Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers

Maud Ehrmann[1][0000−0001−9900−2193], Matteo Romanello[1][0000−0002−1890−2577], Alex Flückiger[2], and Simon Clematide[2][0000−0003−1365−0662]

[1] Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{maud.ehrmann,matteo.romanello}@epfl.ch
[2] University of Zurich, Zurich, Switzerland
{alex.flueckiger,simon.clematide}@uzh.ch

Abstract. This paper presents an overview of the first edition of HIPE (Identifying Historical People, Places and other Entities), a pioneering shared task dedicated to the evaluation of named entity processing on historical newspapers in French, German and English. Since its introduction some twenty years ago, named entity (NE) processing has become an essential component of virtually any text mining application and has undergone major changes. Recently, two main trends characterise its developments: the adoption of deep learning architectures and the consideration of textual material originating from historical and cultural heritage collections. While the former opens up new opportunities, the latter introduces new challenges with heterogeneous, historical and noisy inputs. In this context, the objective of HIPE, run as part of the CLEF 2020 conference, is threefold: strengthening the robustness of existing approaches on non-standard inputs, enabling performance comparison of NE processing on historical texts, and, in the long run, fostering efficient semantic indexing of historical documents. Tasks, corpora, and results of 13 participating teams are presented.

Keywords: Named entity recognition and classification · Entity linking · Historical texts · Information extraction · Digitized newspapers · Digital humanities

1 Introduction

Recognition and identification of real-world entities is at the core of virtually any text mining application. As a matter of fact, referential units such as names of persons, locations and organizations underlie the semantics of texts and guide their interpretation. Around since the seminal Message Understanding Conference (MUC) evaluation cycle in the 1990s [18], named entity-related tasks have

The final authenticated version is available online at https://doi.org/10.1007/978-3-030-58219-7_21

undergone major evolutions until now, from entity recognition and classification to entity disambiguation and linking [33, 43].

Context. Recently, two main trends characterise developments in NE processing. First, at the technical level, the adoption of deep learning architectures and the usage of embedded language representations greatly reshapes the field and opens up new research directions [2, 27, 26]. Second, with respect to application domain and language spectrum, NE processing has been called upon to contribute to the field of Digital Humanities (DH), where massive digitization of historical documents is producing huge amounts of texts [50]. Thanks to large-scale digitization projects driven by cultural institutions, millions of images are being acquired and, when it comes to text, their content is transcribed, either manually via dedicated interfaces, or automatically via Optical Character Recognition (OCR). Beyond this great achievement in terms of document preservation and accessibility, the next crucial step is to adapt and develop appropriate language technologies to search and retrieve the contents of this 'Big Data from the Past' [22]. In this regard, information extraction techniques, and particularly NE recognition and linking, can certainly be regarded among the first and most crucial processing steps.

Motivation. Admittedly, NE processing tools are increasingly being used in the context of historical documents. Research activities in this domain target texts of different nature (e.g., museum records, state-related documents, genealogical data, historical newspapers) and different tasks (NE recognition and classification, entity linking, or both). Experiments involve different time periods, focus on different domains, and use different typologies. This great diversity demonstrates how many and varied the needs—and the challenges—are, but also makes performance comparison difficult, if not impossible.

Furthermore, it appears that historical texts poses new challenges to the application of NE processing [11, 41], as it does for language technologies in general [47]. First, inputs can be extremely noisy, with errors which do not resemble tweet misspellings or speech transcription hesitations, for which adapted approaches have already been devised [29, 7, 46]. Second, the language under study is mostly of earlier stage(s), which renders usual external and internal evidences less effective (e.g., the usage of different naming conventions and presence of historical spelling variations) [5, 4]. Further, beside historical VIPs, texts from the past contain rare entities which have undergone significant changes (esp. locations) or do no longer exist, and for which adequate linguistic resources and knowledge bases are missing [20]. Finally, archives and texts from the past are not as anglophone as in today's information society, making multilingual resources and processing capacities even more essential [34].

Overall, and as demonstrated by Vilain et al. [52], the transfer of NE tools from one domain to another is not straightforward, and the performance of NE tools initially developed for homogeneous texts of the immediate past are affected when applied on historical materials [48]. This echoes the proposition of Plank [42], according to whom what is considered as standard data (i.e. contemporary news genre) is more a historical coincidence than a reality: in NLP non-canonical, heterogeneous, biased and noisy data is rather the norm than the exception.

