HPLT: High Performance Language Technologies

Mikko Aulamo★, Nikolay Bogoychev‡, Shaoxiong Ji★, Graeme Nail‡, Gema Ramírez-Sánchez†, Jörg Tiedemann★, Jelmer van der Linde‡, Jaume Zaragoza†
★University of Helsinki, ‡University of Edinburgh, †Prompsit Language Engineering
https://hplt-project.org/

Abstract

We describe the High Performance Language Technologies project (HPLT), a 3-year EU-funded project started in September 2022. HPLT will build a space combining petabytes of natural language data with large-scale model training. It will derive monolingual and bilingual datasets from the Internet Archive and CommonCrawl and build efficient and solid machine translation (MT) as well as large language models (LLMs). HPLT aims at providing free, sustainable and reusable datasets, models and workflows at scale using high-performance computing (HPC).

1 Introduction

The HPLT project aims at innovating the current language and translation modelling landscape by building the largest collection of free and reproducible models and datasets for around 100 languages. Datasets will be derived from web-crawled data using already established processing pipelines from the ParaCrawl[1] and MaCoCu corpora.[2] They will be adapted and improved to run efficiently on HPC centres in order to produce consistent datasets at scale. HPLT will also build open, sustainable and efficient LLMs and MT models with significant language coverage using the powerful supercomputing infrastructure of European HPC centres. Datasets, models, pipelines and software to build them will be shared along with additional tools to ease data management, model building and evaluation.

An HPC-powered consortium: The consortium gathers research groups, the experience of an industry partner, and the computational infrastructure and involvement of two HPC centres in Europe. Most of the processing will happen on LUMI, a pre-exascale supercomputer, which will be made NLP-aware to pave the way for further initiatives and exploitation of the project outcomes. The 8 partners in the consortium are: Charles University in Prague, University of Edinburgh, University of Helsinki, University of Oslo, University of Turku, Prompsit Language Engineering, CESNET, and Sigma2 HPC centres.

2 Expected Results

Datasets: Starting from 7 PB of web-crawled data from the Internet Archive[3] and 5 from CommonCrawl,[4] we will derive monolingual and bilingual datasets for systematic LLM and MT building with a large language coverage. Data curation, a crucial part of the process, will be based on adapted versions of the Bitextor and Monotextor pipelines[5]. Filtered and anonymized versions enriched with genre information will be released. Output formats will follow commonly adopted standards and their distribution will be handled through OPUS[6] and LINDAT[7] with open-source licenses along with analytics and metadata.

Models: Efficient and high-quality language and translation models will be built and released. Regarding LLMs, when sizes and computational resources allow, we aim at building BERT (Devlin et al., 2019), T5 (Raffel et al., 2020), and GPT-like models (Brown et al., 2020) for all the targeted languages. We will opt for multilingual models where necessary to mitigate the lack of sufficient training data that is expected for some of the targeted languages. For MT models, we plan to build

[1] https://paracrawl.eu/
[2] https://macocu.eu
[3] https://archive.org/
[4] https://commoncrawl.org/
[5] https://github.com/bitextor/
[6] https://opus.nlpl.eu/
[7] https://lindat.mff.cuni.cz/

not only English-centric models but also other language combinations including multilingual models depending on data availability and interest. We will share HPLT models through OPUS-MT and HuggingFace with open-source licenses. The first HPLT LLMs have already been published: GPT3-like models for Finnish[8], still under evaluation.

Pipelines and Tools: HPLT wants to ease data management and model building, making HPC centres in Europe ready to run the same pipelines and tools in a transparent and straightforward manner even on other datasets and languages. Below, we describe two of the tools that HPLT is developing in this direction.

OpusCleaner[9] is a one-stop dataset download/examine/filter toolkit built with modern large-scale NLP models in mind. It is based on *python* and uses a web interface to make it easy to run on HPC clusters. The workflow is as follows: (1) dataset selection: downloads to the host running the web server, not the local machine; (2) filter selection: allows filtering and visualizing the effect interactively on a random sample of each selected dataset; (3) labeling: allows categorising each dataset; (4) batch filter execution: applies filters and labeling to all datasets from a one-line runme command and (5) dataset (near-)deduplication across collections.

OpusTrainer[10] is a large-scale data shuffler/augmenter which takes a collection of datasets and feeds it to a neural network training toolkit according to a set schedule. Its design aims to solve neural network training problems at scale. It features: (1) sampling and mixing of data from multiple sources; (2) per-source shuffling and independent dataset mixing avoiding out-of-memory issues; (3) curriculum learning with the definition of training stages, each one having its own mixture of datasets; (4) stochastic modifications of the training batch to support end-user requirements like support for title case, all caps, placeholders, etc.

3 MT at HPLT

HPLT's ambition is to democratise access to efficient MT. We will use our large curated datasets with robust software pipelines to train high-quality MT systems and, by leveraging the HPC capacity available to the project, over an extensive set of languages. All models will be properly evaluated and documented using standard metrics. Releasing all models with appropriate metadata and optimised training recipes will also help to avoid unnecessary computation for sub-optimal and repetitive procedures. Beyond large systems, we aim to build lightweight models using knowledge distillation (Kim and Rush, 2016). An ensemble of large teacher models can produce compact students that mimic their teacher's quality, with negligible degradation but much lower computational costs during inference. Quantisation and other efficiency techniques can further increase speed and lower the memory footprint, which is essential for responsive and large-scale translation tasks.

[8] https://turkunlp.org/gpt3-finnish
[9] shorturl.at/boLW7
[10] shorturl.at/pDKPT

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]. The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union.

References

[Brown et al.2020] Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

[Devlin et al.2019] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[Kim and Rush2016] Kim, Yoon and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

[Raffel et al.2020] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.