Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents

Maud Ehrmann[1] [0000−0001−9900−2193], Matteo Romanello[2] [0000−0002−7406−6286], Sven Najem-Meyer[1] [0000−0002−3661−4579], Antoine Doucet[3] [0000−0001−6160−3356], and Simon Clematide[4] [0000−0003−1365−0662]

[1] Digital Humanities Laboratory, EPFL, Vaud, Switzerland
maud.ehrmann@epfl.ch, sven.najem-meyer@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. This paper presents an overview of the second edition of HIPE (Identifying Historical People, Places and other Entities), a shared task on named entity recognition and linking in multilingual historical documents. Following the success of the first CLEF-HIPE-2020 evaluation lab, HIPE-2022 confronts systems with the challenges of dealing with more languages, learning domain-specific entities, and adapting to diverse annotation tag sets. This shared task 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 HIPE-2022, run as an evaluation lab of the CLEF 2022 conference, is to gain new insights into the *transferability* of named entity processing approaches across languages, time periods, document types, and annotation tag sets. Tasks, corpora, and results of participating teams are presented.

Keywords: Named entity recognition and classification · Entity linking · Historical texts · Information extraction · Digitised newspapers · 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Through decades of massive digitisation, an unprecedented amount of historical documents became available in digital format, along with their machine-readable texts.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major step forward in terms of preservation and accessibility, it also bears the potential for new ways to engage with historical documents' contents. The application of machine reading to historical documents is potentially transformative and the next fundamental challenge is to adapt and develop appropriate technologies to efficiently search, retrieve and explore information from this 'big data of the past' [21]. 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 [28]. 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 is not straightforward, and performances are not on par with what is usually observed on contemporary well-edited English news material [8]. In particular, NE processing on historical documents faces the challenges of domain heterogeneity, input noisiness, dynamics of language, and lack of resources [9]. 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.

Motivation and Objectives. As the first evaluation campaign of its kind on multilingual historical newspaper material, the CLEF-HIPE-2020 edition[5] [13, 14] proposed the tasks of NE recognition and classification (NERC) and entity linking (EL) in ca. 200 years of historical newspapers written in English, French and German. HIPE-2020 brought together 13 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 attempts to drive further progress on these points, and also confront systems with new challenges. An additional point is that in the meantime several European cultural heritage projects have prepared additional NE-annotated text material, thus opening a unique window of opportunity to organize a second edition of the HIPE evaluation lab in 2022.

HIPE-2022[6] shared task focuses 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:

[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 heterogeneous annotation tag sets and guidelines.

Overall, HIPE-2022 confronts participants with the challenges of dealing with more languages, learning domain-specific entities, and adapting to diverse annotation schemas. The objectives of the evaluation lab are to contribute 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 or model can be optimised to perform well across settings and annotation targets in a cultural heritage context. In particular, the following research questions are 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. [3] proposed a method to deal with related, but heterogeneous tag sets. Several researchers successfully applied meta-learning strategies to NERC to improve transfer learning: Li et al. [23] 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. [36] presented techniques to improve cross-lingual transfer; and Li et al. [24] tackled the problem of domain shifts and heterogeneous label sets using meta-learning, proposing a highly data-efficient domain adaptation approach.

The remainder of this paper is organized as follows. Sections 2 and 3 present the tasks and the material used for the evaluation. Section 4 details the evaluation framework, with evaluation metrics and the organisation of system submissions around tracks and challenges. Section 5 introduces the participating systems, while Section 6 presents and discusses their results. Finally, Section 7 summarizes the benefits of the task and concludes.[8]

[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.

[8] For space reasons, the discussion of related work is included in the extended version of this overview [15].

Table 1: Overview of HIPE-2022 datasets with an indication of which tasks they are suitable for according to their annotation types.

Dataset alias	Document type	Languages	Suitable for
hipe2020	historical newspapers	de, fr, en	NERC-Coarse, NERC-Fine, EL
newseye	historical newspapers	de, fi, fr, sv	NERC-Coarse, NERC-Fine, EL
sonar	historical newspapers	de	NERC-Coarse, EL
letemps	historical newspapers	fr	NERC-Coarse, NERC-Fine
topres19th	historical newspapers	en	NERC-Coarse, EL
ajmc	classical commentaries	de, fr, en	NERC-Coarse, NERC-Fine, EL

2 Task Description

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

Task 1: Named Entity Recognition and Classification (NERC)

- **Subtask 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 NERC-Fine:** includes the recognition and classification of entity mentions according to fine-grained types, plus the detection and classification of nested entities of depth 1. This subtask is 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 value if the mention does not have a corresponding item in the knowledge base (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 Data

HIPE-2022 data consists of six NE-annotated datasets composed of historical newspapers and classic commentaries covering ca. 200 years. Datasets originate from the previous HIPE-2020 campaign, from HIPE organisers' previous research project, and from several European cultural heritage projects which agreed to postpone the publication of 10% to 20% of their annotated material to support HIPE-2022. Original datasets feature several languages and were annotated with different entity tag sets and according to different annotation guidelines. See Table 1 for an overview.

3.1 Original Datasets

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:

- **HIPE-2020 data** corresponds to the datasets of the first HIPE-2020 campaign. They are composed of articles from Swiss, Luxembourgish and American newspapers in French, German and English (19C-20C) that were assembled during the *impresso* project[9] [10]. Together, the train, dev and test `hipe2020` datasets contain 17,553 linked entity mentions, classified according to a fine-grained tag set, where nested entities, mention components and metonymic senses are also annotated [12].
- **NewsEye data** corresponds to a set of NE-annotated datasets composed of newspaper articles in French, German, Finnish and Swedish (19C-20C) [18]. Built in the context of the *NewsEye* project[10], the `newseye` train, dev and test sets contain 36,790 linked entity mentions, classified according to a coarse-grained tag set and annotated on the basis of guidelines similar to the ones used for `hipe2020`. Roughly 20% of the data was retained from the original dataset publication and is published for the first time for HIPE-2022, where it is used as test data (thus the previously published test set became a second dev set in HIPE-2022 data distribution).
- **SoNAR data** is an NE-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[11]. The `sonar` dataset contains 1,125 linked entity mentions, classified according to a coarse-grained tag set. It was thoroughly revised and corrected on NE and EL levels by the HIPE-2022 organisers. It is split in a dev and test set – without providing a dedicated train set.
- **Le Temps data**: a previously unpublished, NE-annotated diachronic dataset composed of historical newspaper articles from two Swiss newspapers in French (19C-20C) [8]. This dataset contains 11,045 entity mentions classified according to a fine-grained tag set similar to `hipe2020`.
- **Living with Machines data** corresponds to an NE-annotated dataset composed of newspaper articles from the British Library newspapers in English (18C-19C) and assembled in the context of the *Living with Machine* project[12]. The `topres19th` dataset contains 4,601 linked entity mentions, exclusively of geographical types annotated following their own annotation guidelines [5]. Part of this data has been retained from the original dataset publication and is used and released for the first time for HIPE-2022.

