

HPLT’s First Release of Data and Models

Nikolay Arefyev[♣] Mikko Aulamo[★] Pinzhen Chen[♡] Ona de Gibert[★]
Barry Haddow[♡] Jindřich Helcl[♣] Bhavitvya Malik[♡] Gema Ramírez-Sánchez[◇]
Pavel Stepachev[♡] Jörg Tiedemann[★] Dušan Variš[♣] Jaime Zaragoza[◇]
[♣]University of Oslo [★]University of Helsinki [♡]University of Edinburgh
[♣]Charles University [◇]Prompsit Language Engineering
<https://hplt-project.org>

Abstract

The High Performance Language Technologies (HPLT) project is a 3-year EU-funded project that started in September 2022. It aims to deliver free, sustainable, and reusable *datasets*, *models*, and *workflows* at scale using high-performance computing. We describe the first results of the project. The data release includes monolingual data in 75 languages at 5.6T tokens and parallel data in 18 language pairs at 96M pairs, derived from 1.8 petabytes of web crawls. Building upon automated and transparent pipelines, the first machine translation (MT) models as well as large language models (LLMs) have been trained and released. Multiple data processing tools and pipelines have also been made public.

1 Introduction

The HPLT project combines petabytes of natural language data and large-scale model training. Focusing on different aspects of the project, there are eight partners in the consortium: Charles University in Prague (coordinator), University of Edinburgh, University of Helsinki, University of Oslo, University of Turku, Prompsit Language Engineering, and CESNET and Sigma2 HPC centres.

The project has achieved several milestones in its first half (Sep 2022–Feb 2024). Specifically, 1) we developed essential data tools and training pipelines; 2) we successfully produced the first edition of datasets for 75 languages and 18 language pairs; 3) we trained and released the first batch of LLMs and MT models.

2 First Release

Datasets For the first release, we processed 1.85 petabytes of the Internet Archive and Common-Crawl to create monolingual and parallel corpora. We release them under the permissive CC0 licence¹ through our project website², OPUS³, and Hugging Face⁴. We also publish through GitHub open-source tools and pipelines to process huge web archive data packages⁵ so that our real use case can serve as an example for others inside and outside of the research community. The monolingual data spans 75 languages and contains roughly 5.6 trillion space-separated tokens after deduplication. The bilingual ones focus on low- to medium-resourced languages and cover 18 language pairs, with roughly 1.4 billion tokens computed on the English side and 96 million sentence pairs.

MT Models The main aim of the first HPLT MT model release was to bring together all the tools in the MT model pipeline, to show that they are capable of building a suite of MT models in a mostly automated fashion. The model building also helped us to extrinsically examine the quality of the first HPLT data release—to see if it influences performance when combined with the much larger existing parallel data on Opus. For this reason, we built bilingual models for all the language pairs included in the first HPLT parallel data release.

The release of the MT model weights is through Hugging Face (HF).⁶ These are available in both HF and Marian formats, compatible with the transformers library and MarianNMT framework.

¹We do not own any of the text from which these text data have been extracted. We license the actual packaging of these text data under the CC0 licence (“no rights reserved”).

²hplt-project.org/datasets/

³opus.nlpl.eu/HPLT.php

⁴huggingface.co/datasets/HPLT/hplt_monolingual_v1_2

⁵github.com/hplt-project

⁶huggingface.co/HPLT

There is also a repository⁷ containing the scripts to download and process the data, and train and evaluate the models. Third-party users can use this repository, together with our tool chain, to completely reproduce our models.

The tooling for the model-building pipeline includes OpusCleaner (for selecting and cleaning training data), OpusTrainer (a data scheduling and data augmenting tool), and OpusPocus (for managing the training process itself). The first two were described in our previous report (Aulamo et al., 2023), whereas OpusPocus is described below.

Pipelines and Tools Besides a significant effort in establishing the data production pipelines, HPLT also develops data analytics, dashboards, and training pipelines.

HPLT Analytics⁸ provides a full range of analytics automatically computed on either monolingual or bilingual datasets to help make informed decisions. It shows corpora details, volumes, language, lengths, noise, quality score distributions, and others. Support for language-dependent components has been added for dozens of languages. Automated reports in YAML and PDF are generated from the web application to which a corpus can be uploaded and processed.

OpusPocus⁹ is an MT training pipeline manager that abstracts and automates the repetitive parts in training: data preparation, model training, and fine-tuning. A user can run the default training pipelines without knowledge about the implementation details, simply having their training data and an execution command. OpusPocus’s main features are: 1) Python implementation given a large user base; 2) modularity: each pipeline step is isolated from others and only requires the outputs from its dependencies; 3) separation of pipeline execution and monitoring; 4) separation of task definition and task execution. In our workflow, OpusCleaner and OpusTrainer could be wrapped in it.

OPUS-MT dashboard¹⁰ (Tiedemann and de Gibert, 2023) is an interface to the OPUS-MT leaderboards that systematically collect benchmark results of publicly available neural MT models. It provides various views on results on a wide range of language pairs for common benchmarks such as WMT test sets, FLORES200, and NTREX. The

dashboard makes it possible to compare models of different sizes and different language coverage to facilitate the selection of appropriate solutions for specific applications. Translations of test sets can also be inspected with highlighted string differences to reference translations or the output of alternative models. The tool currently provides performance information for thousands of open MT models including OPUS-MT models and others in the Hugging Face model hub.

OpusDistillery¹¹ is an end-to-end pipeline for multilingual MT sequence-level distillation to train efficient NMT models. We develop on top of the Firefox Translation Training pipeline (FTT)¹² from the Bergamot project¹³. We have added support for using pre-trained OPUS-MT models, GPU tracking, and multilingual training.

3 Future Plan for Machine Translation

In our next MT model release, we target three aspects. First, we aim to deliver MT models for even lower-resourced languages that HPLT has acquired. Next, we will investigate cost-efficient MT models via distillation, quantization, etc. Finally, in comparison with the current release of unidirectional translation models, we plan to explore massively multilingual models and large language models.

Acknowledgment

This project has received funding from the European Union’s Horizon Europe research and innovation programme under Grant agreement No 101070350 and from UK Research and Innovation (UKRI) under the UK government’s Horizon Europe funding guarantee [grant number 10052546].

References

- Aulamo, Mikko, Nikolay Bogoychev, Shaoxiong Ji, Graeme Nail, Gema Ramírez-Sánchez, Jörg Tiedemann, Jelmer van der Linde, and Jaume Zaragoza. 2023. HPLT: High performance language technologies. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*.
- Tiedemann, Jörg and Ona de Gibert. 2023. The OPUS-MT dashboard – a toolkit for a systematic evaluation of open machine translation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*.

⁷github.com/hplt-project/mt-models

⁸github.com/hplt-project/data-analytics-tool

⁹github.com/hplt-project/OpusPocus

¹⁰github.com/hplt-project/OPUS-MT-dashboard

¹¹github.com/Helsinki-NLP/OpusDistillery

¹²github.com/mozilla/firefox-translations-training

¹³browser.mt