Is there a case for accepting machine translated scholarly content in repositories?

May 8, 2023

Christophe Dony, Iryna Kuchma, Tomasz Neugebauer, Jean-François Nomine, Milica Ševkušić, and Kathleen Shearer

Multilingualism is a critical characteristic of a healthy, inclusive, and diverse research communications landscape. However, multilingualism presents a particular challenge for the discovery of research outputs. Although researchers and other information seekers may only be able to read in one or two languages, they may want to know about all the relevant research in their area, regardless of the language in which it is published. Conversely, information seekers may want to discover research outputs in their own language(s) more easily. To facilitate this, COAR Task Force on Supporting Multilingualism and non-English Content in Repositories has been developing and promoting good practices for repositories in managing multilingual and non-English content. In the course of our work, the topic of **machine translation (MT)** has sparked a heated discussion within the Task Group and we would like to share with you the nature of this discussion.

Should MT scholarly texts be uploaded into repositories? And if so, who would be credited as the author and how? Who would bear the responsibility for the translation quality? The COPE position statement on Authorship and AI tools already states that "AI tools cannot be listed as an author of a paper. AI tools cannot meet the requirements for authorship as they cannot take responsibility for the submitted work. As non-legal entities, they cannot assert the presence or absence of conflicts of interest nor manage copyright and license agreements."(1)

Are there any specific contexts or community use cases – e.g. public health emergencies, climate change issues, or biodiversity loss – that would greatly benefit from machine translated (MTed) content in repositories? Human translation capacities are often not available to support the information needs in terms of time and quantity. Is there a role for repositories in terms of providing access to content and research, regardless of the language in which it is written? What kind of innovations in indexing, translation and searching would we like to see that would enable a more nuanced approach, rather than an outright ban on exclusively MTed content in repositories? If uploading MTed content into repositories were to significantly advance the community to a more accessible multilingual universe, wouldn't it be harmful to preemptively dismiss that approach? Time is important and the search for the perfect solution means potentially decades more of inaccessible content due to language barriers.

But how can we ensure accuracy in translations? How can we avoid cognitive biases (2) and ensure that AI produced content doesn't discriminate or reinforce existing stereotypes (3)? Who will critically analyze the results of MTed content? Or is it sufficient to just tag exclusively MTed content to make the readers aware of it, with a caveat notice? Should it be left up to the users to apply a critical eye and interpret it as they see fit? What would be the ethical and legal implications of this? Would a warning or "disclaimer" notice for MTed content deposited in repositories be sufficient and effective, such as the one below?

"This document/This material is a machine translation [of : [citation of original]] from [source language code] into [target language code]. This machine translation has not been reviewed or edited and is provided "as is" for the sole purpose of assisting users in understanding at least part of the subject matter of the original content expressed in [source language]. This provision does not imply a guarantee of correctness and accuracy of the said machine translation [in target language] by any natural or legal person in any part of this translation. [Consequently, the provision of this translation shall not give rise to any liability on the part of any person to any other person in the event that this translation is used for any purpose whatsoever.] Users of this machine translation are expressly invited to have it checked, revised or edited by a professional translator or relevant expert."

If yes, how would one actually deposit MTed content into a repository – together with the original document – or as a separate, but related record?

What about the author's point of view – the original author(s) might not be happy with the quality of the MT. Authorisation of authors is required for human translation – should this also be the case for MTed content as well? Or is it sufficient to have clear reuse rights (an open content license)?

Shall repositories be mentioned in the relevant institutional policies and regulations on the use of AI? Shall we also have clear recommendations on MTed content in repository policies – e.g. in the quality control and curation sections? Such a complex issue definitely requires institutional-level decisions, policies and regulations.

These are the questions that have arisen in the Task Group as we grapple with recommendations related to repositories and machine translations.

Machine translation today

Translations are invaluable – "today, language translation is the new frontier of information exchange."(4) For decades MT technologies have been widely used as language mediators (5), but there are only about 100 languages for which MT systems have been built, out of more than 7 000 (un)official languages spoken around the world. This results in MT systems being "highly skewed towards European languages due to two main bottlenecks which are lack of digitized languages and lack of translated text",(6)

In addition to language (under)representation within existing MT tools, the quality of existing translations is variable. According to the paper, *Not lost in translation: The implications of machine translation technologies for language professionals and for broader society*, "even when languages are represented, the quality of the translation varies markedly, for example because of differences in the amount of resources that are available to train AI systems". (7) Linguistic limitations and human-coded biases are worth mentioning as well (8). For example, there is still a question whether MT is fully able to translate meaning and not just texts.

