Translations and Open Science

Exploring how translation technologies can support multilingualism in scholarly communication

S. Fiorini[1], A. Tezcan[2], T. Vanallemeersch[3], S. Szoc[3], K. Migdisi[3], L. Meeus[3], and L. Macken[2]

[1] OPERAS, Belgium
[2] Ghent University, Belgium
[3] CrossLang, Belgium
susanna.fiorini@operas-eu.org

Abstract. English is by and large considered as the lingua franca of scholarly communication. Such a generalised use has certainly the advantage of facilitating international exchanges, but it also generates inequalities among researchers, and limits the dissemination of scientific knowledge. Although translation could be promptly identified as the solution, scholarly communication has historically been marked by a shortage of human and financial resources to support traditional translation processes. The goal of this paper is to present a multi-user approach to machine translation evaluation for a use in scholarly communication. In particular, the paper introduces the fine-tuning and evaluation methodology set up to comply with the needs of different target user *personas* (translators, researchers, readers). Given the focus of the conference, the paper will describe in more detail the evaluation methodology related to the "Translator" *persona*. The paper will also include general preliminary conclusions, and information about the on-going evaluation work.

Keywords: translation technology · machine translation · machine translation evaluation · multilingualism · open science.

1 Introduction

English is by and large considered as the *lingua franca* of scholarly communication. Such a generalised use has certainly the advantage of facilitating exchanges in an increasingly internationalised research landscape. However, this linguistic dominance also generates inequalities among researchers [1], and limits the dissemination of scientific knowledge within non-English speaking communities [2, 3]. In this context, translation could be promptly identified as a solution to help eliminate language barriers and inequalities in research, according to open science principles. Yet, scholarly communication has historically been marked by a shortage of human and financial resources to support traditional translation processes.

The *Translations and Open Science* project was launched to promote a more structured implementation of translation technologies [4, 5] in order to foster

Proceedings of the International Conference HiT-IT 2023, pages 41–51,
7-9 July 2023, Naples, Italy
https://doi.org/10.26615/issn.2683-0078.2023_004

language equality in scholarly communication as well as the dissemination of knowledge at lower costs and with greater efficiency. The expected deliverable of the project is a technology-based scientific translation service, combining technology tools, digital language resources and human skills. Besides in-domain multilingual data and collaborative translation features, the service is intended to provide scientific translators and researchers with adapted machine translation engines, which will serve different purposes and usage scenarios. A use-case study [6] conducted as part of the project suggested indeed that machine translation is used not only by translators as a productivity aid, but also by researchers as a foreign-language writing assistant, as well as by readers of different profiles, who leverage machine translation for discoverability and gisting purposes.

The goal of this paper is to present our approach to machine translation evaluation for use in scholarly communication. In particular, we will introduce the methodology we set up in order to fine-tune and evaluate machine translation engines, while taking into account the different target user profiles and the associated needs. Given the focus of the conference, the paper will describe in more detail the evaluation methodology related to the "Translator" *persona* and usage scenario. The paper will also include general preliminary conclusions, as well as information about the on-going evaluation work.

2 Machine translation for scholarly communication

The conducted use-case study [6] allowed us to draft an overview of the current translation practices in scholarly communication across a variety of scientific domains. Based on a series of interviews and workshops involving a total of 30 participants[4], the study revealed different levels of acceptance of translation technologies, and in particular machine translation, among scientific translators and researchers according to their domain of specialisation. Although these differences can be partially explained by the very specific characteristics of disciplinary content and writing standards, we also observed an impact of the subjective user attitudes on the acceptance of machine translation. To rely on objective data, not affected by the enthusiasm or the scepticism of users, we decided to carry out an *ad hoc* evaluation in order to assess the relevance of machine translation deployment in scholarly communication.

