Machine Translation - from research to production

Mikołaj Koszowski, Karol Grzegorczyk, Karol Karpinski

{mikolaj.koszowski,karol.grzegorczyk,karol.karpinski}@allegro.pl

allegro ML Research

Overview

There are over 250 million offers at Allegro. Titles and descriptions for most of them are created in Polish. To enable non-polish speakers to use our platform, we need to translate the offers into other languages. Doing this by human translators would take ages. Therefore we need Machine Translation.

There are many high-quality MT providers. However, relying only on cloud-provided MT solutions has two disadvantages: it's not tailored to the e-commerce domain and is quite expensive at our scale. Therefore, we decided to create our own in-house Machine Translation engine.

Text Corpora

We have multiple sources of training data. They can be divided into 2 groups: **Parallel** and **Monolingual**. Following our previous work [1], we compute features for all of our raw data. Some of the features are specific for parallel corpore while other can ba applied to all our data. Later we experiment with different filtering thresholds. At the beginning we set them based on outliers from clean corpora.

	Data Sources	Data Filtering
Parallel	▪ publicly available open-source corpora ▪ a corpus extracted from a database of human-translated product descriptions	▪ Bicleaner AI [2] score (trained on e-commerce corpora) ▪ probabilistic modification of pair length ratio ▪ edit distance and digit sequence mismatch
Monolingual	▪ backtranslation of e-commerce domain subset of monolingual text data ▪ backtranslation of monolingual product descriptions ▪ parallel en-xx corpora translated from English to Polish	▪ length based ones: max, min, avg of characters and words ▪ character based: alphabet dissimilarity and digit ratios ▪ language identification probability based on fastText

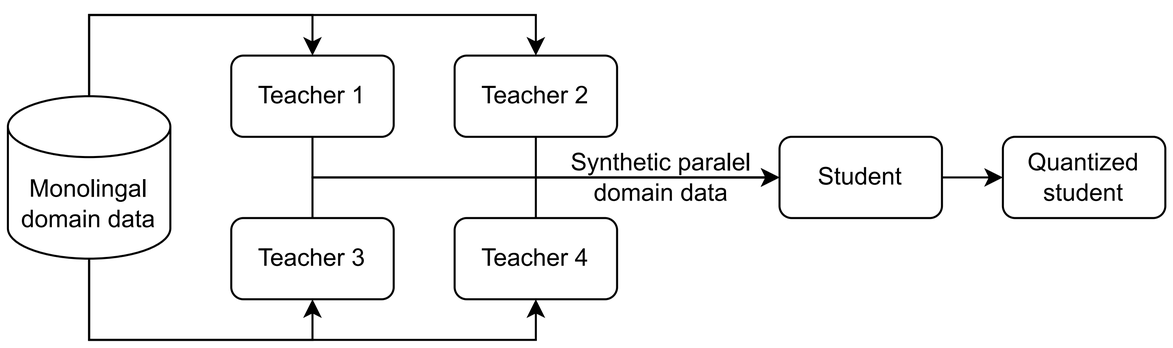
Training

- Training using Marian MT [3] on preemptible VMs with NVIDIA A100 GPUs in Google Cloud Platform
- Model architecture: vanilla **Transformer BIG** (213 mln params) trained from random weights
- Dynamic data filtering for each training job based on JSON filtering definitions
- Automatic conversion of Marian logs to the TensorBoard format and upload to Vertext AI Experiments
- For back-translation we use output sampling with the Gumbel noise

Google Cloud TensorBoard MARIAN NMT

Distillation

1. translate over 40 mln high-quality polish sentences from Allegro product descriptions using an ensemble of teachers into a target language
2. join source sentences with target ones forming a synthetic parallel corpus
3. overfit a student (with the same architecture as the teachers) to this synthetic corpus.