One of the most consequential issues is that, "at the moment, high-resource languages, i.e. languages for which large quantities of training data are available in digital form and for which MT tools work best, are the same languages dominating scientific and social exchanges. Unless this changes, for example because more digital examples of text become available, improvements in AI technologies could further exacerbate the concentration of prominence to a restricted set of languages rather than reduce existing disparities." (9) Translation to/from low-resource languages would require more intensive post-editing, which would have to be done by comparatively small communities familiar with these languages. These communities are more likely to invest their limited time and energy into improving the translation of local language content into English, with the rationale that this would make it accessible to a wider audience.

In his 2017 dissertation, *Integrating machine translation with institutional repositories: a case study of university of Nairobi*, Julius Nyabuti Mugoya explored the viability and challenges of the use of MT for digital collections in the University of Nairobi institutional repository (IR). The premise of the study is that research outputs in a repository should be available in the language of the user. Since university funding is getting tighter, free software should be considered for some tasks such as translation. The specific objectives of the study include: "finding out the suitability and convenience of using MT in translating IR resources in the UoN repository; evaluating the most suitable free MT system for translating IR resources from English to Kiswahili; identifying the benefits and/or constraints encountered when doing so, and; generally identifying the future role of MT in digital information dissemination." (10)

As an example, Mugoya refers to the 2015 Ebola outbreak in Guinea, Liberia and Sierra Leone, where Ebola-related materials were predominantly in English and French, yet this region has over 90 local languages. Biodiversity conservation is another case cited in the study, where about a third of analysed materials was available in languages other than English. Mugoya's proposition is to increase the value of IR resources to researchers through MT by developing a way through which a user can switch between different language versions eliminating the processes involved in copying and pasting into a translation app or even trying to decipher the meaning. Unfortunately, Mugoya's results confirm that most MTs are developed with European languages in mind. The methods and rules of MT for Kiswahili do not work well as there are structural differences between European and African languages (e.g. African languages use verbs with a complex conjugation rule set that cannot be deployed within the rule set of European languages). The findings showed that MT has not gained as much accuracy yet as human translation when translating from English to Kiswahili, however, the translations rendered are still suitable to give an indication of what the content is about. The study recommends that a widget to aid users of the IRs be included that would detect and align the required translations of the source text.

Summing up, having accurate translations of the full text content in repositories is useful. This could potentially improve discoverability of diverse research outputs in aggregators and in web search engines. However, available MT tools are currently unable to ensure equally accurate translations across languages and disciplines. Despite major improvements over the years, MT for translation of scholarly literature is still lacking sufficient training corpora (even in widely-used language combinations) given the vast fields of knowledge across the entire spectrum of academia / research and is still a subject of exploration and experimentation. Caution calls for refraining from relying on "good enough" results, even if improved by apparently fluent machine output. MT, however, is useful when there is no need to rely on the accuracy of information contained in a text, "MT tools can help users understand the broad meaning of a text, a process which is called gisting" (11). Yes, for much of the content in repositories, accuracy and reliability are of critical importance.

Accountability and authorship

Even if the quality were satisfactory, the inclusion of MTed full-text content in repositories may have serious social and legal implications. For example, there are issues of accountability related to the issue of authorship. In cases where a mistranslated document is used, who is accountable? Transferring accountability to the person depositing an exclusively MTed resource into a repository may be problematic or difficult to explain or enforce.

Additionally, there is a high risk of mistranslation or poorly translated scholarly content if there is no human mediation of the MTed text. This is less problematic for shorter texts such as abstracts or non-scientific texts. But the risks of possible misinterpretation/mistranslation seem to be higher when dealing with unverified MTed content: "In particular, as societies struggle to limit the spread and the deleterious effects of misinformation, disinformation and malinformation, the use of MT technologies could have the unintended effect of making such problems more acute. To the extent that un-checked translations of varying levels of accuracy and quality are made available to a large number of individuals and are accessed by individuals without a critical understanding of the nature of the translation process, MT tools could exacerbate existing problems associated with online content." (12)

Another possible consequence might be that professional translators' work is devalued, especially if translations are taken into account in research assessment in the future and if it is not clear that records are just MTs only (when cited).

