3002–3012, Florence, Italy, July. Association for Computational Linguistics.

Goodfellow, I. J., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA, USA. http://www.deeplearningbook.org.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). *Achieving Human Parity on Automatic Chinese to English News Translation*. arXiv. https://arxiv.org/abs/1803.05567.

Jwalapuram, P., Joty, S., and Shen, Y. (2020).

Pronoun-targeted fine-tuning for NMT with hybrid losses. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2267–2279, Online, November. Association for Computational Linguistics.

Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press, Cambridge.

Koponen, M. (2016). Is machine translation post-editing worth the effort? a survey of research into post-editing and effort. *The Journal of Specialised Translation*, 25:131–148.

Luong, N. Q. and Popescu-Belis, A. (2016). A contextual language model to improve machine translation of pronouns by re-ranking translation hypotheses. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 292–304.

Läubli, S., Sennrich, R., and 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 (EMNLP)*, pages 4791–4796, Brussels. Association for Computational Linguistics.

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., and Toral, A. (2020). A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67, Mar.

Matusov, E. (2019). The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland, August. European Association for Machine Translation.

Popel, M., Macháček, D., Auersperger, M., Bojar, O., and Pecina, P. (2019). English-Czech systems in WMT19: Document-level transformer. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 342–348, Florence, Italy, August. Association for Computational Linguistics.

Popel, M., Tomkova, M., Tomek, J., Łukasz Kaiser, Uszkoreit, J., Bojar, O., and Žabokrtský, Z. (2020). Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(4381).

Reiss, K. (1981). Type, kind and individuality of text: Decision making in translation. *Poetics Today*, 2(4):121–131.

Sennrich, R. (2018). Why the time is ripe for discourse in machine translation. In *Second Workshop on Neural Machine Translation and Generation (invited talk)*.

Toral, A., Castilho, S., Hu, K., and Way, A. (2018). Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium, October. Association for Computational Linguistics.

Toral, A., Oliver, A., and Ballestín, P. R. (2020). Machine translation of novels in the age of transformer. In *Maschinelle Übersetzung für Übersetzungsprofis*, pages 276–295, Dublin, Ireland. BDÜ Fachverlag.

Toral, A. (2019). Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 273–281, Dublin, Ireland, August. European Association for Machine Translation.

Toral, A. (2020). Reassessing claims of human parity and super-human performance in machine translation at WMT 2019. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 185–194, Lisboa, Portugal, November. European Association for Machine Translation.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Vojtěchová, T., Novák, M., Klouček, M., and Bojar, O. (2019). Sao wmt19 test suite: Machine translation of audit reports. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.