Assisting the maintenance of multilingual documents

Yehor Korotenko[1]

[1]Université Paris-Saclay | LISN

Adaptive Machine Translation, Large Language Models, Markup Languages, Natural Language Processing

Question

Scientific narratives such as course notes, textbooks or manuals are constantly evolving. Can Machine Translation help author and maintain multilingual versions thereof?

Use cases

- **Educational content**: university lecture notes, online courses, open textbooks.
- **Scientific documentation**: technical manuals, software guides, reproducibility documents.
- **Collaborative platforms**: Markdown/LaTeX-based materials shared by international teams.
- **Needs**:
  - Frequent updates → translations must adapt quickly.

Desirable Features

- **Syntax**
  - Preserve markup.
  - Preserve non-linguistic spans.
- **Style & Long-Term Maintenance**
  - Preserve and learn from post-edits and stylistic decisions.
  - Maintain consistency across revisions.
- **Domain-Specific Vocabulary**
  - Preserve and learn from the terminology chosen by the author.
  - Preserve context-specific meanings.
- **Usability and ethics**
  - Integrate easily in authoring workflows.
  - Preserve Privacy and Intellectual property.
  - Reduce environmental impact.

Model Agnostic Approach

- **Markup-Preserving Pipeline**
  - Parse & segment documents into **chunks**.
  - Quote markup within tags to leverage the widespread ability of translation models to preserve XML.
  - ↳ Preserves **markup** while allowing **correct reordering**.
  - Never alter code/math.

  Example: The following **LaTeX** code

```
Let $f$ be an \textbf{endomorphism} in $\R^n$
```

  is quoted as:

```
<TEXT>Let <PH original="$f$"/> be an <PH original="\textbf{"/
>endomorphism<PH original="}"/> in <PH original="$\R^n$"/></TEXT>
```

- **Translation Memory & Correspondence Database**
  - Record translation pairs of chunks with **checksums**.
  - Detect unchanged or slightly modified chunks.
  - Reuse or slightly modify existing translation to:
    - Speedup translation, reduce cost/energy.
    - Maintain **terminological & stylistic consistency**.
    - Preserve post-edits.
- **Vocabulary Dictionary**
  - User-provided translation pairs for key terminology:
    - Applied only to natural language.
    - Ensures **domain accuracy** and reduces post-editing.
- **Library** providing the core algorithms and approaches for handling marked-up narratives (LaTeX, ...).
- **Command-line** tool built on top of the library, offering direct integration into typical authoring workflows.

Evaluation

- **Models Compared**
  - **Gemini-2.0-flash** → strongest quality, proprietary.
  - **LLaMA-3.3-70B** → robust with markup, hosted at Paris-Saclay.
  - **Gemma3-27B** → runs locally, weaker with syntax preservation.
- **Environmental impact**
  - Translating 72 files of Python course costs around 6 Wh.
  - ↳ That is equivalent to charging a smartphone from 0% to about 15–20% [1].
- **Results Overview**

Task	Best Model	Notes
Plain text translation	Gemma3-27B (fluency)	Struggles with long context
Markup-heavy documents	LLaMA-3.3-70B	More robust than Gemma
Overall quality	Gemini-2.0-flash	Proprietary , less privacy-friendly

Key Contributions

- **XML-based syntax-preserving translation.**
- **Translation memory** for sustainable updates.
- **Vocabulary dictionary** for domain precision.
- **Multilingual** version tracking.
- **Library and command-line tool**.

Project github repository

Extended abstract about the project

Conclusion & Future work

- Open-weight LLMs are competitive in **plain-text**, but limited for **syntax-heavy** tasks.
- Our **model-agnostic strategies** bridge this gap.
- Next steps:
  - Extend support to more markup formats.
  - Support synchronous collaborative authoring.
  - Better learn from post-edits.
  - Include document context when translating chunks.

Bibliography

[1] Y. Korotenko, "Translation Prototype Report." [Online]. Available: https://dobbikov.github.io/sci-trans-git/prototype_report.pdf

[2] Y. Moslem, R. Haque, J. D. Kelleher, and A. Way, "Adaptive Machine Translation with Large Language Models." 2024.

[3] Z. Zhu *et al.*, "LaTeXTrans: Structured LaTeX Translation with Multi-Agent Coordination." 2025.

[4] H. C. Kleidermacher and J. Zou, "Science Across Languages: Assessing LLM Multilingual Translation of Scientific Papers." 2025.

Examples

- Math notes written in LaTeX and their translations:

Original (fr)

Translation (en)

Translation (ua)

LaTeX source code

Acknowledgments

I would like to thank my internship supervisor, Nicolas M. Thiery, for his valuable advice, support and suggesting the project.

Yehor Korotenko

dobbikov.com

yehor.korotenko@universite-paris-saclay.fr