MaTOS: Machine Translation for Open Science

Rachel Bawden[2], Maud Bénard[1], Éric de la Clergerie[2], José Cornejo Cárcamo[1], Nicolas Dahan[2,4], Manon Delorme[3], Mathilde Huguin[3], Natalie Kübler[1], Paul Lerner[4], Alexandra Mestivier[1], Joachim Minder[3], Jean-François Nominé[3], Ziqian Peng[2,4], Laurent Romary[2], Panagiotis Tsolakis[2], Lichao Zhu[1], François Yvon[4]

[1] ALTAE, Université Paris Cité, Paris, France
[2] Inria, Paris, France
[3] INIST, CNRS, Nancy, France
[4] ISIR, CNRS and Sorbonne Université, Paris, France

Correspondence: yvon@isir.upmc.fr

Abstract

This paper is a short presentation of MaTOS (Machine Translation for Open Science), a project focusing on the automatic translation of scholarly documents. Its main aims are (a) to develop resources (term lists and corpora) for high-quality machine translation, (b) to study methods for translating complete, structured documents in a cohesive and consistent manner, (c) to propose novel metrics to evaluate machine translation in technical domains. Publications and resources are available on the project web site: https://anr-matos.fr.

Motivations

MaTOS, Machine Translation for Open Science, is a four-year project (2022-2026) aiming to develop new methods for the full machine translation (MT) of scholarly documents, as well as automatic metrics for evaluating the quality of the translations produced. Our main target application is the translation of scholarly articles between French and English, for which linguistic resources can be exploited to obtain high-quality translations. These translation can both be used to speed up publication in a foreign language, but also to improve the discoverability of scientific information, and facilitate its dissemination to the general public. However, efforts to improve the MT of entire documents are hampered by the inability of existing automatic MT metrics to detect and evaluate translation issues that span multiple sentences. Such issues are not rare and happen due to inadequate modeling of discourse phenomena.

MaTOS is part of a growing trend in research and technology to automate the processing of scholarly articles, providing new tools to discover and process an increasing volume of publications. MT is one the most important technologies in this regard, as it holds the promise of facilitating the global discussion about the current state of scientific knowledge beyond the scientific community, where these discussions take place mostly in English. Using MT, other critical applications of scholarly text mining can also be made available in multiple languages, e.g. bibliometric analysis and the automatic detection of plagiarism and articles reporting falsified conclusions. MaTOS will contribute to this general trend by (a) developing new open resources for specialized MT, (b) improving the description of textual consistency markers for scholarly articles, through the study of terminological variation, (c) studying new multilingual processing methods to handle long documents, and (d) proposing dedicated automatic metrics for these tasks.

Challenges

New neural machine translation (NMT) architectures can handle extended contexts, corresponding to paragraphs or even longer parts of documents. However, notwithstanding the limitations of existing computational architectures, efforts to improve MT for complete documents are hindered on the one hand by a general lack of resources, and on the other by the inability of existing automatic metrics to detect system weaknesses and identify the best ways to remedy them.

MaTOS tackles these two difficulties head-on, paying particular attention to the issue of translating complex terms and their variation within documents. The translation of specialized terms, which is critical for academic texts in particular, remains difficult, due to the specific linguistic structures in which the terms appear (e.g., complex nominal phrases), the lack of a stabilized reference translation for emerging terms and the lack of modeling of their variation within texts and corpora.

Participants

MaTOS is a multidisciplinary project, bringing together teams with diverse scientific backgrounds: the ALMAnaCH project-team at Inria, Paris[1] and the MLIA team at ISIR[2] bring expertise in natural language processing and are mostly involved in the technological workpackages, developing methods and tools to automatically identify terms and their variants, to perform translation at the document level and to automatically evaluate whole document translation. ALTAE[3] (Université Paris-Cité), will focus on the development of resources (annotated corpora and term lists) and conduct fine-grained studies of the terminological variation within and across scholarly documents. INIST-CNRS,[4] will also contribute to resource development, in line with their primary missions related to the dissemination of scientific and technological information.

Results

After two years, the project has produced a set of reports documenting the state of the art, focusing notably on (a) human assessments of translation quality (Bénard et al., 2024), (b) automatic evaluation of translation at the document level (Dahan et al., 2024) and (c) computational architectures for document translation (Peng et al., 2024).

Various resources have also been collected, prepared and formatted. These include terminologies for two domains ("Natural Language Processing" and "Earth and Planet Sciences"), as well as monolingual and bilingual corpora, in particular long documents and their translations for the same domains. They can be downloaded from our website.

Regarding natural language processing, efforts have focused on three aspects: (a) the development of tools to identify terms and their variants in corpora (these will be used to document in detail the spectrum of acceptable terminological variations and to thoroughly evaluate the quality of terminology translation at the document level), (b) the study of methods for the automatical suggestion of neologisms for the translation of emerging terms (Lerner and Yvon, 2025), and (c) the development of specialized MT systems able to translate long documents, based on both encoder-decoder architectures and large multilingual language models. Preliminary results are in (Peng et al., 2025).

Regarding evaluation, two pilot studies involving the post-editing of automatically translated abstracts have been carried out with the involvement of specialized translators and members of the academic community, in anticipation of a larger-scale study (Bawden et al., 2024).

Acknowledgments

This research was funded by the French Agence Nationale de la Recherche (ANR) under the project MaTOS - "ANR-22-CE23-0033-03". R. Bawden's participation is also partly funded by her chair in the PRAIRIE institute funded by the ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001.

References

Rachel Bawden, Ziqian Peng, Maud Bénard, Eric Villemonte de La Clergerie, Raphaël Esamotunu, Mathilde Huguin, Natalie Kübler, Alexandra Mestivier, Mona Michelot, Laurent Romary, Lichao Zhu, and François Yvon. 2024. Translate your Own: a Post-Editing Experiment in the NLP domain. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, page 431–443, Sheffield, United Kingdom.

Maud Bénard, Natalie Kübler, Alexandra Mestivier, Joachim Minder, and Lichao Zhu. 2024. Étude des Protocoles d'Évaluation Humaine pour la Traduction de Documents. Technical Report Delivrable D4.1.1, Projet ANR MaTOS.

Nicolas Dahan, Rachel Bawden, and François Yvon. 2024. Survey of Automatic Metrics for Evaluating Machine Translation at the Document Level. Technical Report Deliverable D4.5.1, Projet ANR MaTOS.

Paul Lerner and François Yvon. 2025. Towards the machine translation of scientific neologisms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 947–963, Abu Dhabi, UAE.

Ziqian Peng, Rachel Bawden, and François Yvon. 2024. Handling Very Long Contexts in Neural Machine Translation: a Survey. Technical Report Delivrable D3.2.1, Projet ANR MaTOS.

Ziqian Peng, Rachel Bawden, and François Yvon. 2025. Investigating length issues in document-level machine translation. In *Proceedings of Machine Translation Summit XX, Vol. 1: Research Track*, Geneva, Switzerland.

[1] https://almanach.inria.fr
[2] https://www.isir.upmc.fr/equipes/mlia
[3] https://clillac-arp.u-paris.fr/
[4] https://www.inist.fr/