There are important nuances in terms of approaches to machine translations. For example, one could consider a structure with different levels of translation, such as: 1. MT only; 2. post-edited after MT – when human translators are called in to work on the outputs of MT (13) (14); and 3. human translation. Yet these levels or degrees of translation would very much be subject to interpretation and would need to be clarified. This also raises questions as to where to draw the line between post-edition and human translation. The extent to which human translators use MT tools in their routine work is unknown.

Machine translation and repository workflows

Repositories are complex social and technical infrastructures focusing on curation and preservation of content. MTs can introduce significant complications and bring many outstanding questions to the institution. In order to be trustworthy, a repository needs transparent policies, quality control and preservation strategies. How does MTed content fit into curation best practice and workflows? What aspects of a submission containing MTed text would be covered by quality checks – just the technical integrity of the file or the accuracy of the content? Would a repository manager be expected to reject such a submission based on the poor quality of translation and, if so, who would set the criteria guiding such decisions? Who would be able to submit an improved version of such a translation and how? Would the old version be preserved in a repository even if removed from public view – is it worthwhile preserving it – or would it be deleted? In the latter case, the institution might not be happy to deal with Persistent ID tombstones, dead external links, and their possible effects on the users' perception of the repository's trustworthiness. Repository and institutional policies would have to be adapted significantly to address these challenging issues.

An "instant" MT could be a useful service (i.e., in line with the current technology at the time of retrieval) when the MTed content isn't deposited in the repository, but there is an API/link to live MT with an appropriate clearly visible notice with a waiver in the language of the search interface. On the other hand, repository users might already use such tools by themselves as there are web services readily available online. Reader's working knowledge of the original language cannot be assumed, because the original language may not be a widely used; and by experience, longer scholarly texts cannot be released simply MTed and unpost-edited because they require post-editing (not to say supplementary scientific revision) to reach a satisfying level of relevance, consistency and quality.

But, what about post-edited MTed content? Could it be of value when produced with scholarly guarantees matching scholarly requirements? Would translated metadata and abstracts be the next option? For example, There is a Europeana Translate Generic Services project to enrich records with MT metadata, which explores how MTs can be used to help make sure everyone can find what they need on Europeana, no matter what language they use. (15)

Conclusions

Given the complexity of the issues as well as the ethical implications, the Task Group has chosen to recommend that repositories not accept exclusively MTed content. (16) This is also in line with the "Report of the Translation and open science' working group" (2020). Rather, MT should be perceived and used as an assistive technology and allowed to change dynamically in real time, transparently and unambiguously labeled as machine assistance, rather than curated and preserved as a primary resource in the repository. The Task Group will monitor this rapidly evolving landscape, and continue to consider the issues and possibly publish further recommendations related to machine translation for scholarly texts, MT-assisted translation, as well as MT of abstracts and metadata in repositories.