[9] https://impresso-project.ch
[10] https://www.newseye.eu/
[11] https://sonar.fh-potsdam.de/
[12] https://livingwithmachines.ac.uk/

Table 2: Entity types used for NERC tasks, per dataset and with information whether nesting and linking apply. ∗: these types are not present in `letemps` data. ∗∗: linking applies, unless the token is flagged as *InSecondaryReference*.

Dataset	**Coarse tag set**	**Fine tag set**	**Nesting**	**Linking**
`hipe2020`	pers	pers.ind	yes	yes
`letemps`		pers.coll		
		pers.ind.articleauthor		
	org∗	org.adm	yes	yes
		org.ent		
		org.ent.pressagency		
	prod∗	prod.media	no	yes
		prod.doctr		
	time∗	time.date.abs	no	no
	loc	loc.adm.town	yes	yes
		loc.adm.reg		
		loc.adm.nat		
		loc.adm.sup		
		loc.phys.geo	yes	yes
		loc.phys.hydro		
		loc.phys.astro		
		loc.oro	yes	yes
		loc.fac	yes	yes
		loc.add.phys	yes	yes
		loc.add.elec		
		loc.unk	no	no
`newseye`	pers	pers.articleauthor	yes	yes
	org	-	yes	yes
	humanprod	-	yes	yes
	loc	-	no	yes
`topres19th`	loc	-	no	yes
	building	-	no	yes
	street	-	no	yes
`ajmc`	pers	pers.author	yes	yes∗∗
		pers.editor		
		pers.myth		
		pers.other		
	work	work.primlit	yes	yes∗∗
		work.seclit		
		work.fragm		
	loc	-	yes	yes∗∗
	object	object.manuscr	yes	no
		object.museum		
	date	-	yes	no
	scope	-	yes	no
`sonar`	pers	-	no	yes
	loc	-	no	yes
	org	-	no	yes

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 [30], 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.

These six datasets compose the HIPE-2022 corpus. They underwent several preparation steps, with conversion to the tab-separated HIPE format, correction of data inconsistencies, metadata consolidation, re-annotation of parts of the datasets, deletion of extremely rare entities (esp. for `topres19th`), and rearrangement or composition of train and dev splits[13].

3.2 Corpora Characteristics

Overall, the HIPE-2022 corpus covers five languages (English, French, Finnish, German and Swedish), with a total of over 2.3 million tokens (2,211,449 for newspapers and 111,218 for classical commentaries) and 78,000 entities classified according to five different entity typologies and linked to Wikidata records. Detailed statistics about the datasets are provided in Table 3 and 4.

The datasets in the corpus are quite heterogeneous in terms of annotation guidelines. Two datasets – `hipe2020` and `letemps` – follow the same guidelines [12, 31], and `newseye` was annotated using a slightly modified version of these guidelines. In the `sonar` dataset, persons, locations and organisations were annotated, whereas in `topres19th` only toponyms were considered. Compared to the other datasets, `ajmc` stands out for having being annotated according to domain-specific guidelines [29], which focus on bibliographic references to primary and secondary literature. This heterogeneity of guidelines leads to a wide variety of entity types and sub-types for the NERC task (see Table 2). Among these types, only persons, locations and organisations are found in all datasets (except for `topres19th`). While nested entities are annotated in all datasets except `topres19th` and `sonar`, only `hipe2020` and `newseye` have a sizable number of such entities.

Detailed information about entity mentions that are affected by OCR mistakes is provided in `ajmc` and `hipe2020` (only for the test set for the latter). As OCR noise constitutes one of the main challenges of historical NE processing [9], this information can be extremely useful to explain differences in performance between datasets or between languages in the same dataset. For instance, looking at the percentage of noisy mentions for the different languages in `ajmc`, we find that it is three times higher in French documents than in the other two languages.

[13] Additional information is available online by following the links indicated for each datasets in Table 1.

Table 3: Overview of newspaper corpora statistics (hipe-2022 release v2.1). NIL percentages are computed based on linkable entities (i.e., excluding time entities for `hipe2020`).

Dataset	Lang.	Fold	Docs	Tokens	Mentions				
					All	Fine	Nested	%noisy	%NIL
`hipe2020`	**de**	Train	103	86,446	3,494	3,494	158	-	15.70
		Dev	33	32,672	1,242	1,242	67	-	18.76
		Test	49	30,738	1,147	1,147	73	12.55	17.40
	Total		185	149,856	5,883	5,883	298	-	16.66
	en	Train	-	-	-	-	-	-	-
		Dev	80	29,060	966	-	-	-	44.18
		Test	46	16,635	449	-	-	5.57	40.28
	Total		126	45,695	1,415	-	-	-	42.95
	fr	Train	158	166,218	6,926	6,926	473	-	25.26
		Dev	43	37,953	1,729	1,729	91	-	19.81
		Test	43	40,855	1,600	1,600	82	11.25	20.23
	Total		244	245,026	10,255	10,255	646	-	23.55
Total			555	440,577	17,553	16,138	944	-	22.82
`newseye`	**de**	Train	7	374,250	11,381	21	876	-	51.07
		Dev	12	40,046	539	5	27	-	22.08
		Dev2	12	39,450	882	4	64	-	53.74
		Test	13	99,711	2,401	13	89	-	48.52
	Total		44	553,457	15,203	43	1,056	-	49.79
	fi	Train	24	48,223	2,146	15	224	-	40.31
		Dev	24	6,351	223	1	25	-	40.36
		Dev2	21	4,705	203	4	22	-	42.86
		Test	24	14,964	691	7	42	-	47.47
	Total		93	74,243	3,263	27	313	-	41.99
	fr	Train	35	255,138	10,423	99	482	-	42.42
		Dev	35	21,726	752	3	29	-	30.45
		Dev2	35	30,457	1,298	10	63	-	38.91
		Test	35	70,790	2,530	34	131	-	44.82
	Total		140	378,111	15,003	146	705	-	41.92
	sv	Train	21	56,307	2,140	16	110	-	32.38
		Dev	21	6,907	266	1	7	-	25.19
		Dev2	21	6,987	311	1	20	-	37.30
		Test	21	16,163	604	0	26	-	35.43
	Total		84	86,364	3,321	18	163	-	32.82
Total			361	1,092,175	36,790	234	2,237	-	44.36
`letemps`	**fr**	Train	414	379,481	9,159	9,159	69	-	-
		Dev	51	38,650	869	869	12	-	-
		Test	51	48,469	1,017	1,017	12	-	-
Total			516	466,600	11,045	11,045	93	-	-
`topres19th`	**en**	Train	309	123,977	3,179	-	-	-	18.34
		Dev	34	11,916	236	-	-	-	13.98
		Test	112	43,263	1,186	-	-	-	17.2
	Total		455	179,156	4,601	-	-	-	17.82
Total			455	179,156	4,601	-	-	-	17.82
`sonar`	**de**	Train	-	-	-	-	-	-	-
		Dev	10	17,477	654	-	-	-	22.48
		Test	10	15,464	471	-	-	-	33.33
	Total		20	32,941	1,125	-	-	-	27.02
Total			20	32,941	1,125	-	-	-	27.02
Grand Total (newspapers)			**1,907**	**2,211,449**	**71,114**	**27,417**	**3,274**		**30.23**

Table 4: Corpus statistics for the `ajmc` dataset (HIPE-2022 release v2.1).

