CONTACT PERSONS
Maud Ehrmann, EPFL
Matteo Romanello, UNIL
Antoine Doucet, ULR
Simon Clematide, UZH

CLEF HIPE 2022 Evaluation Lab

Identifying Historical People, Places and other Entities

STAY CONNECTED
https://hipe-eval.github.io/
https://zenodo.org/communities/hipe-eval
TWITTER @ImpressoProject
#clef2022 #HIPE2022

Nº 0003 STAVANGER (NO) / TUESDAY, APRIL 12TH, 2022 FREE / OPEN-SOURCE

Shared Task on Named Entity Recognition and Linking on Multilingual Historical Documents

GOAL & DESCRIPTION

» HIPE-2022 is a CLEF 2022 Evaluation Lab on named entity processing in **historical documents** from the 18th to the 20th century in several languages.

» Compared to the first edition, HIPE-2022 introduces several **novelties**: 1) two document types, classical commentaries and historical newspapers; 2) a broader language spectrum; 3) heterogeneous tag sets and guidelines.

» The **objective** is to assess and advance the development of robust NE processing systems and gain new insights into the transferability of approaches across languages, time periods, document types and annotation tag sets, thereby supporting text understanding of cultural heritage data.

» **Challenges**: HIPE-2022 will confront participants with the challenges of dealing with more languages, learning domain-specific entities, and adapting to diverse annotation schemas.

TASKS & RESEARCH QUESTIONS

» **Tasks:**

» **Named Entity Recognition and Classification (NERC),** with 2 subtasks, NERC-Coarse and NERC-Fine.

» **Named Entity Linking (EL)**: linking of named entity mentions to a unique referent in Wikidata (or to a NIL entity).

» **Research questions**

» How well can general prior knowledge transfer to historical texts?

» Are in-domain language representations (i.e. LMs learned on historical document collections) beneficial, and under which conditions?

» How can systems adapt and integrate training material with different annotations?

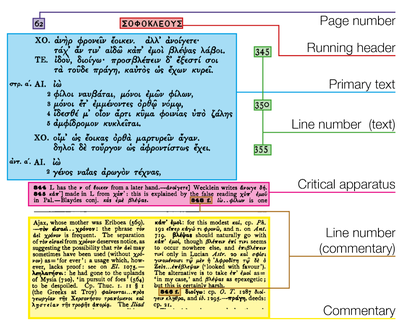
» How can systems, with limited additional in-domain training material, (re)-target models to produce a certain type of annotation?

CORPORA

» The HIPE-2022 datasets are based on six primary datasets that come from several European cultural heritage projects and previous research projects of the HIPE organisers:

» The **historical newspaper data** is composed of several datasets in English, Finnish, French, German and Swedish with: Le Temps data, HIPE-2020 data, NewsEye data, SoNAR data and Living with Machine data.

» The **classical commentaries data** originates from the Ajax Multi-Commentary project and is composed of OCRed 19C commentaries published in French, German and English.



» Textual materials come from different Optical Character Recognition (OCR) softwares and are of varying quality.

» Datasets are prepared and published as **HIPE-2022 releases**, which correspond to a single package composed of neatly structured and homogeneously formatted primary datasets.

EVALUATION

» To accommodate the HIPE-2022 combinatory (tasks, languages, document types, entity tag sets) and foster research on transferability, the evaluation lab is organized around tracks and challenges:

» **Track**: a triple composed of test data [dataset-language-task]

» **Challenge**: a predefined set of tracks (a challenge can be seen as a kind of tournament with multiple tracks).

» HIPE-2022 specifically evaluates **3 challenges**:

» **Multilingual Newspaper Challenge**: newspaper only, min. 2 lang;

» **Multilingual Classical Commentary Challenge**: classical commentary only , min. 3 lang;

» **Global Adaptation Challenge**: both doc types, min. 2 lang.

» Evaluation is based on the open source HIPE scorer which implements (macro and micro) Precision, Recall, and F-measure, with strict and relaxed evaluation regimes.

SIGNIFICANCE OF THE LAB

» **Benefits for the NLP community:** test the robustness of existing approaches and of transfer learning and domain adaptation methods;

» **Benefits for the DH community:** NE processing can support research questions; awareness of performances support informed usage of data.

» **Benefits for cultural heritage professionals**: enhanced access to cultural heritage textual collections.

TIMETABLE

» Early November 2021: registration opens.

» February 2022: release of data.

» Early May 2022: test phase (one week).

» Mid-May 2022: release of results to participants.

» June-July 2022: submission of participants' working notes papers

» September 2022: workshop at the CLEF conference

PARTICIPATION & REGISTRATION

» Teams should register via the CLEF 2022 registration portal.

» It ispossible to participate in one, some or all of the tracks and challenges.

https://hipe-eval.github.io/HIPE-2022/

https://groups.google.com/g/hipe-2022

ORGANIZERS Maud Ehrmann, EPFL - Matteo Romanello, UNIL - Antoine Doucet, ULR - Simon Clematide, UZH
ADVISORY BOARD Sally Chambers, Ghent Centre for Digital Humanities - Frédéric Kaplan, EPFL - Clemens Neudecker, Berlin State Library
CONTRIBUTORS Sven Najem-Meyer, EPFL
PARTNER PROJECTS NEWSEYE, LIVING WITH MACHINE, IMPRESSO, SONAR, HIPE-2020

EPFL · UNIL | Université de Lausanne · La Rochelle Université · University of Zurich[UZH]