Dataset collection for machine translation fine-tuning The use-case study showed that more than 80% of the interviewed translators and researchers who use machine translation work with free, generic online engines. For our evaluation, we decided to assess whether fine-tuning can help to produce better machine translation output, especially by taking into account specialised terminology, which is crucial in scientific texts. In order to do so, we collected in-domain parallel language datasets in the English-French language pair in three

[4] 15 scientific translators, 5 researchers, 8 academic publishers, 1 academic librarian, 1 translation technology engineer, all scientific domains combined.

pilot scientific domains. The pilot domains and the characteristics of the collected resources are the following:

1. Climatology and Climate Change (Physical Sciences): 100,563 collected segments, 397 extracted[5] terms; translation direction of the bilingual corpus and evaluation task → English to French;
2. Neurosciences (Life Sciences): 103,175 collected segments, 415 extracted terms; translation direction of the bilingual corpus and evaluation task→ English to French;
3. Human Mobility, Environment, and Space (Social Sciences and Humanities): 112,963 collected segments, 300 extracted terms; translation direction of the bilingual corpus and evaluation task → French to English.

As an example, the Human Mobility, Environment, and Space corpus was split into four subsets: a training set (104,539 segments), a validation set (1,896 segments), a test set for automatic evaluation (2,183 segments), and an evaluation set for human evaluation (4,345 segments, including 954 segments specially selected and collected for human evaluation). A similar approach was used for the two other domains in order to proceed to the evaluation tasks.

Choice of engines to be evaluated Since openness is a core principle of the *Translations and Open Science* project, we primarily considered open-source engines, allowing for customisation. We decided to pick two engines presenting different customisation methods: an engine based on an open-source library with highly customisable, multi-parameter setup (OpenNMT), and an engine allowing for simplified, user-level adaptation (ModernMT). In this way, we wanted to be able to determine what kind of fine-tuning effort (if any) is necessary in order to produce better output with disciplinary texts.

Although it does not comply with the open-source requirement, we also included in the evaluation the engine which is, according to our use-case study, the most used by our target community (DeepL).

Fine-tuning To train and fine-tune the engine based on the OpenNMT library, we started with a from-scratch training on open-source parallel datasets provided in OPUS [7]. This resulted in a generic machine translation model, which we then fine-tuned on the collected specialised datasets.

Concerning ModernMT, we fine-tuned the baseline engine by uploading the collected corpora in TMX format through the dedicated feature provided in the online user interface.

With regard to DeepL, we fine-tuned the baseline engine through the *Glossary* feature for terminology customisation. However, it should be noted that, according to our use-case study, most of our target users do not work with this feature, which is also not supported yet in all the API configurations available.

[5] Automated term extraction from the collected corpora, with human review.

These points could be potential limitations in the case of a large-scale deployment, so they will have to be taken into account together with the evaluation outcomes.

3 Evaluation of machine translation for scholarly communication

After the fine-tuning, we proceeded to an in-domain evaluation of the selected machine translation engines. According to the use cases identified in our previous research, the evaluation was set up to provide information about the usability of the raw machine translation output generated by the evaluated engines in the three scenarios below:

1. A researcher using machine translation as a support to write a paper in a foreign language or to translate a paper into a foreign language;
2. A translator using machine translation to perform post-editing in a computer-assisted translation environment;
3. A reader using machine translation to get an idea of the content of a scientific publication.

The scores relating to specialised terminology compliance in machine translation output are also leveraged to understand whether raw machine translation can be useful to automatically translate publication metadata and therefore improve the discoverability of research in multiple languages.

Automatic evaluation The engines were submitted for automatic evaluation by producing output for in-domain test datasets with both baseline and fine-tuned engines. The outputs produced by the six engines (three baseline and three fine-tuned engines) were compared to reference translations using automatic evaluation metrics such as the statistical metrics BLEU and TER (Translation Edit Rate) and the neural (deep learning based) metric COMET [9]. The MATEO software [10] was used to calculate these metrics. The comparison between the baseline and fine-tuned engines was intended to provide further insight into fine-tuning needs, and in particular to bring additional information about the relevance and the required level of fine-tuning effort in order to improve machine translation output.