Dataset	Lang.	Fold	Docs	Tokens	Mentions				
					All	Fine	Nested	%noisy	%NIL
ajmc	**de**	Train	76	22,694	1,738	1,738	11	13.81	0.92
		Dev	14	4,703	403	403	2	11.41	0.74
		Test	16	4,846	382	382	0	10.99	1.83
	Total		106	32,243	2,523	2,523	13	13.00	1.03
	en	Train	60	30,929	1,823	1,823	4	10.97	1.66
		Dev	14	6,507	416	416	0	16.83	1.70
		Test	13	6,052	348	348	0	10.34	2.61
	Total		87	43,488	2,587	2,587	4	11.83	1.79
	fr	Train	72	24,670	1,621	1,621	9	30.72	0.99
		Dev	17	5,426	391	391	0	36.32	2.56
		Test	15	5,391	360	360	0	27.50	2.80
	Total		104	35,487	2,372	2,372	9	31.16	1.52
Grand Total (ajmc)			**297**	**111,218**	**7,482**	**7,482**	**26**		**1.45**

HIPE-2022 datasets show significant differences in terms of lexical overlap between train, dev and test sets. Following the observations of Augenstein et al. [2] and Taillé et al. [32] on the impact of lexical overlap on NERC performance, we computed the percentage of mention overlap between data folds for each dataset, based on the number of identical entity mentions (in terms of surface form) between train+dev and test sets (see Table 5). Evaluation results obtained on training and test sets with low mention overlap, for example, can be taken as an indicator of the ability of the models to generalise well to unseen mentions. We find that `ajmc`, `letemps` and `topres19th` have a mention overlap which is almost twice that of `hipe2020`, `sonar` and `newseye`.

Finally, regarding entity linking, it is interesting to observe that the percentage of NIL entities (i.e. entities not linked to Wikidata) varies substantially across datasets. The Wikidata coverage is drastically lower for `newseye` than for the other newspaper datasets (44.36%). Conversely, only 1.45% of the entities found in `ajmc` cannot be linked to Wikidata. This fact is not at all surprising considering that commentaries mention mostly mythological figures, scholars of the past and literary works, while newspapers mention many relatively obscure or unknown individuals, for whom no Wikidata entry exists.

3.3 HIPE-2022 Releases

HIPE-2022 data is released as a single package consisting of the neatly structured and homogeneously formatted original datasets. The data is released in IOB format with hierarchical information, similarly to CoNLL-U[14], and consists

[14] https://universaldependencies.org/format.html

Table 5: Overlap of mentions between test and train (plus dev) sets as percentage of the total number of mentions.

Dataset	Lang.	% overlap	Folds
ajmc	de	31.43	train+dev vs test
	en	30.50	train+dev vs test
	fr	27.53	train+dev vs test
	Total	29.87	
hipe2020	de	16.22	train+dev vs test
	en	6.22	dev vs test
	fr	19.14	train+dev vs test
	Total	17.12	
letemps	fr	25.70	train+dev vs test
sonar	de	10.13	dev vs test
newseye	fr	14.79	train+dev vs test
	de	20.77	train+dev vs test
	fi	6.63	train+dev vs test
	sv	10.36	train+dev vs test
	Total	16.18	
topres19th	en	32.33	train+dev vs test

of UTF-8 encoded, tab-separated values (TSV) files containing the necessary information for all tasks (NERC-Coarse, NERC-Fine, and EL). There is one TSV file per dataset, language and split. Original datasets provide different document metadata with different granularity. This information is kept in the files in the form of metadata blocks that encode as much information as necessary to ensure that each document is self-contained with respect to HIPE-2022 settings. Metadata blocks use namespacing to distinguish between mandatory shared task metadata and dataset-specific metadata.

HIPE-2022 data releases are published on the HIPE-eval GitHub organisation repository[15] and on Zenodo[16]. Various licences (of type CC-BY and CC-BY-NC-SA) apply to the original datasets – we refer the reader to the online documentation.

4 Evaluation Framework

4.1 Task Bundles, Tracks and Challenges

To accommodate the different dimensions that characterise the HIPE-2022 shared task (languages, document types, entity tag sets, tasks) and to foster research on

15 https://github.com/impresso/CLEF-HIPE-2020/tree/master/data
16 https://doi.org/10.5281/zenodo.6579950

transferability, the evaluation lab is organised around **tracks** and **challenges**. Challenges guide participation towards the development of approaches that work across settings, e.g. with documents in at least two different languages or annotated according to two different tag sets or guidelines, and provide a well-defined and multi-perspective evaluation frame.

To manage the total combinations of datasets, languages, document types and tasks, we defined the following elements (see also Figure 1):

- **Task bundle**: a task bundle is a predefined set of tasks as in HIPE-2020 (see bundle table in Fig. 1). Task bundles offer participating teams great flexibility in choosing which tasks to compete for, while maintaining a manageable evaluation frame. Concretely, teams were allowed to submit several 'submission bundles', i.e. a triple composed of dataset/language/taskbundle, with up to 2 runs each.
- **Track**: a track corresponds to a triple composed of dataset/language/task and forms the basic unit for which results are reported.
- **Challenge**: a challenge corresponds to a predefined set of tracks. A challenge can be seen as a kind of tournament composed of tracks.

HIPE-2022 specifically evaluates 3 challenges:

1. **Multilingual Newspaper Challenge** (MNC): This challenge aims at fostering the development of multilingual NE processing approaches on historical newspapers. The requirements for participation in this challenge are that submission bundles consist only of newspaper datasets and include at least two languages for the same task (so teams had to submit a minimum of two submission bundles for this challenge).
2. **Multilingual Classical Commentary Challenge** (MCC): This challenge aims at adapting NE solutions to domain-specific entities in a specific digital humanities text type of classic commentaries. The requirements are that submission bundles consist only of the `ajmc` dataset and include at least three languages for the same task.
3. **Global Adaptation Challenge** (GAC): Finally, the global adaptation challenge aims at assessing how efficiently systems can be retargeted to any language, document type and guidelines. Bundles submitted for this challenge could be the same as those submitted for MNC and MCC challenges. The requirements are that they consist of datasets of both types (commentaries and newspaper) and include at least two languages for the same task.

