European Language Resource Coordination
Connecting Europe Facility

ATHENA Research & Innovation Information Technologies

European Commission

COVID-19 MLIA EVAL
ROUND 2: DATA ACQUISITION

Vassilis Papavassiliou, Stelios Piperidis
{vpapa, spip}@athenarc.gr
Virtual Meeting, 17 February 2022

European Language Resource Coordination
Connecting Europe Facility

European Commission

DATA COLLECTION: 1ST & 2ND ROUND OF THE MT TASK

- From the aspect of data collection,
- how to simulate a very quick response of the MT community in an emergency, like the current pandemic?
  - Generate collections of parallel corpora based on
    - available sources related to health and medicine (“general” subset)
    - content related to COVID-19 of multi/bi-lingual websites (parallel data for domain adaptation)
    - outcomes of available infrastructures (comparable data for domain adaptation)

ATHENA Research & Innovation Information Technologies

DATA COLLECTION: 1ST ROUND ("GENERAL" SUBSET)

- An updated version of the EMEA corpus by
  - harvesting the website of the European Medicines Agency (https://www.ema.europa.eu/en)
  - applying new (more robust and efficient) methods for
    - text extraction from PDF files (enhanced version of the PDFBox library),
    - identification of sentence pairs (by using multilingual embeddings, LASER toolkit)
    - parallel corpus filtering (criteria/findings in WMT parallel corpus filtering shared task).
- Medical/Health-related multilingual collections
  - offered by the Publications Office of EU (https://op.europa.eu/en/home)
  - processed in a similar manner.

European Language Resource Coordination
Connecting Europe Facility

European Commission

DATA COLLECTION: 1ST & 2ND ROUNDS ("COVID-19" SUBSET)

- First step in acquiring COVID-19-related data:
  - Identification of multi/bi-lingual websites with such content.
- With the aim of constructing publicly available data sets,
  - we targeted websites of
    - national authorities and public health agencies (https://www.ecdc.europa.eu/en/COVID-19/national-sources),
    - EU agencies and
    - specific broadcast websites (e.g., Voxeurop, GlobalVoices, etc.)
- In the 2nd round
  - recrawl these websites and
  - crawl websites of several international organizations.

ATHENA Research & Innovation Information Technologies

European Language Resource Coordination
Connecting Europe Facility

European Commission

DATA COLLECTION: 2ND ROUND (NEWS ON COVID-19)

- Exploit the outcomes of an available infrastructure, namely MediSys (Medical Information System).
  - MediSys is one of the publicly accessible systems of the Europe Media Monitor (EMM) which processes media to identify potential public health threats in a fully automated fashion
  - Datasets of metadata which :
    - concern COVID-19 related news
    - contain millions of news articles stored in RSS/XML format
      - https://publications.jrc.ec.europa.eu/repository/handle/JRC120808
      - https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/LANGUAGE-TECHNOLOGY/EMM_collection/2020_MediSys_Covid19_dataset/
      - Jacquet, Guillaume; Verile, Marco (2020): COVID-19 news monitoring with Medical Information System (Medisys). European Commission, Joint Research Centre (JRC) [Dataset] PID: http://data.europa.eu/89h/bd2f71e7-0551-4f57-8e82-fcfca8c1a462

ATHENA Research & Innovation Information Technologies

DATA COLLECTION: 2ND ROUND (NEWS COVID-19)

- Datasets spanning across 10 months (Dec. 2019 – Sep. 2020)
  - grouped according to the publication date of the news articles in batches of a month or two-month period.
- Processing steps:
  1. Parse metadata to extract URL and language of each article.
  2. Fetch news articles of the targeted languages (i.e. AR, DE, EL, EN, ES, FR, IT, SV) as HTML files
  3. Remove the boilerplate (e.g., advertisements, disclaimers, etc.) to keep the main content of each webpage in plain text format
  4. Merge text files to create a single document for each language and each period

DATA COLLECTION: 2ND ROUND (NEWS ARTICLES ON COVID-19)

- These documents :
  - are COVID-19 related monolingual corpora (i.e., in a narrow topic)
  - were published in the same time period
- could be considered comparable (in pairs),

5. Mine sentence alignments for each EN-X language pair.
6. Adopt filtering methods (i.e., thresholding the alignment score by 1.04, removing near deduplicates, etc.).

European Language Resource Coordination
Connecting Europe Facility

European Commission

DATA COLLECTION: DATASETS FOR 2ND ROUND OF THE MT TASK

- Nature of the source data :
  - news articles that might have been (partially) translated many times and republished on several websites
- Aim:
  - Automatic extraction of parallel sentence pairs from comparable corpora
- Resulted corpora mostly include
  - correct sentence pairs
- but also
  - Partially translated pairs
  - “free” translations

ATHENA Research & Innovation Information Technologies