Introducing the HIPE 2022 Shared Task: Named Entity Recognition and Linking in Multilingual Historical Document *

Maud Ehrmann[1], Matteo Romanello[2], Antoine Doucet[3], and Simon Clematide[4]

[1] Digital Humanities Laboratory, EPFL, Vaud, Switzerland
maud.ehrmann@epfl.ch
[2] University of Lausanne, Lausanne, Switzerland
matteo.romanello@unil.ch
[3] University of La Rochelle, La Rochelle, France
antoine.doucet@univ-lr.fr
[4] Department of Computational Linguistics, University of Zurich, Zurich, Switzerland
simon.clematide@uzh.ch

Abstract. We present the HIPE-2022 shared task on named entity processing in multilingual historical documents. Following the success of the first CLEF-HIPE-2020 evaluation lab, this edition confronts systems with the challenges of dealing with more languages, learning domain-specific entities, and adapting to diverse annotation tag sets. HIPE-2022 is part of the ongoing efforts of the natural language processing and digital humanities communities to adapt and develop appropriate technologies to efficiently retrieve and explore information from historical texts. On such material, however, named entity processing techniques face the challenges of domain heterogeneity, input noisiness, dynamics of language, and lack of resources. In this context, the main objective of the evaluation lab is to gain new insights into the *transferability* of named entity processing approaches across languages, time periods, document types, and annotation tag sets.

Keywords: named entity processing, information extraction, text understanding, historical documents, digital humanities

1 Introduction

Through decades of massive digitisation, an unprecedented amount of historical documents became available in digital format, along with their machine-readable texts. While this represents a major step forward in terms of preservation and accessibility, it also bears the potential for new ways to engage with historical documents' contents. The application of machine reading to historical documents is potentially transformative and the next fundamental challenge is to adapt and

* The version of record is available at https://doi.org/10.1007/978-3-030-99739-7_4.

develop appropriate technologies to efficiently search, retrieve and explore information from this 'big data of the past' [9]. Semantic indexing of historical documents is in great demand among humanities scholars, and the interdisciplinary efforts of the digital humanities (DH), natural language processing (NLP), computer vision and cultural heritage communities are progressively pushing forward the processing of facsimiles, as well as the extraction, linking and representation of the complex information enclosed in transcriptions of digitised collections [14]. In this regard, information extraction techniques, and particularly named entity (NE) processing, can be considered among the first and most crucial processing steps.

Yet, the recognition, classification and disambiguation of NEs in historical texts are not straightforward, and performances are not on par with what is usually observed on contemporary well-edited English news material [3]. In particular, NE processing on historical documents faces the challenges of domain heterogeneity, input noisiness, dynamics of language, and lack of resources [6]. Although some of these issues have already been tackled in isolation in other contexts (with e.g., user-generated text), what makes the task particularly difficult is their simultaneous combination and their magnitude: texts are severely noisy, and domains and time periods are far apart.

In this regard, the first CLEF-HIPE-2020 edition[5] [5] proposed the tasks of NE recognition and classification (NER) and entity linking (EL) in ca. 200 years of historical newspapers written in English, French and German and successfully showed that the progress in neural NLP – specifically driven by Transformer-based approaches – also translates into improved performances on historical material, especially for NER. In the meantime, several European cultural heritage projects have prepared additional annotated text material, thereby opening a unique window of opportunity for organising a second edition of the HIPE evaluation lab in 2022.

2 Motivation and Objectives

As the first evaluation campaign of its kind on multilingual historical newspaper material, HIPE-2020 brought together 13 enthusiastic teams who submitted a total of 75 runs for 5 different task bundles. The main conclusion of this edition was that neural-based approaches can achieve good performances on historical NERC when provided with enough training data, but that progress is still needed to further improve performances, adequately handle OCR noise and small-data settings, and better address entity linking. HIPE-2022 will attempt to drive further progress on these points, and also confront systems with new challenges.

HIPE-2022[6] will focus on named entity processing in historical documents covering the period from the 18th to the 20th century and featuring several languages. Compared to the first edition, HIPE-2022 introduces several novelties, with:

[5] https://impresso.github.io/CLEF-HIPE-2020
[6] https://hipe-eval.github.io/HIPE-2022/

- the addition of a new type of document alongside historical newspapers, namely classical commentaries[7];
- the consideration of a broader language spectrum, with 5 languages for historical newspapers (3 for the previous edition), and 3 for classical commentaries;
- the confrontation with the issue of the heterogeneity of annotation tag sets and guidelines.

