Translating Mathematical Texts

Clemens H. Cap

September 2024

1 Introduction

Machine translation has reached a pretty advanced stage with DeepL probably being one of the best systems available on the market. However, the translation of texts which contain mathematical formulae, diagrams or images still poses problems. This is especially true when the document is given as a specification language, such as LaTeX, TikZ, or XML description languages.

For example, the text

```
\section{Example}
```

should not be translated into

```
\Abschnitt{Beispiel}
```

but rather into

```
\section{Beispiel}
```

When tables, diagrams, references or footnotes are used, more problems arise. Parameters and portions of the LaTeXlanguage must be identified and properly translated or protected against translation. The core issue is the proper separation of textual contents from markup, diagram and formula contents. Only the text portions should be translated and the other parts should be placed into their proper places of the document flow.

Currently, DeepL offers only rudimentary support for translating documents which contain XML-tags. This feature can be used to protect formula content from tampering by the translation mechanism.

Other approaches may be found in the literature.

2 Research Questions

1. Provide an overview on the state of the art for translating mathematical texts.
2. Analyze how the XML-translation features of DeepL can be used for protecting the non-textual portions of a document from being treated incorrectly by the translation system.
3. Develop an algorithm which works with the core features of LaTeXand a small number of its most important extensions.
4. Implement this algorithm, preferably in PHP, as an integration with a Mediawiki-based system is our primary application.
5. Develop test documents for evaluating the capabilities of your concepts.
6. Provide a report on the limits of the system you have developed.