MLLP* Pilot

By

Ruben Gaspar – IT-CDA-IC

*Machine Learning and Language Processing: https://www.mllp.upv.es/

21.09.2020

Agenda

- Contract
- Collecting data:
  - CDS
  - Digital Memory
  - E-learning
  - LHCP conference
- Work so far

All my notes can be found at:
https://codimd.web.cern.ch/CkA_VyauS_CYqXZrqPzPQg

MLLP Pilot

- Contract and DAI (https://edh.cern.ch/Document/SupplyChain/DAI/8261751) finally sent end of May (27.05.2020)
  - Kind of worrying ahead of time
- NDA signed with MLLP on 10.06.2020

Collecting data

- Gathering information: CDS, Digital Memory, E-learning, LHCP online conference, Noemi from IR/ECO
  - Please follow MLLP guidelines about how to generate transcripts
- Creating the different group sets: training/development, test for the AI models.

Collected information stats: CERN Document Server

First, text data gathered from the CERN Document Server, corresponding to the "Articles & Preprints" and "Books & Proceedings" categories (543K records/xml files), this is, scientific papers, reports, PhD Theses, etc. published in all time:

	*Objects	Sentences.	Tokens	Vocab.
Titles	519K	519K	4.6M	228K
Abstracts	130K	652K	15.6M	393K
Documents (PDFs)	296K	48.9M	1.1G	10.9M
OVERALL	543K	50.0M	1.1G	11.0M

- *Objects: original files processed
- Sentences: sentences
- Tokens: total number of words, numbers, expressions, ...; this is, whatever string surrounded by spaces.
- Vocabulary: number of unique words contained in the data source.
- Units: K -> Kilo (x1000), M -> Mega (x1000000), G -> Giga (x1000000000)

Collected information stats: LHCP 2020

Second, the LHCP 2020 conference task, where videos last on average 26min, and every video features a slides PDF file. It is an speaker-independent dev/test partition (i.e. speakers in dev are not present in test, and viceversa):

	Videos	Length	Speakers
dev	13	5.8h	13 (9 M*, 4 F)
test	14	5.9h	14 (9 M, 5 F)

M*: male, F: femail

Collected information stats: E-learning

First, the e-Learning task, consisting of single-speaker, short formative video tutorials that lasts, on average, 5min:

	Videos	Length	Speakers
test	34	2.8h	9 (6 M, 3 F)

Collected information stats: Digital Memory

Second, regarding text data from the training set of the Digital Memory (all folders but "audioverbatimRef", which is given as a predefined test set; see below the discussion about this particular data source), this is, recordings of internal CERN meetings:

	Objects	Sentences.	Tokens	Vocab.
train	26K	26K	869K	27K

Transcripts are not reliable, due two to reasons: approximate non-verbatim transcripts (avoid wrong utterance-to-phoneme mappings) & result of an automatic speech-to-text alignment, consequences of the former alignments are not accourate.

	Audios	Length	Speakers
train	25.7K	61.6h	N/A
test	598	1.3h	N/A

Further work done

- MLLP Accounts created to John Pym and James Gilles
- Initial testing of the code base in python and nodejs being done for online transcription
  - Basic aim was to test stability of the MLLP online transcript (gRPC calls)
  - It should lead to an app working in similar way as otter.ai.
  - RTMP audio extraction to be researched.
- Working on baseline WER calculations (point 1.2 working plan, see DAI)

Opencast

- Miguel Angel joined (after mandatory quarantine) in September
  - Great news!