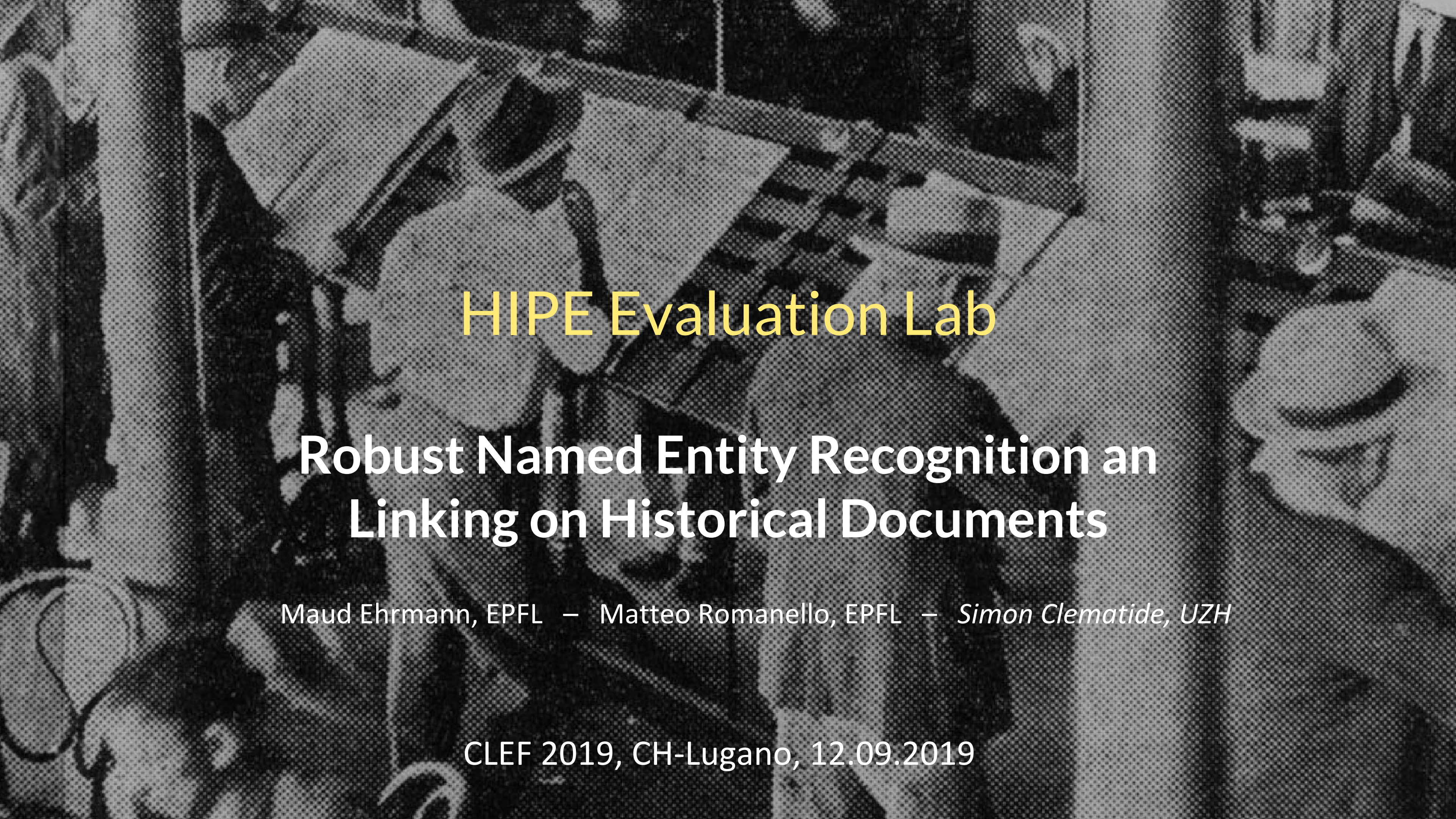


HIPE Evaluation Lab

Robust Named Entity Recognition an Linking on Historical Documents

Maud Ehrmann, EPFL – Matteo Romanello, EPFL – *Simon Clematide, UZH*

CLEF 2019, CH-Lugano, 12.09.2019

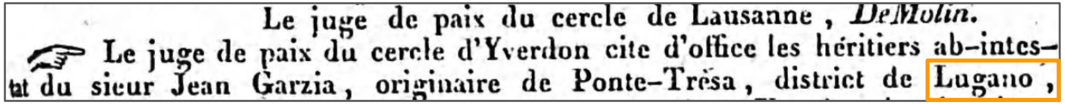
Identifying Historical People, Places and other Entities

What - NE evaluation campaign on historical newspapers in French, German and English.