Objectives. In this context of new needs and materials emerging from the humanities, the HIPE shared task[3] puts forward for the first time the systematic evaluation of NE recognition and linking on diachronic historical newspaper material in French, German and English. In addition to the release of a multilingual, historical NE-annotated corpus, the objective of this shared task is threefold:

1. strengthening the robustness of existing approaches on non-standard inputs;
2. enabling performance comparison of NE processing on historical texts;
3. fostering efficient semantic indexing of historical documents in order to support scholarship on digital cultural heritage collections.

Even though many evaluation campaigns on NE were organized over the last decades[4], only one considered French historical texts [16]. To the best of our knowledge, no NE evaluation campaign ever addressed multilingual, diachronic historical material. The present shared task is organized as part of "*impresso* - Media Monitoring of the Past", a project which tackles information extraction and exploration of large-scale historical newspapers.[5]

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 metrics and the organisation of system submissions. Section 5 introduces the 13 participating systems while Section 6 presents and discusses their results. Finally, Section 7 summarizes the benefits of the task and concludes.[6]

2 Task Description

The HIPE shared task includes two NE processing tasks with sub-tasks of increasing level of difficulty.

Task 1: Named Entity Recognition and Classification (NERC)

- **Subtask 1.1 - NERC coarse-grained** (NERC-Coarse): this task includes the recognition and classification of entity mentions according to high-level entity types.
- **Subtask 1.2 - NERC fine-grained** (NERC-Fine): this task includes the recognition and classification of mentions according to finer-grained entity types, as well as of nested entities and entity mention components (e.g. function, title, name).

[3] `https://impresso.github.io/CLEF-HIPE-2020/`

[4] MUC, ACE, CONLL, KBP, ESTER, HAREM, QUAERO, GERMEVAL, etc.

[5] `https://impresso-project.ch/`

[6] For space reasons, the discussion of related work is included in the extended version of this overview [12].

Types	Sub-types	
pers	pers.ind	pers.ind.articleauthor
	pers.coll	
org	org.ent	org.ent.pressagency
	org.adm	
prod	prod.media	
	prod.doctr	
date	time.date.abs	
loc	loc.adm	loc.adm.town
		loc.adm.reg
		loc.adm.nat
		loc.adm.sup
	loc.phys	loc.geo
		loc.hydro
		loc.astro
	loc.oro	
	loc.fac	
	loc.add	loc.add.phys
		loc.add.elec

Table 1: Entity types used for NERC tasks.

Task 2: Named Entity Linking (EL). This task requires the linking of named entity mentions to a unique referent in a knowledge base – here Wikidata – or to a NIL node if the mention's referent is not present in the base. The entity linking task includes two settings: without and with prior knowledge of mention types and boundaries, referred to as end-to-end EL and EL only respectively.

3 Data

3.1 Corpus

The shared task corpus is composed of digitized and OCRized articles originating from Swiss, Luxembourgish and American historical newspaper collections and selected on a diachronic basis.[7]

Corpus selection. The corpus was compiled based on systematic and purposive sampling. For each newspaper and language, articles were randomly sampled among articles that a) belong to the first years of a set of predefined decades covering the life-span of the newspaper (longest duration spans ca. 200 years),

[7] From the Swiss National Library, the Luxembourgish National Library, and the Library of Congress (Chronicling America project), respectively. Original collections correspond to 4 Swiss and Luxembourgish titles, and a dozen for English. More details on original sources can be found in [12].

and b) have a title, have more than 50 characters, and belong to any page. For each decade, the set of selected articles was additionally manually triaged in order to keep journalistic content only. Items corresponding to feuilleton, tabular data, cross-words, weather forecasts, time-schedules, obituaries, and those with contents that a human could not even read because of extreme OCR noise were therefore removed. Different OCR versions of same texts are not provided, and the OCR quality of the corpus therefore corresponds to real-life setting, with variations according to digitization time and preservation state of original documents. The corpus features an overall time span of ca. 200 years, from 1798 to 2018.

Corpus annotation. The corpus was manually annotated according to the HIPE annotation guidelines [14]. Those guidelines were derived from the Quaero annotation guide, originally designed for the annotation of named entities in French speech transcriptions and already used on historical press corpora [45, 44]. HIPE slightly recast and simplified this guide, considering only a subset of entity types and components, as well as of linguistic units eligible as named entities. HIPE guidelines were iteratively consolidated via the annotation of a "mini-reference" corpus – consisting of 10 content items per language –, where annotation decisions were tested and difficult cases discussed. Despite these adaptations, the HIPE corpus mostly remain compatible with Quaero-annotated data, as well as with the NewsEye project's NE data sets[8], annotated with guidelines derived from HIPE.