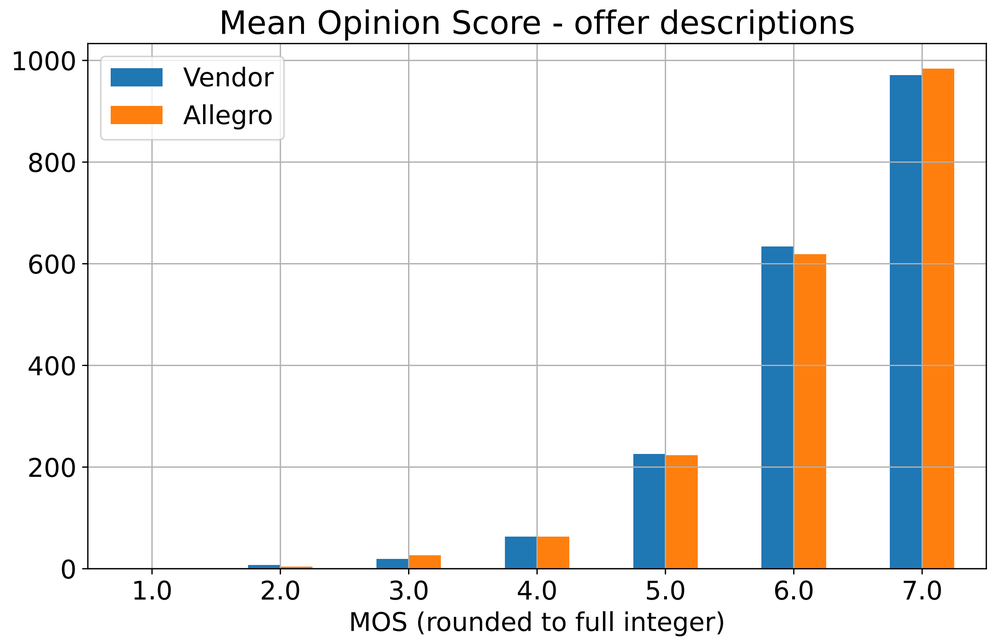


Collaboration

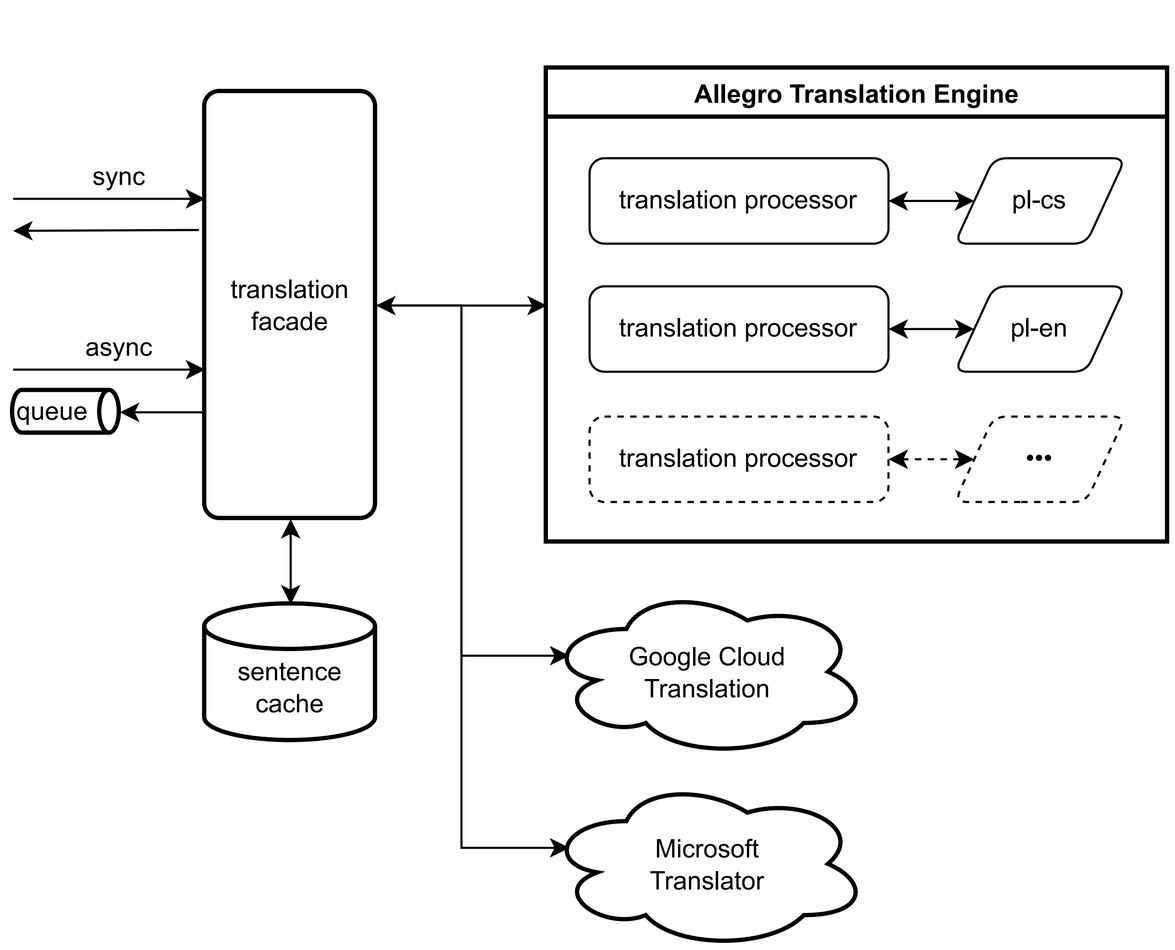
THE UNIVERSITY of EDINBURGH · ADAM MICKIEWICZ UNIVERSITY POZNAŃ · DEFINED.AI

Evaluation

We have validation-sets made from representative sample of **a few thousand sentences** from a product descriptions. During training we calculate **BLEU, chrF** and when finished we calculate **COMET** [4] which is our main automatic metric, due to high correlation with human judgment. We base our decision about starting a new evaluation campaign on it. Our annotators are **multiple bilingual experts**, they evaluate each translation in isolation, but have access to: a title, a product picture and a full description. Additionally, we are working on more qualitative evaluation setup based on **MQM**, but for now we use **Likert scale**:
1 - nonsensical translation to **7** - perfect, the meaning and grammar fully correct.



Serving



- synchronous and asynchronous endpoints for client convenience
- sentence caching for cost and computation power savings
- multiple translation providers integrated
- in-house translation provider implemented - **Allegro Translation Engine**
- serving with CTranslate2 on CPUs in our own DCs
- can be deployed in GCP on GPU
- translation processor enables placeholder support for glossaries, no translate phrases, URLs, e-mails, emojis, etc.

References

[1] M. Koszowski, K. Grzegorczyk, and T. Hadeliya, "Allegro. eu submission to wmt21 news translation task," in *Proceedings of the Sixth Conference on Machine Translation*, pp. 140–143, 2021.

[2] J. Zaragoza-Bernabeu, G. Ramírez-Sánchez, M. Bañón, and S. Ortiz Rojas, "Bicleaner ai: Bicleaner goes neural," in *Proceedings of the Language Resources and Evaluation Conference*, (Marseille, France), pp. 824–831, European Language Resources Association, June 2022.

[3] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in C++," in *Proceedings of ACL 2018, System Demonstrations*, (Melbourne, Australia), pp. 116–121, Association for Computational Linguistics, July 2018.

[4] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "Comet: A neural framework for mt evaluation," *arXiv preprint arXiv:2009.09025*, 2020.