Overall, HIPE-2022 will confront participants with the challenges of dealing with more languages, learning domain-specific entities, and adapting to diverse annotation schemas. The evaluation lab will therefore contribute to gain new insights on how best to ensure the transferability of NE processing approaches across languages, time periods, document and annotation types, and to answer the question whether one architecture/model can be optimised to perform well across settings and annotation targets in a cultural heritage context. In particular, the following research questions will be addressed:

1. How well can general prior knowledge transfer to historical texts?
2. Are in-domain language representations (i.e. language models learned on the historical document collections) beneficial, and under which conditions?
3. How can systems adapt and integrate training material with different annotations?
4. How can systems, with limited additional in-domain training material, (re)-target models to produce a certain type of annotation?

Recent work on NERC showed encouraging progress on several of these topics: Beryozkin et al. [1] proposed a method to deal with related, but heterogeneous tag sets. Several researchers successfully applied meta-learning strategies to NERC in order to improve transfer learning: Li et al. [10] improved results for extreme low-resource few-shot settings where only a handful of annotated examples for each entity class are used for training; Wu et al. [17] presented techniques to improve cross-lingual transfer; and Li et al. [11] tackled the problem of domain shifts and heterogeneous label sets using meta-learning, proposing a highly data-efficient domain adaptation approach.

2.1 Significance of the Evaluation Lab

HIPE-2022 will benefit the NLP and DH communities, as well as cultural heritage professionals.

Benefits for the NLP community - NLP and information extraction practitioners will have the possibility to test the robustness of existing approaches and to experiment with transfer learning and domain adaptation methods, whose

[7] Classical commentaries are scholarly publications dedicated to the in-depth analysis and explanation of ancient literary works. As such, they aim to facilitate the reading and understanding of a given literary text. More information on the HIPE-2022 classical commentaries corpus in Section 3.2.

performances could be systematically evaluated and compared on broad historical and multilingual data sets. Beside gaining new insights with respect to domain and language adaptation and advancing the state of the art in semantic indexing of historical material, the lab will also contribute a set of multilingual NE-annotated datasets that could be used for further training and benchmarking.

Benefits for the DH community - DH researchers are in need of support to explore the large quantities of text they currently have at hand, and NE processing is high on their wish list. Such processing can support research questions in various domains (e.g. history, political science, literature and historical linguistics) and knowing about performances is a must in order to do an informed usage of the enriched data. This lab's outcome (datasets and systems) will be beneficial to DH practitioners insofar as it will help identify state-of-the-art solutions for NE processing of historical texts.

Benefits for cultural heritage professionals - Libraries, archives and museums (LAM) increasingly focus on advancing the usage of artificial intelligence methods on cultural heritage text collections, in particular NE processing [13, 7]. This community is eager to collaborate and provide data (when copyright allows) for high-quality semantic enrichment.

3 Overview of the Evaluation Lab

3.1 Task Description

HIPE-2022 focuses on the same tasks as CLEF-HIPE-2020, namely:

Task 1: Named Entity Recognition and Classification (NERC)

Subtask 1.1 - NERC-Coarse: this task includes the recognition and classification of high-level entity types (Person, Organisation, Location, Product and domain-specific entities, e.g. mythological characters or literary works in classical commentaries).

Subtask 1.2 - NERC-Fine: includes 'NERC-Coarse', plus the detection and classification at sub-type level and the detection of NE components (e.g. function, title, name). This subtask will be proposed for English, French and German only.

Task 2: Named Entity Linking (EL) This task corresponds to the linking of named entity mentions to a unique item ID in Wikidata, our knowledge base of choice, or to a NIL node if the mention does not have a corresponding item in the KB. We will allow submissions of both end-to-end systems (NERC and EL) and of systems performing exclusively EL on gold entity mentions provided by the organizers (EL-only).

3.2 Data Sets

Corpora The lab's corpora will be composed of historical newspapers and classic commentaries covering ca. 200 years. We benefit from published and to-date unpublished NE-annotated data from organisers' previous research project,

from the previous HIPE-2020 campaign, as well as from several ongoing research projects which agreed to postpone the publication of 10% to 20% of their annotated material in order to support HIPE-2022.