Besides calculating metric scores, we also visualised the differences between machine translation outputs. The software used shows the difference on character level between the reference translation and the machine translation output, as well as the character-based edit distance between the two sentences.

Human evaluation As part of the human evaluation task, we evaluated the output of the three machine translation engines which obtained the best scores in automatic evaluation. The evaluation was set up to assess machine translation output usability for the three following *personas* and usage scenarios:

Persona 1 - "Translator": professional translator who masters the source language, is a native speaker of the target language, and has a good knowledge of the domain in question. This *persona* performed an adequacy assessment task, as well as a post-editing task in a dedicated evaluation tool.
Persona 2 - "Expert": researcher specialised in the domain in question, who uses machine translation to (a) translate their scientific publication, (b) write an article in the target language (writing aid), or (c) gist scientific texts that are not written in their native language (reading aid). Having a good to native knowledge of the source and target languages, as well as a perfect command of specialised terminology in both languages, this *persona* performed the same evaluation tasks assigned to the "Translator" *persona*: adequacy and post-editing (see section 3.1.3).
Persona 3 - "Layperson": a person who has at most basic knowledge in the domain (e.g. a non-academic reader or a researcher in a different scientific domain). This *persona* has good to excellent knowledge of the target language and makes use of machine translation to gist educational scientific texts. The participants to this task read text excerpts of 100-200 words, drawn from the evaluation set, in a cumulative self-paced reading view. Based on text characteristics - such as the origin of the excerpt (abstract or full text), sentence length, and lexical variety - the texts were classified into different sets which were submitted to different user groups. The human reference translation was used as a benchmark. Reading time was measured. After reading each excerpt, the users were asked to answer multiple-choice comprehension questions as an incentive to read the text attentively.

The "Translator" evaluation setup Given the focus of the conference, we only present in detail the evaluation methodology for the "Translator" *persona* and usage scenario. As part of this human evaluation subtask, two professional translators, specialising in the domains in question, performed for each domain adequacy and post-editing tasks.

The adequacy task consisted in judging the adequacy of the machine translated segments (sentences) of scientific publications, by assigning a score between 1 and 5. The aim of this task was to assess how adequately the machine translation of the segment expressed the source segment's meaning, and, by consequence, how useful the translation was for gisting and discoverability purposes.

Around 500 segments extracted from scientific papers, reviews and abstracts were shown to each evaluator in the order they appear in the document. For each segment, the evaluators were provided with: (1) the part of the paragraph preceding the evaluated segment, (2) the segment itself, (3) the remainder of the paragraph, and (4) machine translation outputs, randomly ordered to avoid evaluator bias (in this way, the evaluators did not only judge the overall quality of machine translation outputs, but also ranked them implicitly). The evaluators were also provided with the abstracts of the documents from which the evaluation segments originate in order to provide more context. Reference translations were not shown to avoid bias.

The post-editing task consisted in asking the evaluators to produce a publishable translation (a terminologically valid, grammatically correct, fluent translation conveying the meaning of the source sentence), based on a source segment, its context, and a machine translation output. The evaluator was also asked to provide a score for perceived post-editing effort for each segment. This task was performed on a different test set than the one used for the adequacy task. As in the adequacy task, around 500 segments were shown, without reference translation, and in the order they appear in the document. However, only one machine translation output for each segment was provided (the evaluators were provided with output from different engines without knowing which engine had been used to translate a specific source segment).

Three metrics were applied to assess the productivity with each machine engine:

1. temporal effort (average time per word);
2. technical effort based on human-target TER (HTER) scores via measurement of post-editing difference (PEdiff) between machine output and the translation produced by the evaluator;
3. perception of effort (see above).

Samples drawn from the post-edited outputs were annotated using the MQM framework. Error annotations were performed using the seven high-level error dimensions: terminology, accuracy, linguistic conventions, style, locale conventions, audience appropriateness, and design and markup. The output edited as part of the "Expert" *persona* setup were also annotated according to the same standards, in order to determine the relations between error types and editing behaviours based on user profiles (for instance, determining whether one *persona* is more likely to correct terminology errors rather than style).