4.2 Evaluation Measures

As in HIPE-2020, NERC and EL tasks are evaluated in terms of Precision, Recall and F-measure (F1-score) [25]. Evaluation is carried out at entity level according to two computation schemes: micro average, based on true positives, false

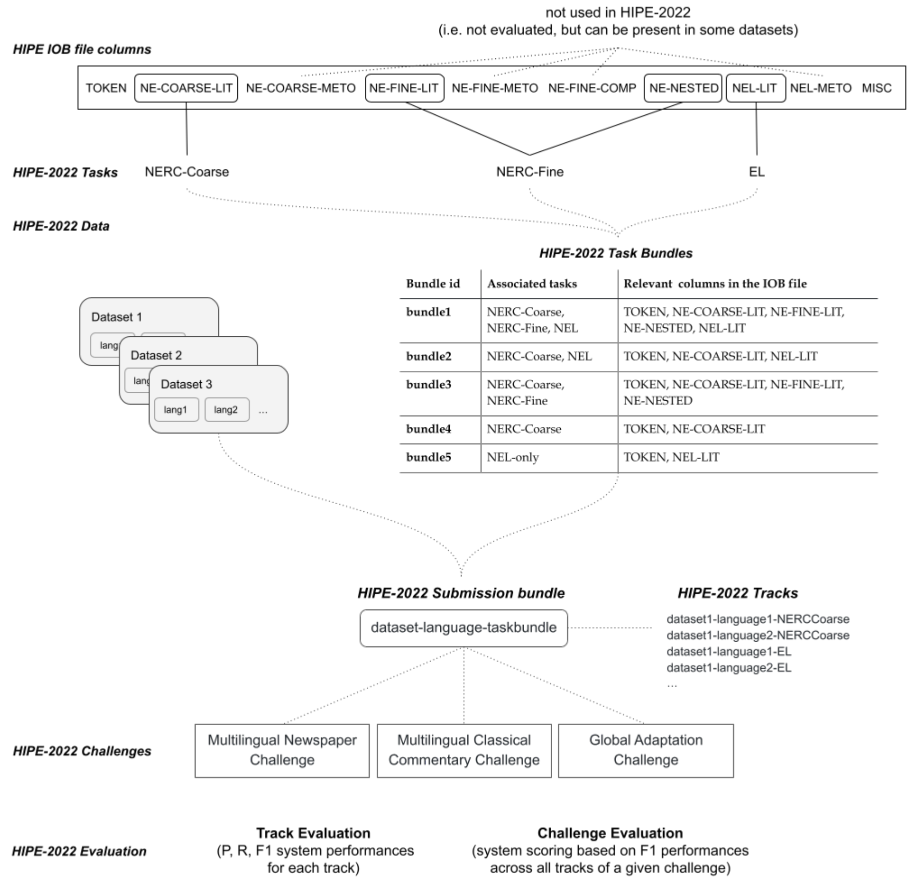

Fig. 1: Overview of HIPE-2022 evaluation setting.

positives, and false negative figures computed over all documents, and macro average, based on averages of micro figures per document. Our definition of macro differs from the usual one: averaging is done at document level and not at entity type level. This allows to account for variance in document length and entity distribution within documents and avoids distortions that would occur due to the unevenly distributed entity classes.

Both NERC and EL benefit from strict and fuzzy evaluation regimes, depending on how strictly entity type and boundaries correctness are judged. For NERC (Coarse and Fine), the strict regime corresponds to exact type and boundary matching, and the fuzzy to exact type and overlapping boundaries. It is to be noted that in the strict regime, predicting wrong boundaries leads to a 'double' punishment of one false negative (entity present in the gold standard but not predicted by the system) and one false positive (entity predicted by the system but not present in the gold standard). Although it penalizes harshly, we keep this metric to be consistent with CoNLL and refer to the fuzzy regime when boundaries are of less importance.

The definition of strict and fuzzy regimes differs for entity linking. In terms of boundaries, EL is always evaluated according to overlapping boundaries in both regimes (what is of interest is the capacity to provide the correct link rather than the correct boundaries). EL strict regime considers only the system's top link prediction (NIL or Wikidata QID), while the fuzzy regime expands system predictions with a set of historically related entity QIDs. For example, "Germany" QID is complemented with the QID of the more specific "Confederation of the Rhine" entity and both are considered as valid answers. The resource allowing for such historical normalization was compiled by the task organizers for the entities of the test data sets (for `hipe2020` and `ajmc` datasets), and are released as part of the HIPE-scorer. For this regime, participants were invited to submit more than one link, and F-measure is additionally computed with cut-offs @3 and @5 (meaning, counting a true positive if the ground truth QID can be found within the first 3 or 5 candidates).

4.3 System Evaluation, Scorer and Evaluation Toolkit

Teams were asked to submit system responses based on submission bundles and to specify at least one challenge to which their submitted bundles belong. Micro and macro scores were computed and published for each track, but only micro figures are reported here.

The evaluation of challenges, which corresponds to an aggregation of tracks, was defined as follows: given a specific challenge and the tracks submitted by a team for this challenge, the submitted systems are rewarded points according to their F1-based rank for each track (considering only the best of the submitted runs for a given track). The points obtained are summed over all submitted tracks, and systems/teams are ranked according to their total points. Further details on system submission and evaluation can be found in the HIPE Participation Guidelines [11].

The evaluation is performed using the **HIPE-scorer**[17]. Developed during the first edition of HIPE, the scorer has been improved with minor bug fixes and additional parameterisation (input format, evaluation regimes, HIPE editions). Participants could use the HIPE-scorer when developing their systems. After the evaluation phase, a complete **evaluation toolkit** was also released, including the data used for evaluation (v2.1), the system runs submitted by participating teams, and all the evaluation recipes and resources (e.g. historical mappings) needed to replicate the present evaluation[18].

5 System Descriptions

In this second HIPE edition, 5 teams submitted a total of 103 system runs. Submitted runs do not cover all of the 35 possible tracks (dataset/language/task

[17] https://github.com/hipe-eval/HIPE-scorer

[18] https://github.com/hipe-eval/HIPE-2022-eval

combinations), nevertheless we received submission for all datasets, with most of them focusing on NERC-Coarse.

5.1 Baselines

As a neural baseline (**Neur-bsl**) for NERC-Coarse and NERC-fine, we fine-tuned separately for each HIPE-2022 dataset XLM-R$_{BASE}$, a multilingual transformer-based language representation model pre-trained on 2.5TB of filtered CommonCrawl texts [6]. The models are implemented using HuggingFace[19] [35]. Since transformers rely primarily on subword-based tokenisers, we chose to label only the first subwords. This allows to map the model outputs to the original text more easily. Tokenised texts are split into input segments of length 512. For each HIPE-2022 dataset, fine-tuning is performed on the train set (except for `sonar` and `hipe2020-en` which has only dev sets) for 10 epochs using the default hyperparameters (Adam $\epsilon = 10\text{e}{-8}$, Learning rate $\alpha = 5\text{e}{-5}$). The code of this baseline (configuration files, scripts) is published in a dedicated repository on the HIPE-eval Github organisation[20], and results are published in the evaluation toolkit.