Historical newspapers. The historical newspaper data is composed of several datasets in English, Finnish, French, German and Swedish which originate from various projects and national libraries in Europe:

- *Le Temps* data: an unpublished, annotated diachronic dataset composed of historical newspaper articles from two Swiss newspapers in French (19C-20C) [3]. This dataset contains 10,580 entity mentions and will be part of the training, dev and test sets.
- *HIPE-2020* data: the datasets used during the first HIPE-2020 campaign, composed of newspaper articles from Swiss, Luxembourgish and American newspapers in French, German and English (19C-20C). These datasets contain 19,848 linked entities and will be part of the training sets.
- *HIPE-2020* unpublished data: a set of unpublished diachronic annotated data composed of newspaper articles from Swiss newspapers in French and German (19C-20C). These data will be part of the test sets.
- *NewsEye* data[8]: a partially published annotated dataset composed of newspaper articles from newspapers in French, German, Finnish and Swedish (19C-20C) [8]. The already published part contains 30,580 entities and will be part of the training and dev sets. The unpublished one (roughly 20% of the total) will be part of the test set.
- *SoNAR* data: an annotated dataset composed of newspaper articles from the Berlin State library newspaper collections in German (19C-20C), produced in the context of the SoNAR project[9]. The (soon to be) published part of this dataset will be part of the training and dev sets, while the unpublished lot will integrate the HIPE-2022 test set.
- *Living With Machines* data[10]: an annotated dataset composed of newspaper articles from the British Library newspapers in English (18C-19C), and annotated exclusively with geographical locations following ad-hoc annotation guidelines. The already published portion of the data [2] contains 3,355 annotated toponyms and will be included in the training and dev sets. The unpublished portion will be part of the test set.

Historical commentaries. The classical commentaries data originates from the *Ajax Multi-Commentary* project and is composed of OCRed 19C commentaries published in French, German and English [15], annotated with both universal NEs (person, location, organisation) and domain-specific NEs (bibliographic references to primary and secondary literature). In the field of classical studies, commentaries constitute one of the most important and enduring forms of scholarship, together with critical editions and translations. They are information-rich texts, characterised by a high density of NEs.

[8] https://www.newseye.eu/
[9] https://sonar.fh-potsdam.de/
[10] https://livingwithmachines.ac.uk/

Annotation In terms of annotation, the common requirement for most of these datasets is to have person, location and organisation entity types, and entity links towards Wikidata. The guidelines used to annotate many datasets derive from related directives (Quaero and *impresso*-HIPE-2020 guidelines)[11], yet they do differ in some respects such as granularity of annotations (coarse vs fine-grained), treatment of metonymy and inclusion of entity components.

3.3 Evaluation

To accommodate the different dimensions that characterise our datasets (languages, document types, domains, entity tag sets) and foster research on transferability, the evaluation lab will be organised around two main 'challenges', namely a *multilingual challenge* and an *adaptation challenge*, each featuring several task 'tracks'. They will ensure that participants will have to work across settings, e.g. with documents in at least two different languages or annotated according to two different tag sets or guidelines, while keeping a clear and defined evaluation frame.

Evaluation will be performed with the open source HIPE scorer[12], which was developed for the first edition of the shared task. Evaluation metrics implemented in the scorer include (macro and micro) Precision, Recall, and F-measure and evaluation settings will include strict (exact matching) and relaxed (fuzzy matching) evaluation scenarios.

4 Conclusion

Following a first and successful shared task on NE processing on historical newspapers, the HIPE-2022 evaluation lab proposes to confront systems with the new challenges of dealing with more languages, learning domain-specific entities, and adapting to diverse annotation tag sets. The overall objective is to assess and advance the development of robust, adaptable and transferable named entity processing systems in order to support information extraction and text understanding of cultural heritage data.

Acknowledgements

We are grateful to the research project consortia and teams who kindly accepted to retain the publication of part of their NE-annotated datasets to support HIPE-2022: the NewsEye project[13]; the Living with Machine project, in particular Mariona Coll'Ardanuy; and the SoNAR project, in particular Clemens Neudecker. We also thank Sally Chambers, Clemens Neudecker and Frédéric Kaplan for their support and guidance as part of the lab's advisory board.

[11] *Impresso* [4] and SoNAR guidelines [12] were derived from Quaero guidelines [16], while NewsEye guidelines correspond to a subset of the *impresso* guidelines.

[12] https://github.com/impresso/CLEF-HIPE-2020-scorer

[13] The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770299

Bibliography

[1] Genady Beryozkin, Yoel Drori, Oren Gilon, Tzvika Hartman, and Idan Szpektor. A joint named-entity recognizer for heterogeneous tag-sets using a tag hierarchy. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 140–150, Florence, Italy, July 2019. URL https://aclanthology.org/P19-1014.

