High-Performance Language Technologies (HPLT) project case: addressing privacy in large language and translation models

Jan Hajič

Institute of Formal and Applied Linguistics
Computer Science School
Faculty of Mathematics and Physics
Charles University, Prague, Czech Republic

The HPLT project

- High-Performance Language Technology
- Horizon Europe DATA call, 2022-2025
- Goals
  - Collect large data from Internet Archive (San Francisco, CA, USA)
    - Approx. 12 PB
  - Extract text, clean, identify, deduplicate, pseudonymize, describe, …
    - Both monolingual and bilingual (parallel) data (texts only)
  - Train language and translation models: 24 EU + min. 16 other
    - xBERTy, GPT-x, Transformer, future SoTA
    - make them openly available (OpusMT, Huggingface, possibly other repos)
  - Evaluate models – keep a dashboard
  - Demonstrate use of EU HPC Centres in a distributed manner
    - Huge compute demands: just for cleaning, 20 mil. CPU hours

HPLT project partners

- Charles University
  - UFAL/LINDAT, Jan Hajic, Dusan Varis, Jindrich Helcl, Martin Popel, Pavel Stranak, Barbora Hladka)
  - coordinator
- University of Edinburgh (Scotland, UK, Barry Haddow / formerly Ken Heafield)
- University of Helsinki (Finland, Jorg Tiedemann, OpusMT)
- University of Turku (Finland, Sampo Pyysalo, Filip Ginter)
- University in Oslo (Norway, Stephan Oepen)
- Prompsit (Spain, Gema Ramirez)
- HPCs:
  - CESNET (Czechia, Ludek Matyska, David Antos)
  - Sigma2 (Norway, Hans Eide)
  - Cooperation with LUMI, EuroHPC, Karolina (IT4Innovations), possibly others

The HPLT project status

- Almost 5 PB downloaded from IA
  - Under specific contract rules
  - Will be complemented from CommonCrawls
  - IA advantage:
    - Whole documents (CC limits size of each document)
- Data published (now at v1.2, Dec. 2023)
  - Plain text, cleaned (of HTML, scripting etc.), deduplicated, language ID assigned, filtered (UT1)
    - 8.4 TB of data (from 1.7 PB), 75 languages, JSONL with metadata
    - Bilingual data only for small languages (for now)
  - https://hplt-project.org/datasets/v1.2
  - Legal conditions: “as is” (~ ParaCrawl distribution)
  - Full description: Deliverable 2.1 (report on v1.0)

HPLT data processing

- Monolingual / bilingual pipelines
  - `warc2text` tool (IA distributed in WARC packages)
  - `Monofixer` tool for fixing character encoding, removing HTML
  - `FastText` for language ID, plus `CLD2`
  - Fluency score by Knesser-Ney character language model
  - Packaging (”sharding”) into equal-sized chunks (a few GBs)
- Bilingual
  - Follows after the monolingual pipeline runs
  - Separated into sentences, translated to English (MarianNMT / OpusMT data), determined similarity / match, cleaned, packaged
  - Using `Bitextor` pipeline (incl. `Bluealign, Bifixer, Bicleaner`)

HPLT additional processing

- Going form v1.0 to v1.2
  - Deduplicated
  - Further cleaning
    - Filtered out documents based on the UT1 blocklist of adult sites
    - Filtered out segments with words per segment < 5, or 200 characters
    - Filtered out mixed language documents (20% of segments or less share the lang ID of document)
    - Parallel data (bitexts) anonymized using the **BiROAMer** tool
      - Mix of Named Entity Recognition and regular expressions

HPLT: data privacy issues (tech)

- Pseudonymization vs. anonymization
  - In text data, similar issue (vs. speech etc.)
    - Could still contain unanonymizable data in some text types
- Standard methodology
  - Named Entity Recognition
  - Removal (by regular expressions)
- Problems faced
  - Non-English languages
    - Scarcity or non-existence of NER training data, person names wrongly marked or mixed up
    - Circumventing: translate, solve NER, align and map back to original language
      - Prone to errors
  - Identifying information outside of names
    - Place of birth, date of birth, IDs, …
    - Some of it solved by regular expressions – problem of multilinguality, errors

HPLT: data privacy issues (legal)

- Right to forget / removal
  - All data marked by a privacy clause that requires us to remove any offending data as reported by eligible users:
  - *Take down: We will comply to legitimate requests by removing the affected sources from the next release of the corpora.*
- Much bigger problem: copyright
  - As with all internet data today wrt Generative models
  - Texts least problematic
  - Specific evaluation of LLMs needed (all aspects)
    - Will be developed later in the project when LLMs are built

Thank you!

https://hplt-project.org
Twitter: @hplt_eu

https://ufal.mff.cuni.cz
https://lindat.cz
https://lindat.cz/services
Twitter: @LindatClariahCZ
Twitter: @ufal_cuni