For entity linking in EL-only setting, we provide the NIL baseline (**Nil-bsl**), where each entity link is replaced with the NIL value.

5.2 Participating Systems

The following system descriptions are compiled from information provided by the participants. More details on the implementation and results can be found in the system papers of the participants [16].

Team **L3i**, affiliated with *La Rochelle University* and with the *University of Toulouse*, France, successfully tackled an impressive amount of multilingual newspaper datasets with strong runs for NERC-coarse, NERC-fine and EL. For the classical commentary datasets (`ajmc`) the team had excellent results for NERC[21]. For NERC, L3i – the winning team in HIPE's 2020 edition – builds on their transformer-based approach [4]. Using transformer-based adapters [19], parameter-efficient fine-tuning in a hierarchical multitask setup (NERC-coarse and NERC-fine) has been shown to work well with historical noisy texts [4]. The innovation for this year's submission lies in the addition of context information in the form of external knowledge from two sources (inspired by [34]). First, French and German Wikipedia documents based on dense vector representations computed by a multilingual Sentence-BERT model [27], including a k-Nearest-Neighbor search functionality provided by ElasticSearch framework. Second, English Wikidata knowledge graph (KG) embeddings that are combined with the first paragraph of English Wikipedia pages (Wikidata5m) [33]. For the knowledge graph embeddings, two methods are tested on the HIPE-2022 data:

[19] https://github.com/huggingface/transformers/

[20] https://github.com/hipe-eval/HIPE-2022-baseline/

[21] The EL results for `ajmc` were low, probably due to some processing issues.

1) the one-stage KG Embedding Retrieval Module that retrieves top-k KG "documents" (in this context, a document is an ElasticSearch retrieval unit that consists of an entity identifier, an entity description and an entity embedding) via vector similarity on the dense entity embedding vector space; 2) the two-stage KG Embedding Retrieval Module that retrieves the single top similar document first and then in a second retrieval step gets the k most similar documents based on that first entity. All context enrichment techniques work by simply concatenating the original input segment with the retrieved context segments and processing the contextualized segments through their "normal" hierarchical NER architecture. Since the L3i team's internal evaluation on HIPE-2022 data (using a multilingual BERT base pre-trained model) indicated that the two-stage KG retrieval was the best context generator overall, it was used for one of the two officially submitted runs. The other "baseline" run did not use any context enrichment techniques. Both runs additionally used stacked monolingual BERT embeddings for English, French and German, for the latter two languages in the form of Europeana models that were built from digitized historical newspaper text material. Even with improved historical monolingual BERT embeddings, the context-enriched run was consistently better in terms of F1-score in NERC-Coarse and -Fine settings.

Team **HISTeria**, affiliated with the *Bayerische Staatsbibliothek München*, Germany, the *Digital Philology* department of the University of Vienna, Austria and the *NLP Expert Center, Volkswagen AG* Munich, Germany, focused on the `ajmc` dataset for their NERC-coarse submission (best results for French and English, second best for German), but also provides experimental results for all languages of the `newseye` datasets[22]. Their NER tagging experiments tackle two important questions:

a) How to build an optimal multilingual pre-trained BERT language representation model for historical OCRized documents? They propose and release hmBERT[23], which includes English, Finnish, French, German and Swedish in various model and vocabulary sizes, and specifically apply methods to deal with OCR noise and imbalanced corpus sizes per language. In the end, roughly 27GB of text per language is used in pre-training.

b) How to fine-tune a multilingual pre-trained model given comparable NER annotations in multiple languages? They compare a single-model approach (training models separately for each language) with a one-model approach (training only one model that covers all languages). The results indicate that, most of the time, the single-model approach works slightly better, but the difference may not be large enough to justify the considerably greater effort to train and apply the models in practice.

[22] Note that these experiments are evaluated using the officially published *Newseye* test sets [18] (released as dev2 dataset as part of HIPE-2022) and not the HIPE-2022 `newseye` test sets, which were unpublished prior to the HIPE 2022 campaign.

[23] For English data, they used the Digitised Books. c. 1510 - c. 1900, all other languages use Europeana newspaper text data.

HISTeria submitted two runs for each ajmc datasets, using careful hyperparameter grid search on the dev sets in the process. Both runs build on the one-model approach in a first multilingual fine-tuning step. Similar to [34], they build monolingual models by further fine-tuning on language-specific training data[24]. Run 1 of their submission is based on hmBERT with vocabulary size 32k, while run 2 has a vocabulary size of 64k. Somewhat unexpectedly, the larger vocabulary does not improve the results in general on the development set. For the test set, though, the larger vocabulary model is substantially better overall. Similar to the team L3i, HISTeria also experimented with context enrichment techniques suggested by [34]. However, for the specific domain of classical commentaries, general-purpose knowledge bases such as Wikipedia could not improve the results. Interestingly, L3i also observed much less improvement with Wikipedia context enrichment on ajmc in comparison to the hipe2020 newspaper datasets. In summary, HISTeria outperformed the strong neural baseline by about 10 F1-score percentage points in strict boundary setting, thereby demonstrating the importance of carefully constructed domain-specific pre-trained language representation models.

Team **Aauzh**, affiliated with *University of Zurich*, Switzerland and *University of Milan*, Italy, focused on the multilingual newspaper challenge in NERC-coarse setting and experimented with 21 different monolingual and multilingual, as well as contemporary and historical transformer-based language representation models available on the HuggingFace platform. For fine-tuning, they used the standard token classification head of the transformer library for NER tagging with default hyperparameters and trained each dataset for 3 epochs. In a preprocessing step, token-level NER IOB labels were mapped onto all subtokens. At inference time, a simple but effective summing pooling strategy for NER for aggregating subtoken-level to token-level labels was used [1]. Run 2 of Aauzh are the predictions of the best single model. Run 1 is the result of a hard-label ensembling from different pre-trained models: in case of ties between O and B/I labels, the entity labels were preferred. The performance of the submitted runs varies strongly in comparison with the neural baseline: for German and English it generally beats the baseline clearly for hipe2020 and sonar datasets, but suffers on French hipe2020 and German/Finnish newseye datasets. This again indicates that in transfer learning approaches to historical NER, the selection of pre-trained models has a considerable impact. The team also performed some post-submission experiments to investigate the effect of design choices: Applying soft-label ensembling using averaged token-level probabilities turned out to improve results on the French newseye datasets by 1.5 percentage point in micro average and 2.4 points in macro average (F1-score). For all languages of the newseye, they also tested a one-model approach with multilingual training. The best multilingual dbmdz Europeana BERT model had a better performance on average (58%) than the best monolingual models (56%). However, several other multilingual pre-trained language models had substantially worse performance,

[24] This improves the results by 1.2% on average on the HIPE-2022 data.

resulting in 57% ensemble F1-score (5 models), which was much lower than 67% achieved by the monolingual ensemble.