Table 1 presents the entity types and sub-types used for annotation, which participant systems had to recognize for NERC-Coarse (types) and NERC-Fine (most fine-grained sub-types). Named entity components, annotated for the type `Person` only, correspond to `name`, `title`, `function`, `qualifier` and `demonym`. Nested entities were annotated for `Person`, `Organization` and `Location` (a depth of 1 was considered during the evaluation), as well as metonymic senses, producing double tags for those entities referring to something intimately associated (metonymic sense) to the concept usually associated with their name (literal sense). As per entity linking, links correspond to Wikidata QID[9].

The annotation campaign was carried out by the task organizers with the contribution of trilingual collaborators. We used the INCEpTION annotation tool [23], which allows the visualisation of image segments alongside OCR transcriptions. Before starting annotating, each annotator was first trained on the mini-reference corpus in order to ensure a good understanding of the guidelines. The inter-annotator agreement rates between 2 annotators was computed on a selection of documents (test set) using Krippendorf's α [25]. Scores correspond to, for Fr, De and En respectively: .81, .79 and .80 for NERC, .73, .69 and .78 for linking towards a QID, and .95, .94 and .90 for linking towards NIL. NERC and linking towards NIL show a good agreement between annotators. The lower

[8] `https://www.newseye.eu/`

[9] The November 2019 dump used for annotation is available at `https://files.ifi.uzh.ch/cl/impresso/clef-hipe`.

	Lang.	docs	tokens	mentions	nested	comp.	% meto.	% NIL	% noisy
Train	Fr	158	129,925	7885	480	3091	12.10	22.04	-
	De	103	71,507	3988	160	1494	16.75	13.67	-
	All	261	201,432	11,873	640	4585	13.66	19.23	-
Dev	Fr	43	29,571	1938	98	743	11.76	17.75	-
	De	33	27,032	1403	65	489	13.68	16.25	-
	En	80	24,266	1032	-	-	4.26	40.79	-
	All	156	80,869	4373	163	1232	10.61	22.71	-
Test	Fr	43	32,035	1802	83	732	13.32	17.70	12.15
	De	48	24,771	1317	64	431	18.45	14.35	13.74
	En	46	13,925	483	-	-	10.77	36.02	6.21
	All	137	70,731	3602	147	1163	14.85	18.93	11.10
All	Fr	244	191,531	11,625	661	4566	13.39	18.87	-
	De	184	123,310	6708	289	2414	16.44	14.34	-
	En	126	38,191	1515	-	-	11.17	24.82	-
	All	554	353,032	19,848	950	6980	13.38	19.67	-

Table 2: Overview of corpus statistics (v1.3).

scores on entity linking confirm the difficulty of the task, especially in the context of historical documents where, almost as a detective, one has to research the correct entities. The low score observed on German (.69) is due to annotation discrepancies with respect to the linking of metonymic entities. The historical normalization of the fuzzy evaluation regime for EL (see Section 4.1) helps mitigate these flaws.

Corpus characteristics. For each task and language—with the exception of English—the HIPE corpus was divided into training, dev and test data sets (70/15/15). English was included later in the shared task and only dev and test sets were released for this language. The overall corpus consists of 554 annotated documents, for a total of 353,032 tokens and 19,848 (linked) mentions (see Table 2 for detailed overview statistics). With 11,625 and 6,708 mentions, French and German corpora are larger than the English one (1,515). Despite our efforts to devise a balanced sampling strategy, the diachronic distribution of mentions is not entirely uniform across languages (see Fig. 1). This is mainly due to the following factors: the temporal boundaries of data to sample from (the German corpus stops at 1950, and the English one shortly afterwards); the varying content of newspaper articles; and, finally, the difficulty of sampling enough materials for certain decades due to OCR noise, such is the case with years 1850-1879 in the English corpus.

An important aspect of the HIPE corpus, and of historical newspaper data in general, is the noise generated by OCR. Annotators were asked to transcribe the surface form of noisy mentions so as to enable studying the impact of noisy

	Lang.	mentions	% nested	% meto.	% NIL
Person	Fr	3745	1.04	0.40	44.49
	De	1867	1.39	0.27	29.24
	En	558	-	0.18	72.40
	All	6170	1.05	0.34	42.40
Location	Fr	5278	10.27	13.02	4.53
	De	3148	6.26	17.15	3.91
	En	599	-	6.01	13.69
	All	9025	8.19	13.99	4.92
Organisation	Fr	1873	3.74	0.16	19.54
	De	1213	4.29	0.25	17.07
	En	241	-	5.81	31.95
	All	3327	3.67	0.60	19.54
Date	Fr	399	0.00	0.00	-
	De	241	2.49	0.00	-
	En	46	-	0.00	-
	All	686	0.87	0.00	-
Media	Fr	313	0.96	0.32	24.92
	De	227	1.32	0.88	31.28
	En	52	-	0.00	61.54
	All	592	1.01	0.51	30.57

Table 3: Statistics per coarse entity type (all data sets).

mentions on NERC and EL tasks. In the test set—where we manually verified the consistency of annotators' transcriptions—about 11% of all mentions contain OCR mistakes.