[2] Mariona Coll Ardanuy, David Beavan, Kaspar Beelen, Kassra Hosseini, and Jon Lawrence. Dataset for Toponym Resolution in Nineteenth-Century English Newspapers, 2021. URL https://https//doi.org/10.23636/b1c4-py78.

[3] Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. Diachronic evaluation of NER systems on old newspapers. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 97–107, Bochum, 2016. Bochumer Linguistische Arbeitsberichte. URL https://infoscience.epfl.ch/record/221391.

[4] Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. Impresso Named Entity Annotation Guidelines. Annotation guidelines, Ecole Polytechnique Fédérale de Lausanne (EPFL) and Zurich University (UZH), January 2020. URL https://zenodo.org/record/3585750.

[5] Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. In Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol, editors, *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, volume 2696, page 38, Thessaloniki, Greece, 2020. CEUR-WS. https://doi.org/10.5281/zenodo.4117566. URL https://infoscience.epfl.ch/record/281054.

[6] Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. Named Entity Recognition and Classification on Historical Documents: A Survey. *arXiv:2109.11406 [cs]*, September 2021. URL http://arxiv.org/abs/2109.11406. arXiv: 2109.11406.

[7] Markus Gregory, Clemens Neudecker, Antoine Isaac, Giles Bergel, and Others. AI in relation to GLAMs task FOrce - Report and Recommendations. Technical report, Europeana Network ASsociation, 2021. URL https://pro.europeana.eu/project/ai-in-relation-to-glams.

[8] Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi Tuyet Hai Nguyen, Günter Hackl, Jose G. Moreno, and Antoine Doucet. A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 2328–2334, New York, NY, USA, July 2021. Association for Computing Machinery. ISBN 978-1-4503-8037-9. https://doi.org/10.1145/3404835.3463255.

[9] Frédéric Kaplan and Isabella di Lenardo. Big Data of the Past. *Frontiers in Digital Humanities*, 4:1–21, 2017. ISSN 2297-2668.

https://doi.org/10.3389/fdigh.2017.00012. URL https://www.frontiersin.org/articles/10.3389/fdigh.2017.00012/full. Publisher: Frontiers.

[10] Jing Li, Billy Chiu, Shanshan Feng, and Hao Wang. Few-shot named entity recognition via meta-learning. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2020.

[11] Jing Li, Shuo Shang, and Ling Shao. Metaner: Named entity recognition with meta-learning. In *Proceedings of The Web Conference 2020*, WWW '20, page 429–440, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. https://doi.org/10.1145/3366423.3380127. URL https://doi.org/10.1145/3366423.3380127.

[12] Sina Menzel, Josefine Zinck, Hannes Schnaitter, and Vivien Petras. Guidelines for Full Text Annotations in the SoNAR (IDH) Corpus. Technical report, Zenodo, July 2021. URL https://zenodo.org/record/5115933.

[13] Thomas Padilla. Responsible Operations: Data Science, Machine Learning, and AI in Libraries. Technical report, OCLC Research, USA, May 2020. URL https://www.oclc.org/content/research/publications/2019/oclcresearch-responsible-operations-data-science-machine-learning-ai.html.

[14] Mia Ridge, Giovanni Colavizza, Laurel Brake, Maud Ehrmann, Jean-Phillipe Moreux, and Andrew Prescott. The past, present and future of digital scholarship with newspaper collections. In *DH 2019 book of abstracts*, pages 1–9, Utrecht, The Netherlands, 2019. URL http://infoscience.epfl.ch/record/271329.

[15] Matteo Romanello, Najem-Meyer Sven, and Bruce Robertson. Optical Character Recognition of 19th Century Classical Commentaries: the Current State of Affairs. In *The 6th International Workshop on Historical Document Imaging and Processing (HIP '21)*, Lausanne, September 2021. Association for Computing Machinery. https://doi.org/10.1145/3476887.3476911. URL https://doi.org/10.1145/3476887.3476911.

[16] Sophie Rosset, Grouin, Cyril, and Zweigenbaum, Pierre. Entités nommées structurées : Guide d'annotation Quaero. Technical Report 2011-04, LIMSI-CNRS, Orsay, France, 2011.

[17] Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F. Karlsson, Biqing Huang, and Chin-Yew Lin. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. *CoRR*, abs/1911.06161, 2019. URL http://arxiv.org/abs/1911.06161.