4 Preliminary conclusions and work in progress

According to the use-case study conducted as part of the *Translations and Open Science* project, machine translation is frequently used to produce multilingual publications in some domains (100% of the life sciences and physical sciences researchers and translators interviewed use machine translation), while in other domains we observed a more reluctant attitude towards this technology (20% of the humanities and social sciences researchers and translators interviewed use machine translation). This data suggests that, besides the required technical efforts, investments in training and literacy programs are also needed in order to efficiently deploy translation technologies, and in particular machine translation, in scholarly communication.

The collection of bilingual scientific datasets for machine translation fine-tuning also raised various challenges. Firstly, the amount of bilingual data available in scientific publications is limited. As we said, translation is not a systematic activity in scholarly communication due to disciplinary standards and

a shortage of resources. Moreover, in most of the cases only abstracts are translated, the full text papers being only available in one language[6]. Also, it can be difficult to determine the origin of the translation, which means that it is not always possible to easily identify and exclude low quality-translations or *translationese* from the test sets. Secondly, a considerable portion of the bilingual data we found is published under licences which expressly forbid data collection and processing, or which do not provide clear information about the authorised uses (∼40% of the identified data sources). We were mainly able to collect and process the publications under Creative Commons licences, according to the conditions established by the applied licence type. For the remaining data, in a few cases we received the authorisation to collect and process data from the right owners, otherwise we relied on the Text and Data Mining exception (TDM), introduced by the European directive 2019/790 and transposed into French law in 2021. When the publication did not fall under the TDM exception and we could not get the required authorisations, we simply did not collect any data (∼30% of the identified data sources). Finally, we observed a general lack of standardisation among data sources when it comes to formats and keyword classification of scientific publications, which can complicate data collection through automated processes.

As for the evaluation task, the automatic evaluation results seem to show that there is no significant improvement in the performance of the fine-tuned versions of DeepL and ModernMT, while the OpenNMT engine does perform better after fine-tuning with the specialised datasets collected as part of the *Translations and Open Science* project, and performs even better after also adding the SciPar corpus [11], which contains parallel corpora from scientific abstracts, all domains combined (Fig. 1-3). However, the overall performance of the OpenNMT engine remains lower than DeepL and ModernMT even after fine-tuning, except for the "Thesis abstracts" document type in the SH7 discipline as well as for the "Review abstracts" and "Thesis abstracts" document types in the LS5 discipline.

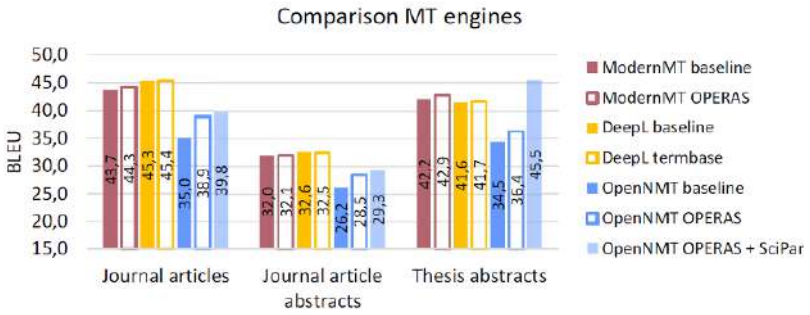

Fig. 1. BLEU scores by engine and document type for the SH7 discipline

[6] Out of the 23 sources from which we collected data, only 9 had some full text papers translated.