Together with OCR, the limited coverage of knowledge bases such as Wikidata tends to have an impact on historical NE processing, and especially on linking. In our corpus, entities that cannot be linked to a Wikidata entry (NIL entities) constitute 30% of the total. Interestingly, and contrary to our initial assumption, NIL entities are uniformly distributed across time periods (see Fig 2). The NIL ratio is higher for `Person`, `Media` and `Organisation` entities, whereas for geographic places (`Location`) Wikidata shows a substantial coverage (see Table 3). `Date` mentions were not linked as per HIPE annotation guidelines.

Corpus release. Data sets were released in IOB format with hierarchical information, in a similar fashion to CoNLL-U[10], and consist of UTF-8, tab-separated-values files containing the necessary information for all tasks (NERC-Coarse, NERC-Fine, and EL) [13].

Given the noisy quality of the material at hand, we chose not to apply sentence splitting nor sophisticated tokenization but, instead, to provide all nec-

[10] `https://universaldependencies.org/format.html`

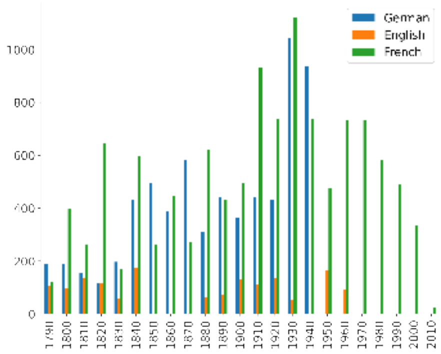


Fig. 1: Diachronic distribution of mentions across languages.

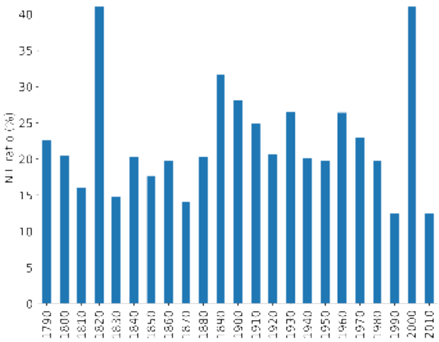


Fig. 2: Diachronic ratio of NIL entities.

essary information to rebuild the OCR text. Alongside each article, metadata (journal, date, title, page number, image region coordinates) and IIIF links to original page images are additionally provided when available.

The HIPE corpus, comprising several versions of each data set for the 3 languages, is released under a CC BY-NC 4.0 license[11] and is available on Zenodo[12] as well as on the HIPE GitHub repository[13].

3.2 Auxiliary Resources

In order to support participants in their system design and experiments, we provided auxiliary resources in the form of 'in-domain' word and character-level embeddings acquired from the same *impresso* newspapers titles and time periods from which HIPE training and development sets were extracted. Those embeddings correspond to fastText word embeddings [3] and flair contextualized string embeddings [1], both for French, German and English.

More specifically, fastText embeddings came in two versions, with subword 3-6 character n-grams and without, and were computed after a basic pre-processing (i.e., lower-casing, replacement of digits by 0 and deletion of all tokens/punctuation of length 1) that also tried to imitate the tokenization of the shared task data. Flair character embeddings were computed using flair 0.4.5[14] with a context of 250 characters, a batch size of 400-600 (depending on the GPU's memory), 1 hidden layer (size 2048), and a dropout of 0.1. Input was normalized with lower-casing, replacement of digits by 0, and of newlines by spaces; everything else was kept as in the original text (e.g. tokens of length 1). It is to be noted that the amount of training material greatly differed between languages (20G for French

[11] `https://creativecommons.org/licenses/by-nc/4.0/legalcode`

[12] `https://zenodo.org/deposit/3706857`

[13] `https://github.com/impresso/CLEF-HIPE-2020/tree/master/data`

[14] `https://github.com/flairNLP/flair`

and 8.5G for German taken from Swiss and Luxembourgish newspapers; 1.1G for English taken from Chronicling America material).

These embeddings are released under a CC BY-SA 4.0 license[15] and are available for download.[16] Contextualized character embeddings were also integrated into the flair framework[17].

