WP4: High Performance Language Models

HPLT online kickoff 19.9.2022

TURKUNLP
.ORG

Group

TurkuNLP is 20+ years old and has more than 20 members

Substantial focus on **large LM training** and use in last ~4 years, with perhaps half the group working with large LMs

Two members **working on HPLT now**; both have been working on large LM training on supercomputers

Overview

WP4 optimizes, builds and evaluates language models (LMs). (cf. WP5: machine translation models)

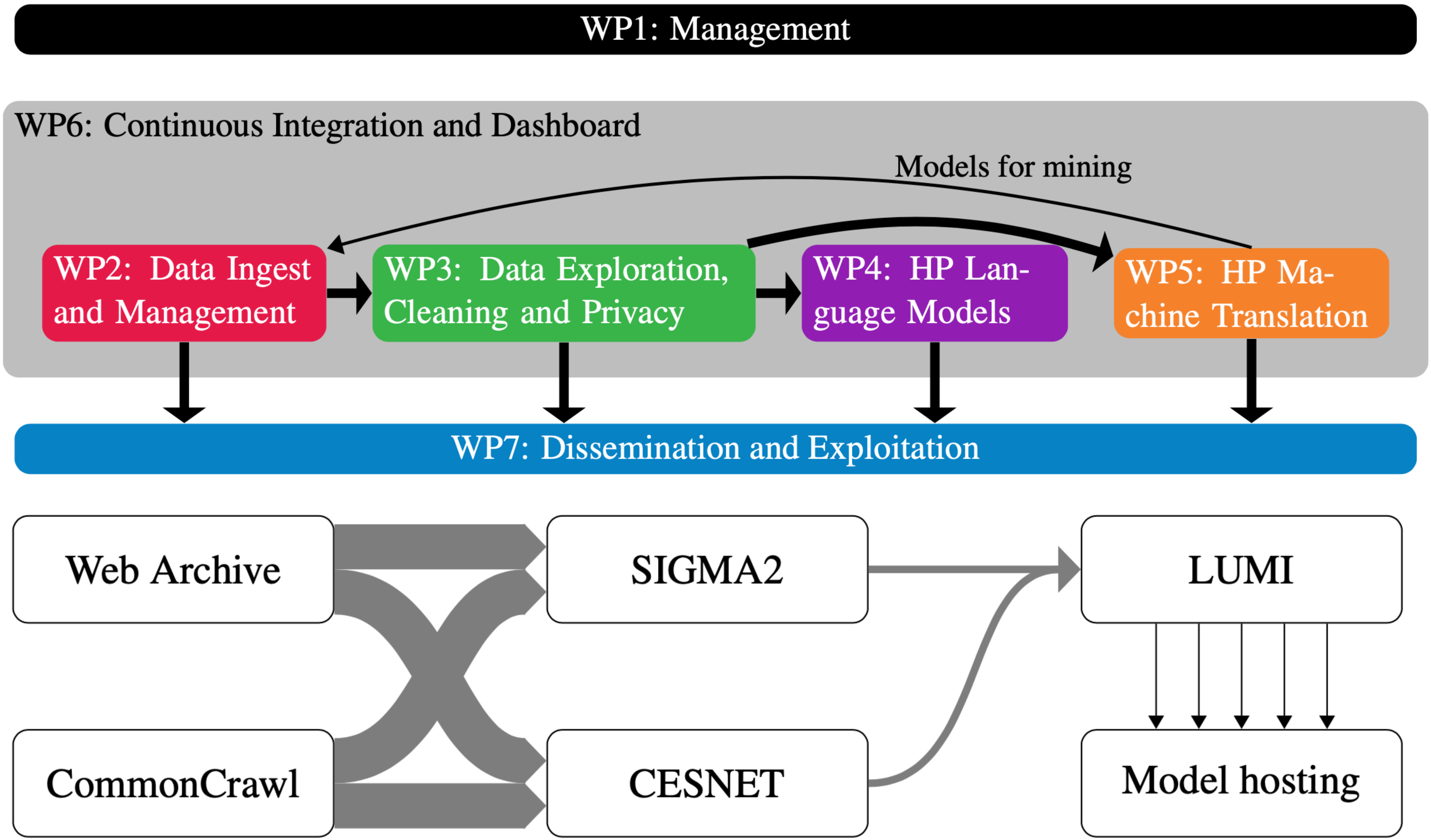
- Pretrain **BERT**-, **GPT**-, and **T5**-like models
- Cover **~80 languages** + multilingual LMs
- **Variations**: model sizes, efficient models, etc.

→ **100s to 1000s** of models in total(!)