Team **Sbb**, affiliated with the *Berlin State Library*, Germany, participated exclusively in the EL-only subtask, but covered all datasets in English, German and French. Their system builds on models and methods developed in the HIPE-2020 edition [22]. Their approach uses Wikipedia sentences with an explicit link to a Wikipedia page as textual representations of its connected Wikidata entity. The system makes use of the metadata of the HIPE-2022 documents to exclude entities that were not existing at the time of its publication. Going via Wikipedia reduces the amount of accessible Wikidata IDs, however, for all datasets but `ajmc` the coverage is still 90%. Given the specialised domain of `ajmc`, a coverage of about 55% is to be accepted. The entity linking is done in the following steps: a) A candidate lookup retrieves a given number of candidates (25 for submission run 1, 50 for submission run 2) using a nearest neighbour index based on word embeddings of Wikipedia page titles. An absolute cut-off value is used to limit the retrieval (0.05 for submission 1 and 0.13 for submission 2). b) A probabilistic candidate sentence matching is performed by pairwise comparing the sentence with the mention to link and a knowledge base text snippet. To this end, a BERT model was fine-tuned on the task of whether or not two sentences mention the same entity. c) The final ranking of candidates includes the candidate sentence matching information as well as lookup features from step (a) and more word embedding information from the context. A random forest model calculates the overall probability of a match between the entity mention and an entity linking candidate. If the probability of a candidate is below a given threshold (0.2 for submission run 1 and 2), it is discarded. The random forest model was trained on concatenated training sets of the same language across datasets.

There are no conclusive insights from HIPE-2022 EL-only results whether run 1 or 2 settings are preferable. Post-submission experiments in their system description paper investigate the influence of specific hyperparameter settings on the system performances.

Team **WLV**, University of Wolverhampton, UK, applied classical BERT-based [7] as well as BiLSTM-based NER architectures [20] with a CRF layer to HIPE-2022 English and French newspaper datasets `hipe2020`, `letemps`, and `topres19th` in the NERC-coarse subtask.

6 Results and Discussion

We report results for the best run of each team and consider micro Precision, Recall and F1-score exclusively. Results for NERC-Coarse and NERC-Fine for all languages and datasets according to both evaluation regimes are presented in Table 6 and 7 respectively. Table 8 reports performances for EL-only, with a cut-off @1. We refer the reader to the HIPE-2022 website and the evaluation toolkit for more detailed results[25], and to the extended overview paper for fur-

[25] See https://hipe-eval.github.io/HIPE-2022 and https://github.com/hipe-eval/HIPE-2022-eval

ther discussion of the results [15].

General observations. All systems now use transformer-based approaches with strong pre-trained models. The choice of the pre-trained model – and the corresponding text types used in pre-training – have a strong influence on performance.

The quality of available multilingual pre-trained models for fine-tuning on NER tasks proved to be competitive compared to training individual monolingual models. However, to get the maximum performance out of it, the multilingual fine-tuning in a first phase must be complemented by a monolingual second phase.

NERC. In general, the systems demonstrated a good ability to adapt to heterogeneous annotation guidelines. They achieved their highest F1-scores for the NERC-Coarse task on `ajmc`, a dataset annotated with domain-specific entities and of relatively small size compared to the newspaper datasets, thus confirming the ability of strong pre-trained models to achieve good results when fine-tuned on relatively small datasets. The good results obtained on `ajmc`, however, may be partly due to the relatively high mention overlap between train and test sets (see Section 3.2). Moreover, it is worth noting that performances on the French subset of the `ajmc` dataset do not substantially degrade despite the high rate of noisy mentions (three times higher than English and German), which shows a good resilience of transformer-based models to OCR noise on this specific dataset.

EL-only. Entity linking on already identified mentions appears to be considerably more challenging than NERC, with F1-scores varying considerably across datasets. The linking of toponyms in `topres19th` is where systems achieved the overall best performances. Conversely, EL-only on historical commentaries (`ajmc`) appears to be the most difficult, with the lowest F1-scores compared to the other datasets.

The EL-only performances of the SBB system on the `ajmc` dataset deserve some further considerations, as they are well representative of the challenges faced when applying a generic entity linking system to a domain-specific dataset. Firstly, SBB team reported that `ajmc` is the dataset with the lowest Wikidata coverage: only 57% of the Wikidata IDs in the test set are found in the knowledge base used by their system (a combination of Wikidata record and Wikipedia textual content), whereas the coverage for all other datasets ranges between 86% (`hipe2020`) and 99% (`topres19th`). The reason for the low coverage in `ajmc` is that, when constructing the knowledge base, only Wikidata records describing persons, locations and organisations were kept. In contrast, a substantial number of entities in `ajmc` are literary works, which would have required to retain also records with Wikidata type "literary work" (`Q7725634`) when building the KB.

Secondly, a characteristic of `ajmc` is that both person and work mentions are frequently abbreviated, and these abbreviations tend to be lacking as lexical information in large-scale KBs such as Wikidata. Indeed, an error analysis of SBB's system results shows that only 1.4% of the correctly predicted entity

Table 6: Results for NERC-Coarse (micro P, R and F1-score). Bold font indicates the highest, and underlined font the second-highest value.

	Strict			Fuzzy			Strict			Fuzzy			Strict			Fuzzy		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
hipe2020																		
	French						German						English					
Aauzh	.718	.675	.696	.825	.776	.800	.716	.735	<u>.725</u>	.812	.833	<u>.822</u>	.538	490	<u>.513</u>	.726	.661	<u>.692</u>
L3i	.786	.831	**.808**	.883	.933	**.907**	.784	.805	**.794**	.865	.888	**.876**	.624	.617	**.620**	.793	.784	**.788**
Wlv	.640	.712	.674	.767	.853	.808	-	-	-	-	-	-	.400	.430	.414	.582	.626	.603
Neur-bsl	.730	.785	<u>.757</u>	.836	.899	<u>.866</u>	.665	.746	.703	.750	.842	.793	.432	.532	.477	.564	.695	.623
	letemps						**sonar**						**topRes19th**					
	French						German						English					
Aauzh	.589	.710	<u>.644</u>	.642	.773	<u>.701</u>	.512	.548	**.529**	.655	.741	**.695**	.816	.760	**.787**	.869	.810	**.838**
Wlv	.581	.659	.618	.627	.711	.666	-	-	-	-	-	-	.712	.771	.740	.765	.829	.796
Neur-bsl	.595	.744	**.661**	.639	.800	**.711**	.267	.361	.307	.410	.554	.471	.747	.782	<u>.764</u>	.798	.836	<u>.816</u>
ajmc																		
	French						German						English					
HISTeria	.834	.850	**.842**	.874	.903	**.888**	.930	.898	.913	.938	.953	.945	.826	.885	**.854**	.879	.943	**.910**
L3i	.810	.842	.826	.856	.889	.872	.946	.921	**.934**	.965	.940	**.952**	.824	.876	.850	.868	.922	.894
Neur-bsl	.707	.778	.741	.788	.867	.825	.792	.846	.818	.846	.903	.873	.680	.802	.736	.766	.902	.828
newseye																		
	French						German											
Aauzh	.655	.657	**.656**	.785	.787	**.786**	.395	.421	.408	.480	.512	.495						
Neur-bsl	.634	.676	.654	.755	.805	.779	.429	.537	**.477**	.512	.642	**.570**						
	Finnish						Swedish											
Aauzh	.618	.524	.567	.730	.619	.670	.686	.604	.643	.797	.702	.746						
Neur-bsl	.605	.687	**.644**	.715	.812	**.760**	.588	.728	**.651**	.675	.836	**.747**						

Table 7: Results for NERC-Fine and Nested (micro P, R and F1-score).