4 Evaluation Framework

4.1 Evaluation Measures

NERC and EL tasks are evaluated in terms of Precision, Recall and F-measure (F1) [30]. Evaluation is done at entity level according to two metrics: micro average, with the consideration of all TP, FP, and FN[18] over all documents, and macro average, with the average of document's micro figures. Our definition of macro differs from the usual one: averaging is done at document-level and not across entity-types, and allows to account for (historical) variance in document length and entity distribution within documents instead of overall class imbalances.

Both NERC and EL benefit from strict and fuzzy evaluation regimes. For NERC (Coarse and Fine), the strict regime corresponds to exact boundary matching and the fuzzy to 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 punishes harshly, we keep this metric to be in line 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 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, and is 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.

[15] https://creativecommons.org/licenses/by-sa/4.0/legalcode

[16] https://files.ifi.uzh.ch/cl/siclemat/impresso/clef-hipe-2020/flair/

[17] https://github.com/flairNLP/flair

[18] True positive, False positive, False negative.

The HIPE scorer was provided to the participants early on, and the full evaluation toolkit (including all recipes and resources to replicate the present evaluation) is published under MIT license[19].

4.2 Task Bundles

In order to allow the greatest flexibility to participating teams as to which tasks to compete for while keeping a manageable evaluation frame, we introduced a system of task bundles offering different task combinations (see Table 4). Teams were allowed to choose only one bundle per language and to submit up to 3 runs per language. Only Bundle 5 (EL only) could be selected in addition to another one; this exception was motivated by the intrinsic difference between end-to-end linking and linking of already extracted entity mentions. Detailed information on system submission can be found in the HIPE Participation Guidelines [13].

Bundle	**Tasks**	**# teams**	**# runs**
1	NERC coarse, NERC fine and EL	2	10
2	NERC coarse and EL	3	10
3	NERC coarse and NERC fine	1	8
4	NERC coarse	7	27
5	EL only	5	20

Table 4: Task bundles.

5 System Descriptions

In this first HIPE edition, 13 participating teams submitted a total of 75 system runs. All teams participated to NERC-Coarse, 3 to NERC-Fine, and 5 to end-to-end EL and EL only. The distribution of runs per language reflects the data, with 35 runs for French (42%), 26 for German (31%), and 22 for English (26%). Besides, six teams worked on all 3 languages. For NERC, all but 2 teams applied neural approaches, and most of them also worked with contextualized embeddings.

5.1 Baselines

As a baseline for NERC-Coarse, we trained a traditional CRF sequence classifier [37] using basic spelling features such as a token's character prefix and suffix, the casing of the initial character, and whether it is a digit. The model, released to participating teams as part of the HIPE scorer, dismisses the segmentation

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

structure and treats any document as a single, long sentence. No baseline is provided for the NERC-Fine sub-task.

The baseline for entity linking (end-to-end EL and EL only) corresponds to AIDA-light [35], which implements the collective mapping algorithm by [19]. The wikimapper[20] tool was used to map Wikipedia URLs onto Wikidata QIDs, and the end-to-end EL baseline run relied on the CRF-based NERC baseline. Given the multilingual nature of the HIPE shared task, it is worth noting that AIDA-light was trained on a 2014 dump of the English Wikipedia, therefore accounting for a generous baseline.

5.2 Participating Systems

The following system descriptions are compiled from information provided by the participants. More accurate implementation details are available in the participants' system papers [6].

Cisteria, a collaboration of the *Ludwig-Maximilians Universität* and the *Bayerische Staatsbibliothek München* from Germany, focused on NERC-coarse for German. They experimented with external and HIPE character and word embeddings as well as several transformer-based BERT-style language models (e.g., German Europeana BERT[21]), all integrated by the neural flair NER tagging framework [1]. Interestingly, they trained different models with different embeddings for literal and metonymic NERC. No additional training material was used.

Ehrmama, affiliated with the University of Amsterdam, tackled coarse and fine-graind NERC for all languages. They build on the LSTM-CRF architecture of [27] and introduce a multi-task approach by splitting the top layers for each entity type. Their general embedding layer combines a multitude of embeddings, on the level of characters, sub-words and words; some newly trained by the team, as well as pre-trained BERT and HIPE's in-domain fastText embeddings. No additional training material was applied.

Ertim, affiliated with *Inalco*, Paris, applied their legacy (2010-13) NER system mXS[22] [36] for contemporary texts on the historical French HIPE data without any adaptation or training. The system uses pattern mining and non-neural machine learning for NERC and their model is based on the QUAERO standard [45], which is the basis for the HIPE annotation guidelines. For EL, only the type `Person` was considered. The resolution is done in two steps, first an approximate string match retrieves French Wikipedia pages, second the Wikidata item is selected whose Wikipedia article has the highest cosine similarity with the HIPE newspaper article containing the mention.

