POLENG MT: An Adaptive MT Platform

Artur Nowakowski[1,2]**, Krzysztof Jassem**[1,2]**, Maciej Lison**[1]**, Kamil Guttmann**[1,2]**, Mikołaj Pokrywka**[1,2]

[1] Poleng, Poznań, Poland

[2] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland

`{name.surname}@poleng.pl`

Abstract

We introduce POLENG MT, an MT platform that may be used as a cloud web application or as an on-site solution. The platform is capable of providing accurate document translation, including the transfer of document formatting between the input document and the output document. The main feature of the on-site version is dedicated customer adaptation, which consists of training on specialized texts and applying forced terminology translation according to the user's needs.

1 General Description

POLENG MT is an MT translation platform available in two versions. Using PaaS (Platform as a Service), the translations are delivered via a cloud web application. In the on-site scenario, the customer organization receives an installation package to be used in the customer's infrastructure. In this case, access to the service is specifically limited to the customer's employees. The following features are shared by both versions of the platform:

- user registration and login;

- document import in `.txt`, `.docx`, `.pptx` and `.xlsx` formats;

- document editing in sentence-by-sentence mode;

- machine translation in an editing window;

- machine translation of entire documents;

- export of the translated document in a format compatible with the imported document;

- pre-translation of documents using translation memory fuzzy search matches;

- ability to proofread and approve translations of sentences;

- expanding translation memory with approved translations;

- transfer of document formatting (fonts, styling, text placement) between input and output document;

- archiving of translated documents per user.

POLENG MT translation models are based on the Marian (Junczys-Dowmunt et al., 2018) and fairseq (Ott et al., 2019) NMT frameworks.

2 Customer Adaptation

Adaptation for specific users is carried out in the on-site versions. The task includes the following processes:

- SSO (single sign-on) login integration, if applicable;

- delivery of a translation engine specialized in the customer's domain, fine-tuned on documents provided by the customer;

- incorporation of a customized lexicon into the NMT engine;

- automatic generation of a lexicon from the customer's documents.

The latter two processes take into account the recognition and generation of inflected forms of lexicon entries. This problem is addressed in (Nowakowski and Jassem, 2021) and (Bergmanis and Pinnis, 2021).

3 Supported Languages

Currently, POLENG MT supports the following language pairs, in both directions:

- Polish–English;

- Polish–Ukrainian;

- Polish–Russian.

In the near future, we plan to add support for language pairs with other Eastern European languages, including Czech, Romanian, Bulgarian and Belarusian.

Upon the customer's request, the POLENG MT platform can support any translation direction, on condition that the customer provides suitable parallel data (for example, in the form of business documents and their translations).

References

Bergmanis, Toms and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online, April. Association for Computational Linguistics.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Nowakowski, Artur and Krzysztof Jassem. 2021. Neural machine translation with inflected lexicon. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 282–292, Virtual, August. Association for Machine Translation in the Americas.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.