	French						**German**						**English**					
	Strict			Fuzzy			Strict			Fuzzy			Strict			Fuzzy		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
hipe2020 (Fine)																		
L3i	.702	.782	**.740**	.784	.873	**.826**	.691	.747	**.718**	.776	.840	**.807**						
Neur-bsl	.685	.733	.708	.769	.822	.795	.584	.673	.625	.659	.759	.706						
hipe2020 (Nested)																		
L3i	.390	.366	.377	.416	.390	.403	.714	.411	.522	.738	.425	.539						
ajmc (Fine)																		
L3i	.646	.694	**.669**	.703	.756	**.728**	.915	.898	**.906**	.941	.924	**.933**	.754	.848	**.798**	.801	.899	**.847**
Neur-bsl	.526	.567	.545	.616	.664	.639	.819	.817	.818	.866	.864	.865	.600	.744	.664	.676	.839	.749

Table 8: Results for EL-only (micro P, R and F1-score @1). Bold font indicates the highest value.

	Strict			Fuzzy			Strict			Fuzzy			Strict			Fuzzy		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
							hipe2020											
	French						German						English					
L3i	.602	.602	**.602**	.620	.620	**.620**	.481	.481	.481	.497	.497	.497	.546	.546	**.546**	.546	.546	**.546**
SBB	.707	.515	.596	.730	.532	.616	.603	.435	**.506**	.626	.452	**.525**	.503	.323	.393	.503	.323	.393
Nil-bsl	.209	.209	.209	.209	.209	.209	.481	.314	.380	.481	.314	.380	.228	.228	.228	.228	.228	.228
							sonar						topres19th					
							German						**English**					
SBB							.616	.446	**.517**	.616	.446	**.517**	.778	.559	.651	.781	.562	.654
Nil-bsl							.333	.333	.333	.333	.333	.333	-	-	-	-	-	-
							newseye											
	French						German											
SBB	.534	.361	.431	.539	.364	.435	.522	.387	.444	.535	.396	.455						
Nil-bsl	.448	.448	**.448**	.448	.448	**.448**	.485	.485	**.485**	.485	.485	**.485**						
							ajmc											
	French						German						English					
SBB	.621	.378	**.470**	.614	.373	**.464**	.712	.389	**.503**	.712	.389	**.503**	.578	.284	**.381**	.578	.284	**.381**
Nil-bsl	.037	.037	.037	.037	.037	.037	.049	.049	.049	.049	.049	.049	.046	.046	.046	.046	.046	.046

links (true positives) correspond to abbreviated mentions, which nevertheless represent about 47% of all linkable mentions.

Unfortunately, no team has worked on adapting annotation models to be able to use different NER training datasets with sometimes incompatible annotations and benefit from a larger dataset overall. Tackling this challenge remains future work.

7 Conclusion and Perspectives

From the perspective of natural language processing, this second edition of HIPE provided the possibility to test the robustness of existing approaches and to experiment with transfer learning and domain adaptation methods, whose performances could be systematically evaluated and compared on broad historical and multilingual data sets. Besides 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 also contributed an unprecedented set of multilingual and historical NE-annotated datasets that can be used for further experimentation and benchmarking.

From the perspective of digital humanities, the lab's outcomes will help DH practitioners in mapping state-of-the-art solutions for NE processing of historical texts, and in getting a better understanding of what is already possible as opposed to what is still challenging. Most importantly, digital scholars are in need of support to explore the large quantities of digitised text they currently have at hand, and NE processing is high on the agenda. Such processing can support research questions in various domains (e.g. history, political science, literature, historical linguistics) and knowing about their performance is crucial in order to make an informed use of the processed data.

From the perspective of cultural heritage professionals, who increasingly focus on advancing the usage of artificial intelligence methods on cultural heritage text collections [26, 17], the HIPE-2022 shared task and datasets represent an excellent opportunity to experiment with multilingual and multi-domain data of various quality and annotation depth, a setting close to the real-world scenarios they are often confronted with.

Overall, HIPE-2022 has contributed to further advance the state of the art in semantic indexing of historical documents. By expanding the language spectrum and document types and integrating datasets with various annotation tag sets, this second edition has set the bar high, and there remains much to explore and experiment.

Acknowledgements The HIPE-2022 team expresses her greatest appreciation to the HIPE-2022 partnering projects, namely AjMC, impresso-HIPE-2020, Living with Machines, NewsEye, and SoNAR, for contributing (and hiding) their NE-annotated datasets. We particularly thank Mariona Coll-Ardanuy (LwM), Ahmed Hamdi (NewsEye) and Clemens Neudecker (SoNAR) for their support regarding data provision, and the members of the HIPE-2022 advisory board, namely Sally Chambers, Frédéric Kaplan and Clemens Neudecker.

Bibliography

[1] Ács, J., Kádár, Á., Kornai, A.: Subword pooling makes a difference. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 2284–2295. Association for Computational Linguistics, Online (Apr 2021). https://doi.org/10.18653/v1/2021.eacl-main.194, https://aclanthology.org/2021.eacl-main.194

[2] Augenstein, I., Derczynski, L., Bontcheva, K.: Generalisation in named entity recognition: A quantitative analysis. Computer Speech & Language **44**, 61–83 (Jul 2017). https://doi.org/10.1016/j.csl.2017.01.012, http://www.sciencedirect.com/science/article/pii/S088523081630002X

[3] Beryozkin, G., Drori, Y., Gilon, O., Hartman, T., Szpektor, I.: 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. pp. 140–150. Florence, Italy (Jul 2019), https://aclanthology.org/P19-1014

[4] Boros, E., Hamdi, A., Linhares Pontes, E., Cabrera-Diego, L.A., Moreno, J.G., Sidere, N., Doucet, A.: Alleviating digitization errors in named entity recognition for historical documents. In: Proceedings of the 24th Conference on Computational Natural Language Learning. pp. 431–441. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.conll-1.35