Inria, by the *ALMAnaCH* project team affiliated at *Inria*, Paris, used DeLFT (Deep Learning Framework for Text)[23] for NERC tagging of English and French.

[20] `https://github.com/jcklie/wikimapper`

[21] `https://huggingface.co/dbmdz`

[22] `https://github.com/eldams/mXS`

[23] `https://github.com/kermitt2/delft`

For English, the pre-trained Ontonotes 5.0 CoNLL-2012 model was used with a BiLSTM-CRF architecture. For EL, the system entity-fishing[24] was used.

Irisa, by a team from *IRISA*, Rennes, France, focused on French NERC and EL. For NERC, they improved the non-neural CRF baseline system with additional features such as context tokens, date regex match, ASCII normalization of the focus token, and the 100 most similar words from the HIPE fastText word embeddings provided by the organizers. For EL, a knowledge-base driven approach was applied to disambiguate and link the mentions of their NERC systems and the gold oracle NERC mentions [15]. Their experiments with the HIPE data revealed that collective entity linking is also beneficial for this type of texts—in contrast to linking mentions separately.

L3i, affiliated with *La Rochelle University*, France, tackled all prediction tasks of HIPE for all languages and achieved almost everywhere the best results. They used a hierarchical transformer-based model [51] built upon BERT [9] in a multi-task learning setting. On top of the pre-trained BERT blocks (German Europeana BERT, French CamemBERT, Multilingual BERT), several transformer layers were added to alleviate data sparsity issues, out-of-vocabulary words, spelling variations, or OCR errors in the HIPE dataset. A CRF was added on top to model the context dependencies between entity tags. An important preprocessing step for NERC was sentence segmentation and the reconstruction of words with hyphenation. For their EL approach, which is based on [24], the team built a Wikipedia/Wikidata knowledge base per language and trained entity embeddings for the most frequent entries [17]. Based on Wikipedia co-occurrence counts, a probabilistic mapping table was computed for linking mentions with entities—taking several mention variations (e.g. lowercase, Levenshtein distance) into account to improve the matching. The candidates were filtered using DBpedia and Wikidata by prioritizing those that corresponded to the named entity type. For persons, they analysed the date of birth to discard anachronistic entities. Finally, the five best matching candidates were predicted.

Limsi, affiliated with *LIMSI, CNRS*, Paris, France, focused on coarse NERC for French and achieved second best results there. They submitted runs from 3 model variations: a) A model based on CamemBERT [31] that jointly predicts the literal and metonymic entities by feeding into two different softmax layers. This model performed best on the dev set for metonymic entities. b) The model (a) with a CRF layer on top, which achieved their best results on literal tags (F1=.814 strict). c) A standard CamemBERT model that predicts concatenated literal and metonymic labels directly as a combined tag (resulting in a larger prediction tagset). This model performed best (within Limsi's runs) on the test set for metonymic entities (F1=.667 strict).

Nlp-uqam, affiliated with *Université du Quebec*, Montréal, Canada, focused on coarse NERC for French. Their architecture involves a BiLSTM layer for word-level feature extraction with a CRF layer on top for capturing label dependencies [27], and an attention layer in between for relating different positions of a sequence [51]. For their rich word representation, they integrate a character-

[24] `https://github.com/kermitt2/entity-fishing`

based CNN approach [8] and contextualized character-based flair embeddings [2] as provided by the HIPE organizers.

Sbb, affiliated with the Berlin State Library, Berlin, focused on coarse NERC and EL for all languages. For NERC, they applied a model based on multilingual BERT embeddings, which were additionally pre-trained on OCRed historical German documents from the SBB collection and subsequently fine-tuned on various multilingual NER data sets [26]. For EL, they constructed a multilingual knowledge base from Wikipedia (WP) articles roughly resembling the categories `Person`, `Location`, and `Organization`. The title words of these pages were embedded by BERT and stored in a nearest neighbor lookup index. A lookup applied to a mention returns a set of linked entity candidates. The historical text segment containing *the mention* and sentences from WP containing *a candidate* are then scored by a BERT sentence comparison model. This model was trained to predict for arbitrary WP sentence pairs whether they talk about the same entity or not. A random forest classifier finally ranks the candidates based on their BERT sentence comparison scores.