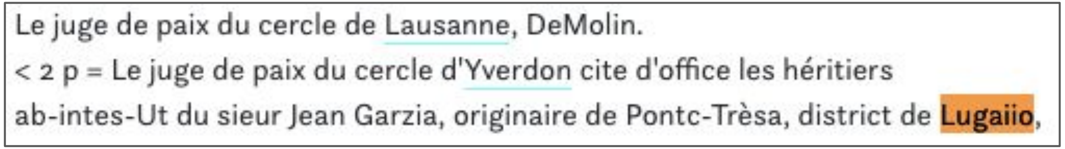
Objectives - Assess and advance the development of **robust named entity processing** systems. Deal with challenging historical material, thereby supporting information extraction and text understanding of cultural heritage data.

Challenges

- Multilingual corpora
- Noisy OCR
- Partial coverage of KBs with respect to historical entities



Le juge de paix du cercle de Lausanne , *DeMolin.*
☞ Le juge de paix du cercle d'Yverdon cite d'office les héritiers ab-intes-
tat du sieur Jean Garzia , originaire de Ponte-Trésa , district de Lugano ,



Le juge de paix du cercle de Lausanne, DeMolin.
< 2 p = Le juge de paix du cercle d'Yverdon cite d'office les héritiers
ab-intes-Ut du sieur Jean Garzia, originaire de Pontc-Trèsa, district de Lugaiio,

Example from Gazette de Lausanne 1829

Tasks

Task 1: Named Entity Recognition and Classification (NERC)

Subtask 1.1 - ‘NERC Essentials’: recognition and classification of high-level entity types.

Subtask 1.2 - ‘NERC fine-grained’: fine-grained entity types + components (e.g. function, title, name).

Task 2 : Named Entity Linking (EL)

Subtask 2.1 - Entity coreference resolution: given a set of mentions in a document, cluster them into coreferent sets.

Subtask 2.2 - Entity Linking: linking of NE mentions to a unique referent in Wikidata or to a NIL node if the mention does not have a referent in the KB.

Data

Corpora - articles sampled among several Swiss, Luxembourgish and British/American (still TBD) historical newspapers on a diachronic basis, released with accompanying metadata.

Datasets

- sample
- dev/train
- test

Auxiliary resources

- HIPE team will provide in-domain word embeddings
- Participants will be encouraged (but not forced) to share any external resource they might use, during and/or after the evaluation campaign.

Timeline

- Early November 2019: registration opens.
- End of November 2019: release of sample data for all tasks and languages.
- End of January 2020: release of full data (training and dev).
- Early May 2020: test phase (one week).
- Early June 2020: release of results to participants.
- June-July 2020: submission of papers (participant and lab- overview papers).
- 22-25 Sept 2020: CLEF 2020 Conference in Thessaloniki, Greece.

Contacts

🌐 HIPE website: https://impresso.github.io/CLEF-HIPE-2020/

🐦 Twitter: @ImpressoProject

✉ Mailing list: https://groups.google.com/forum/#!forum/clef-hipe-2020

🌐 Impresso project: https://impresso-project.ch

Organizers

- Maud Ehrmann, EPFL, maud.ehrmann@epfl.ch
- Matteo Romanello, EPFL, matteo.romanello@epfl.ch
- Simon Clematide, UZH, simon.clematide@uzh.ch

Advisory Board

- R. Eckart de Castilho, TU Darmstadt.
- Clemens Neudecker, Berlin State Library
- Sophie Rosset, LIMSI-CNRS
- David Smith, NorthEastern University