[5] Coll Ardanuy, M., Beavan, D., Beelen, K., Hosseini, K., Lawrence, J.: Dataset for Toponym Resolution in Nineteenth-Century English Newspapers (2021). https://doi.org/10.23636/b1c4-py78, https://https//doi.org/10.23636/b1c4-py78

[6] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised Cross-lingual Representation Learning at Scale (Apr 2020)

[7] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423

[8] Ehrmann, M., Colavizza, G., Rochat, Y., Kaplan, F.: Diachronic evaluation of NER systems on old newspapers. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016). pp. 97–107. Bochumer Linguistische Arbeitsberichte, Bochum (2016), https://infoscience.epfl.ch/record/221391

[9] Ehrmann, M., Hamdi, A., Pontes, E.L., Romanello, M., Doucet, A.: Named Entity Recognition and Classification on Historical Documents: A Survey.

arXiv:2109.11406 [cs] (Sep 2021 (to appear in ACM journal *Computing Surveys* in 2022), http://arxiv.org/abs/2109.11406, arXiv: 2109.11406

[10] Ehrmann, M., Romanello, M., Clematide, S., Ströbel, P.B., Barman, R.: Language Resources for Historical Newspapers: The *Impresso* Collection. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 958–968. European Language Resources Association, Marseille, France (May 2020)

[11] Ehrmann, M., Romanello, M., Doucet, A., Clematide, S.: HIPE 2022 Shared Task Participation Guidelines. Tech. rep., Zenodo (Feb 2022). https://doi.org/10.5281/zenodo.6045662, https://zenodo.org/record/6045662

[12] Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Impresso Named Entity Annotation Guidelines. Annotation guidelines, Ecole Polytechnique Fédérale de Lausanne (EPFL) and Zurich University (UZH) (Jan 2020). https://doi.org/10.5281/zenodo.3604227, https://zenodo.org/record/3585750

[13] Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Overview of CLEF HIPE 2020: Named entity recognition and linking on historical newspapers. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 288–310. Lecture Notes in Computer Sciences, Springer International Publishing, Cham (2020)

[14] Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. vol. 2696, p. 38. CEUR-WS, Thessaloniki, Greece (2020). https://doi.org/10.5281/zenodo.4117566, https://infoscience.epfl.ch/record/281054

[15] Ehrmann, M., Romanello, M., Najem-Meyer, S., Doucet, A., Clematide, S.: Extended Overview of HIPE-2022: Named Entity Recognition and Linking on Multilingual Historical Documents. In: Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (eds.) Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum. CEUR-WS (2022)

[16] Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (eds.): Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum. CEUR-WS (2022)

[17] Gregory, M., Neudecker, C., Isaac, A., Bergel, G., Others: AI in relation to GLAMs task FOrce - Report and Recommendations. Tech. rep., Europeana Network ASsociation (2021), https://pro.europeana.eu/project/ai-in-relation-to-glams

[18] Hamdi, A., Linhares Pontes, E., Boros, E., Nguyen, T.T.H., Hackl, G., Moreno, J.G., Doucet, A.: 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. pp. 2328–2334. SIGIR

'21, Association for Computing Machinery, New York, NY, USA (Jul 2021). https://doi.org/10.1145/3404835.3463255

[19] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 2790–2799. PMLR (09–15 Jun 2019), https://proceedings.mlr.press/v97/houlsby19a.html

[20] Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. ArXiv **abs/1508.01991** (2015)

[21] Kaplan, F., di Lenardo, I.: Big Data of the Past. Frontiers in Digital Humanities **4**, 1–21 (2017). https://doi.org/10.3389/fdigh.2017.00012, https://www.frontiersin.org/articles/10.3389/fdigh.2017.00012/full, publisher: Frontiers

[22] Labusch, K., Neudecker, C.: Named entity disambiguation and linking on historic newspaper OCR with BERT. In: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. No. 2696 in CEUR Workshop Proceedings, CEUR-WS (September 2020), http://ceur-ws.org/Vol-2696/paper_163.pdf

[23] Li, J., Chiu, B., Feng, S., Wang, H.: Few-shot named entity recognition via meta-learning. IEEE Transactions on Knowledge and Data Engineering pp. 1–1 (2020)

[24] Li, J., Shang, S., Shao, L.: Metaner: Named entity recognition with meta-learning. In: Proceedings of The Web Conference 2020. p. 429–440. WWW '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3366423.3380127, https://doi.org/10.1145/3366423.3380127

[25] Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R.: Performance measures for information extraction. In: In Proceedings of DARPA Broadcast News Workshop. pp. 249–252 (1999)

[26] Padilla, T.: Responsible Operations: Data Science, Machine Learning, and AI in Libraries. Tech. rep., OCLC Research, USA (May 2020). https://doi.org/10.25333/xk7z-9g97

[27] Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4512–4525. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.365, https://aclanthology.org/2020.emnlp-main.365

[28] Ridge, M., Colavizza, G., Brake, L., Ehrmann, M., Moreux, J.P., Prescott, A.: The past, present and future of digital scholarship with newspaper collections. In: DH 2019 book of abstracts. pp. 1–9. Utrecht, The Netherlands (2019), http://infoscience.epfl.ch/record/271329

[29] Romanello, M., Najem-Meyer, S.: Guidelines for the Annotation of Named Entities in the Domain of Classics (Mar 2022). https://doi.org/10.5281/zenodo.6368101, https://doi.org/10.5281/zenodo.6368101

[30] Romanello, M., Sven, N.M., Robertson, B.: 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). Association for Computing Machinery, Lausanne (Sep 2021). https://doi.org/10.1145/3476887.3476911, https://doi.org/10.1145/3476887.3476911
[31] Rosset, S., Grouin, Cyril, Zweigenbaum, Pierre: Entités nommées structurées : Guide d'annotation Quaero. Tech. Rep. 2011-04, LIMSI-CNRS, Orsay, France (2011)
[32] Taillé, B., Guigue, V., Gallinari, P.: Contextualized Embeddings in Named-Entity Recognition: An Empirical Study on Generalization. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval. pp. 383–391. Lecture Notes in Computer Science, Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_48
[33] Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., Tang, J.: Kepler: A unified model for knowledge embedding and pre-trained language representation (11 2019), https://arxiv.org/pdf/1911.06136.pdf
[34] Wang, X., Shen, Y., Cai, J., Wang, T., Wang, X., Xie, P., Huang, F., Lu, W., Zhuang, Y., Tu, K., Lu, W., Jiang, Y.: DAMO-NLP at SemEval-2022 task 11: A knowledge-based system for multilingual named entity recognition (2022). https://doi.org/10.48550/ARXIV.2203.00545, https://arxiv.org/abs/2203.00545
[35] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020), https://www.aclweb.org/anthology/2020.emnlp-demos.6
[36] Wu, Q., Lin, Z., Wang, G., Chen, H., Karlsson, B.F., Huang, B., Lin, C.: Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. CoRR **abs/1911.06161** (2019), http://arxiv.org/abs/1911.06161