SinNer, affiliated with *INRIA* and Paris-Sorbonne University, Paris, France, focused on coarse literal NERC for French and German. They provided 2 runs based on a BiLSTM-CRF architecture, which combines fastText [3] and contextualized ELMo [40] embeddings[25]. For run 2, which performs better than their run 1 and is the one reported here, they applied propagation of entities at the document level. They optimized hyperparameters by training each variant three times and by selecting on F-score performance on the dev set. For run 3, they retrained SEM[26] with the official HIPE data sets and applied entity propagation. For German, they augmented SEM's gazetteers with location lexicons crawled from Wikipedia. The considerably lower performance of run 3 illustrates the advantage of embedding-based neural NER tagging.

Upb, affiliated with the *Politehnica University of Bucharest*, Bucarest, Bulgaria, focused on coarse literal NERC for all languages. Their BERT-based model centers around the ideas of transfer and multi-task learning as well as multilingual word embeddings. Their best performing runs combine multilingual BERT embeddings with a BiLSTM layer followed by a dense layer with local SoftMax predictions or alternatively, by adding a CRF layer on top of the BiLSTM.

Uva-ilps, affiliated with the University of Amsterdam, The Netherlands, focused on coarse NERC and end-to-end EL for all languages, and EL-only on English. They fine-tuned BERT models for token-level NERC prediction. Their EL approach was implemented by searching for each entity mention in the English Wikidata dump indexed by ElasticSearch[27]. The main problem there was the lack of German and French entities, although person names still could be found. For run 1 and 2 of EL only on English, they improved the candidate entity ranking by calculating cosine similarities between the contextual embeddings of a sentence containing the target entity mention and a modified sentence where

[25] [38] for French, [32] for German.

[26] SEM [10] is a CRF-based tool using Wapiti [28] as its linear CRF implementation.

[27] `https://www.elastic.co/`

the mention was replaced with a candidate entity description from Wikidata. The semantic similarity scores were multiplied by relative Levenshtein similarity scores between target mention and candidate labels to prefer precise character-level matches. Run 2 added historical spelling variations, however, this resulted in more false positives. Run 3 used REL [21], a completely different neural NERC and EL system. Candidate selection in REL is twofold, 4 candidates are selected by a probabilistic model predicting entities given a mention, and 3 candidates are proposed by a model predicting entities given the context of the mention. Candidate disambiguation combines local compatibility (prior importance, contextual similarity) and global coherence with other document-level entity linking decisions. Their REL-based run 3 outperformed their runs 1 and 2 clearly.

Webis, by the *Webis* group affiliated with the *Bauhaus University Weimar*, Germany, focused on coarse NERC for all languages. For each language, they trained a flair NERC sequence tagger [1] with a CRF layer using a stack of 4 embeddings: Glove embeddings [39], contextual character-based flair embeddings, and the forward and backward HIPE character-based flair embeddings. Their pre-processing included sentence reconstruction (by splitting the token sequence on all periods, except after titles, month abbreviations or numbers), and dehyphenation of tokens at the end of lines. For German, they experimented with data augmentation techniques by duplicating training set sentences and replacing the contained entities by randomly chosen new entities of the same type retrieved from Wikidata. A post-processing step resolved IOB tag sequence inconsistencies and applied a pattern-based tagging for time expressions. Although internal dev set validation F-scores looked promising, their official results on the test set had a bias towards precision. This could be due to format conversion issues.

6 Results and Discussion

We report results for the best run of each team and consider micro Precision, Recall and F1 exclusively. Results for NERC-Coarse and NERC-Fine for the three languages, both evaluation regimes and the literal and metonymic senses are presented in Table 5 and 6 respectively, while results for nested entities and entity components are presented in Table 7. Table 8 reports performances for end-to-end EL and EL only, with a cut-off @1. We refer the reader to the HIPE 2020 website[28] for more detailed results, and to the extended HIPE overview for a more in-depth discussion [12].

General observations. Neural systems with strong embedding resources clearly prevailed in HIPE NERC, beating symbolic CRF or pattern-matching based approaches by a large margin (e.g., compare baseline performance in Table 5). However, we also notice performance differences between neural systems that rely on BiLSTMs or BERT, the latter generally performing better.