Endnotes

1. https://publicationethics.org/cope-position-statements/ai-author
2. "It is important to note that ChatGPT is not governed by ethical principles and cannot distinguish between right and wrong, true and false. This tool only collects information from the databases and texts it processes on the internet, so it also learns any cognitive bias found in that information. It is therefore essential to critically analyse the results it provides and compare them with other sources of information." – ChatGPT and Artificial Intelligence in higher education: Quick start guide. Published in 2023 by the United Nations Educational, Scientific and Cultural Organization, 7, place de Fontenoy, 75352 Paris 07 SP, France and the UNESCO International Institute for Higher Education in Latin America and the Caribbean (IESALC), Edificio Asovincar, Av. Los Chorros con Calle Acueducto, Altos de Sebucán. Caracas, 1071, Venezuela https://www.iesalc.unesco.org/wp-content/uploads/2023/04/ChatGPT-and-Artificial-Intelligence-in-higher-education-Quick-Start-guide_EN_FINAL.pdf
3. "On the one hand, this reflects the lack of female participation in subjects related to AI and in research/development on AI and on the other hand, the power of generative AI to produce and disseminate content that discriminates or reinforces gendered and other stereotypes."; "Machine translation technology performs less accurately when words or texts that are translated from a text that is rather gender-neutral are translated to a language that is not. For example, this could lead to machine translation technology only providing a single-gendered translation or using masculine translation by default in male-dominated fields." – Borgonovi, F., J. Hervé and H. Seitz (2023), "Not lost in translation: The implications of machine translation technologies for language professionals and for broader society", OECD Social, Employment and Migration Working Papers, No. 291, OECD Publishing, Paris, https://doi.org/10.1787/e1d1d170-en
4. Ibid
5. "On one hand, the number of unique translatable language pairs increased from around 16 000 in 2019 to around 150 000 in 2022. On the other hand, MT technologies today can translate some texts with a high level of accuracy – although the quality of translations is variable and depends on the algorithms being used by the language models, the quantity, quality, and variety of the translations used to train the machine learning algorithms that are at the basis of AI MT tools as well as the complexity of the text that is translated." Ibid
6. Ibid
7. Ibid
8. "Linguistic limitations include obstacles related to untranslatable words and contextual meanings in cross-language translations, cultural context and cultural expectations, and ever-evolving nature of languages. In every language, words can be found that capture a very fine-grained meaning and for which no translation to another language exists…Additionally, MT technologies have difficulties evaluating or recognising metaphorical meanings, interpreting and translating hidden or subtle messages, or identifying contextual meanings that are not literal, such as humour, irony, or sarcasm." Ibid
9. Ibid
10. http://erepository.uonbi.ac.ke/handle/11295/104580
11. Borgonovi, F., J. Hervé and H. Seitz (2023), "Not lost in translation: The implications of machine translation technologies for language professionals and for broader society", OECD Social, Employment and Migration Working Papers, No. 291, OECD Publishing, Paris, https://doi.org/10.1787/e1d1d170-en
12. Ibid
13. "Post-editing skills are among the extended set of linguistic skills which are demanded when MT tools are integrated into professional translators' workflow… Post-editing means that once MT output is generated, an editing, amending and correction process is conducted by human translators to achieve high quality translations. Because post-editing needs to be done by someone proficient in the target language, post-editing is mostly carried out by professional translators… Post-editing training for example includes knowledge on various kinds of machine translation systems or machine translation error analysis. This is important since machine translation accuracy has, despite strong advances over the past years, not been able to reach a human level of language proficiency, due to limitations related to linguistics, and biases." Ibid
14. "On the broader labor market effects of MT, the emergence of AI might lead to the emergence of a new category of language professionals, not exactly translators but rather "post-editors" who would be paid less than traditional translators. Post-editors (PE) would devote less time and effort on the MT pre-translated text than to traditional translation projects, given that PE tasks tend to have relatively lower pay but identical, if not tighter, deadlines. Furthermore, if the jobs of translators were to become more digitized, the profession may face increasing expectations for remote working arrangements" Ibid
15. https://pro.europeana.eu/post/building-on-state-of-the-art-machine-translation-services
16. Some of the reasons exposed above are advanced in the report "Translations and open science" (p.19). Here is the excerpt in French that addresses some of theses concerns – though the report does adopt a more nuanced stance on this. "Il est cependant fondamental de se poser les questions suivantes : ces traductions automatiques instantanées, qui de facto seraient des textes « éphémères » ne faisant pas l'objet d'un travail éditorial, peuvent-elles être citées ? Peuvent-elles être indexées et référencées systématiquement ? Sans doute non. Par ailleurs, il ne faudrait pas non plus favoriser la circulation de fausses informations pouvant découler de faux positifs de traduction automatique ; parfois, le texte brut traduit automatiquement est tellement fluide que les erreurs deviennent presque indétectables lors d'une simple lecture." – Susanna Fiorini, Franck Barbin, Martine Garnier-Rizet, Katell Hernandez Morin, Franziska Humphreys, et al.. Rapport du groupe de travail "Traductions et science ouverte". [Rapport Technique] Comité pour la science ouverte. 2020, 44 p. (hal-03640511)

See this table for a summary of the Task Group Discussions