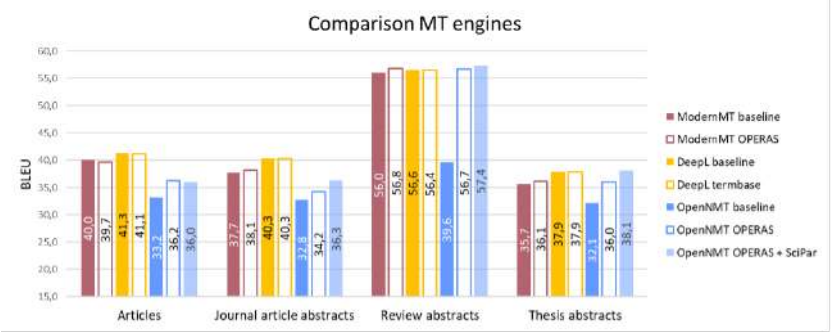

Fig. 2. BLEU scores by engine and document type for the LS5 discipline

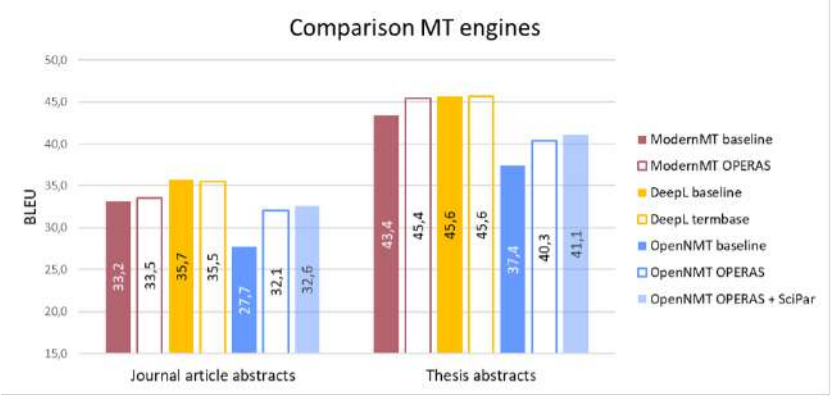

Fig. 3. BLEU scores by engine and document type for the PE3_10 discipline

At a first glance, these results seem to suggest that, given the effort required to collect and prepare parallel data, fine-tuning might not be the most effective strategy to improve machine translation output. We even found cases where the baseline DeepL engine performed better than the fine-tuned one: however, given that fine-tuning in DeepL only covers terminology, this could be due to terminology inconsistencies in the reference translations.

In the light of the evaluation outcome after adding the SciPar corpus to the OpenNMT engine, another hypothesis is that data collection for fine-tuning should not be strictly narrowed to in-domain texts only. This is key information for the general sustainability of our approach.

The human evaluation performed by professional translators overall confirmed the ranking established by the automatic evaluation: for the three disciplines, DeepL tends to have on average the lowest post-edit time and perceived effort as well as the highest user rating in adequacy tasks, followed by

ModernMT and OpenNMT. The MQM annotation results in the same machine translation engine ranking, at least for disciplines SH7 and LS5 (data for the PE3_10 discipline is still under production at the time the present paper is being drafted). Given the importance of specialised terminology in scientific texts, we focused our analysis on terminology errors, which might discourage the use of raw machine translation to automatically translate publication metadata for discoverability purposes (Fig. 4 and 5).