- **Evaluation**: perplexity + downstream tasks

UTURKU, UOSLO CUNI; 36 months; 78PM

Overview


WP1: Management
WP6: Continuous Integration and Dashboard
Models for mining
WP2: Data Ingest and Management
WP3: Data Exploration, Cleaning and Privacy
WP4: HP Language Models
WP5: HP Machine Translation
WP7: Dissemination and Exploitation
Web Archive
CommonCrawl
SIGMA2
CESNET
LUMI
Model hosting

Overview

Four tasks:

- T4.1: Building/Training Language Models (UTURKU, UOSLO)
- T4.2: Efficient Data Usage & HPC utilization (UOSLO)
- T4.3: Evaluating Large Language Models (UTURKU, UOSLO)
- T4.4: Ethical Considerations (UOSLO, CUNI)

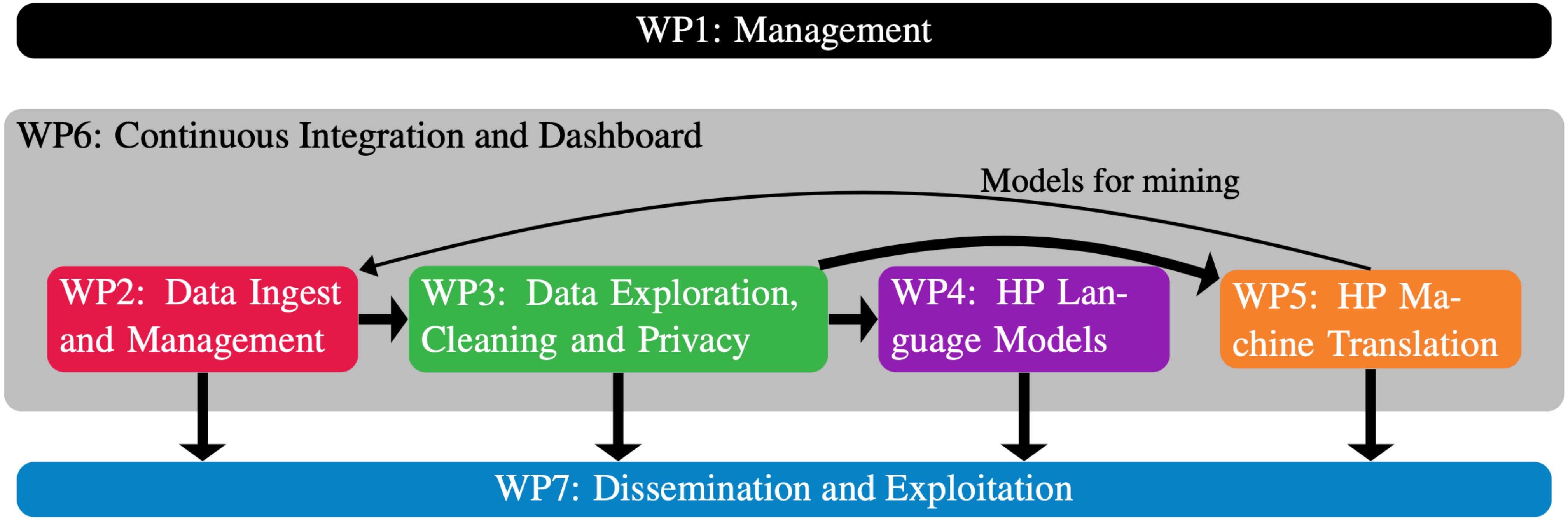
Two deliverables:

- D4.1: First trained language models (UTURKU, M18)
- D4.2: Report on language model evaluation (UTURKU, M35)

Implementation

Key components:

- Monolingual **datasets** (WP2 → WP3 → WP4)
- **Compute** (LUMI-G)
- **Model** and pre-training implementations


WP1: Management
WP6: Continuous Integration and Dashboard
Models for mining
WP2: Data Ingest and Management
WP3: Data Exploration, Cleaning and Privacy
WP4: HP Language Models
WP5: HP Machine Translation
WP7: Dissemination and Exploitation

Compute

HPLT has **3M GPU-hours on LUMI**, the 3rd fastest supercomputer in the world

(Likely possible to get substantial amount of additional GPU time on LUMI)

WP4 and WP5 likely to be the heaviest users of compute

Technology

In UTURKU, currently focusing on **Pytorch + Megatron-DeepSpeed**; interest also in TensorFlow, JAX.

Current status by model class:

- **Causal** (GPT-like): fully functional, scaled to 800 GPUs
- **Bidirectional** (BERT-like): fully functional, not scaled
- **Encoder-Decoder** (T5-like): not running on ROCm, technical challenges remain with large model training

Discussion

- **What languages to focus on first**?
- **Which multilingual model to train**? Balance between limited and massively multilingual?
- **What model sizes to train, and when**? Focus on largest feasible first, or work up from smaller models?
- **Which additional LMs to explore**? Interest in memory/retrieval-augmented models?
- **Which downstream tasks to target in evaluation**?
- **How to split compute budget**? WP4/WP5/others, project participants, GPT/BERT/T5/others?

Discussion

- **Apply for additional compute**? (HPLT members already have several million GPU-h in separate projects!)
- **How generic should pretraining implementations be?** e.g. LUMI only / ROCm+Slurm platforms / supercomputers / any computer?
- **How generic should evaluation implementations be?**
- (Related: How to prioritize training efficiency vs. generality of implementation?)
- **How to coordinate technical work** on WP4/WP5 to minimize duplication of effort?