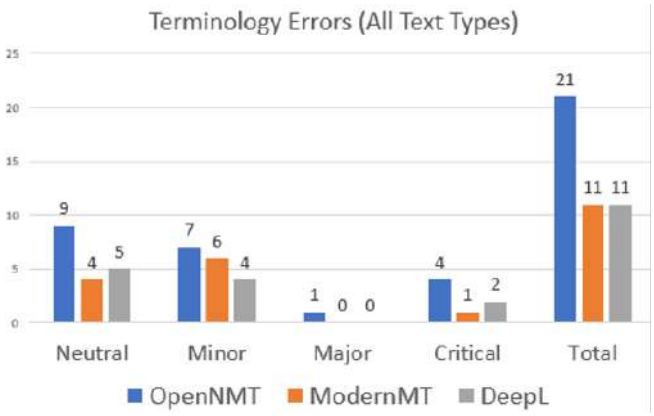

Fig. 4. Terminology errors for SH7 discipline

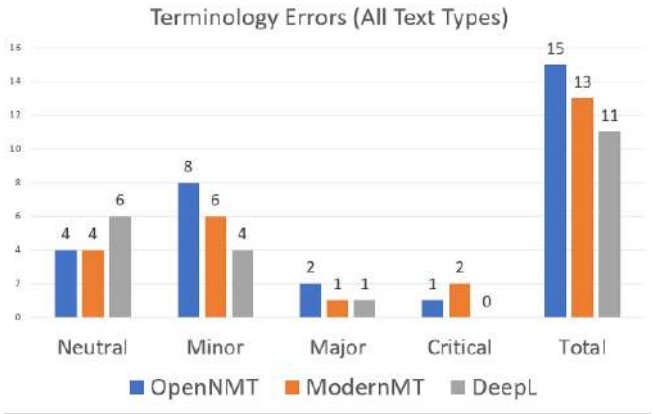

Fig. 5. Terminology errors for LS5 discipline

When it comes to correlations between translators regarding post-edit time, perceived effort, HTER and adequacy ratings (more specifically, Pearson product-

moment correlation coefficient[7]), they range from 35% to 45% in the SH7 discipline, from 35% to 52% in the LS5 discipline, and from 29% to 52% in the PE3_10 discipline. Regarding the machine translation engine ranking obtained, it is plausible to assume that DeepL may have benefited from the extensive use over time of its free version by researchers, while OpenNMT might have been more affected by the variable quality of fine-tuning data. In order to improve the reliability of the results in the future, the study could therefore benefit from the inclusion of a larger panel of evaluators per profile and a clearer view of the nature and quality of the data used for the fine-tuning of machine translation engines.

Acknowledgements The authors acknowledge F. Barbin and K. Hernandez-Morin (Université Rennes 2) for their contribution to the evaluation task of the *Translations and Open Science* project. The project is funded by the French National Fund for Open Science, which includes contributions from the the French Ministry of Higher Education and Research, universities and other research institutions. The project also received a special contribution from the French Ministry of Culture.

References

1. V. Ramírez-Castañeda, 2020, Disadvantages in preparing and publishing scientific papers caused by the dominance of the English language in science: The case of Colombian researchers in biological sciences. PLoS ONE 15(9): e0238372. https://doi.org/10.1371/journal.pone.0238372
2. Di Bitetti, Mario S., and Julián A. Ferreras, 2017, Publish (in English) or perish: The effect on citation rate of using languages other than English in scientific publications, Ambio 46.1: 121-127
3. Z. Taşkın, G. Doğan, E. Kulczycki, A. Zuccala, 2020, Science needs to inform the public. That can't be done solely in English, LSE blog. https://blogs.lse.ac.uk/covid19/2020/06/18/long-read-science-needs-to-inform-the-public-that-cant-be-done-solely-in-english/
4. L. Bowker, J. Ciro, 2019, Machine translation and global research: Towards improved machine translation literacy in the scholarly community, Bingley, UK: Emerald Publishing
5. S. Fiorini et al., 2020, Rapport du groupe de travail "Traductions et science ouverte", Comité pour la science ouverte. 44 p.
6. Use case study for a technology-based scientific translation service, report to be published
7. J. Tiedemann, 2012, Parallel Data, Tools and Interfaces in OPUS. Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), 2214-2218
8. Papineni et al,. 2002,BLEU: a Method for Automatic Evaluation of Machine Translation Kishore Papineni, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318

[7] A measure of linear correlation between two sets of data.

9. Rei et al., 2020, COMET: A Neural Framework for MT Evaluation Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)
10. B. Vanroy, A. Tezcan, L. Macken, 2023, MATEO: MAchine Translation Evaluation Online. In M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, ... H. Moniz (Eds.), Proceedings of the 24th Annual Conference of the European Association for Machine Translation (pp. 499-500). Tampere, Finland: European Association for Machine Translation (EAMT).
11. D. Roussis et al., 2022, SciPar: A Collection of Parallel Corpora from Scientific Abstracts, Proceedings of the Thirteenth Language Resources and Evaluation Conference