[28] `https://impresso.github.io/CLEF-HIPE-2020/`

	French						German						English					
(a) Literal	Strict			Fuzzy			Strict			Fuzzy			Strict			Fuzzy		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Cisteria	-	-	-	-	-	-	.745	.578	.651	**.880**	.683	.769	-	-	-	-	-	-
Ehrmama	.793	.764	.778	.893	.861	.877	.697	.659	.678	.814	.765	.789	.249	.439	.318	.405	.633	.494
Ertim	.435	.248	.316	.604	.344	.439	-	-	-	-	-	-	-	-	-	-	-	-
Inria	.605	.675	.638	.755	.842	.796	-	-	-	-	-	-	.461	.606	.524	.568	.746	.645
Irisa	.705	.634	.668	.828	.744	.784	-	-	-	-	-	-	-	-	-	-	-	-
L3i	**.831**	**.849**	**.840**	**.912**	**.931**	**.921**	**.790**	**.805**	**.797**	.870	**.886**	**.878**	**.623**	**.641**	**.632**	.794	**.817**	**.806**
Limsi	.799	.829	.814	.887	.909	.898	-	-	-	-	-	-	-	-	-	-	-	-
Nlp-uqam	.705	.634	.668	.828	.744	.784	-	-	-	-	-	-	-	-	-	-	-	-
Sbb	.530	.477	.502	.765	.689	.725	.499	.484	.491	.730	.708	.719	.347	.310	.327	.642	.572	.605
SinNer	.788	.802	.795	.886	.902	.894	.658	.658	.658	.775	.819	.796	-	-	-	-	-	-
Upb	.693	.686	.689	.825	.817	.821	.677	.575	.621	.788	.740	.763	.522	.416	.463	.743	.592	.659
Uva-ilps	.656	.719	.686	.794	.869	.830	.499	.556	.526	.689	.768	.726	.443	.508	.473	.635	.728	.678
Webis	.731	.228	.347	.876	.273	.416	.695	.337	.454	.833	.405	.545	.476	.067	.117	**.873**	.122	.215
Baseline	.693	.606	.646	.825	.721	.769	.643	.378	.476	.790	.464	.585	.531	.327	.405	.736	.454	.562
Median	.705	.680	.677	.828	.829	.808	.686	.576	.636	.801	.752	.766	.461	.439	.463	.642	.633	.645
(b) Meto.	Strict			Fuzzy			Strict			Fuzzy			Strict			Fuzzy		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Cisteria	-	-	-	-	-	-	.738	.500	.596	.787	.534	.636	-	-	-	-	-	-
Ehrmama	.697	.554	.617	.708	.562	.627	.696	.542	.610	.707	.551	.619	-	-	-	-	-	-
L3i	**.734**	**.839**	**.783**	**.734**	**.839**	**.783**	.571	**.712**	**.634**	.626	**.780**	**.694**	.667	**.080**	**.143**	**1.00**	**.120**	**.214**
Limsi	.647	.688	.667	.655	.696	.675	-	-	-	-	-	-	-	-	-	-	-	-
Nlp-uqam	.423	.420	.422	.468	.464	.466	-	-	-	-	-	-	-	-	-	-	-	-
Baseline	.541	.179	.268	.541	.179	.268	**.814**	.297	.435	**.814**	.297	.435	**1.00**	.040	.077	**1.00**	.040	.077
Median	.647	.554	.617	.655	.562	.627	-	-	-	-	-	-	-	-	-	-	-	-

Table 5: Results for NERC-Coarse (micro P, R and F-measure). Bold font indicates the highest, and underlined font the second-highest value.

	French						German					
(a) Literal	Strict			Fuzzy			Strict			Fuzzy		
	P	R	F	P	R	F	P	R	F	P	R	F
Ehrmama	.696	.724	.710	.776	.807	.791	**.650**	.592	.620	**.754**	.687	.719
Ertim	.418	.238	.303	.568	.324	.412	-	-	-	-	-	-
L3i	**.772**	**.797**	**.784**	**.843**	**.869**	**.856**	.628	**.712**	**.668**	.734	**.813**	**.771**
(b) Metonymic												
Ehrmama	.667	.554	.605	.667	.554	.605	**.707**	.551	.619	**.717**	.559	.629
L3i	**.718**	**.661**	**.688**	**.738**	**.679**	**.707**	.601	**.703**	**.648**	.659	**.771**	**.711**

Table 6: Results for NERC-Fine.

	French						German					
(a) Comp.	Strict			Fuzzy			Strict			Fuzzy		
	P	R	F	P	R	F	P	R	F	P	R	F
Ehrmama	**.695**	.632	**.657**	**.801**	.707	.751	**.681**	.494	.573	**.735**	.534	.618
Ertim	.042	.045	.043	.074	.080	.077	-	-	-	-	-	-
L3i	.680	**.732**	**.657**	.773	**.832**	**.801**	.595	**.698**	**.642**	.654	**.768**	**.707**
(b) Nested												
Ehrmama	**.397**	.280	.329	**.448**	.317	.371	-	-	-	-	-	-
L3i	.337	**.402**	**.367**	.357	**.427**	**.389**	.471	.562	.513	.517	.616	.562

Table 7: Results for nested entities and entity components.

In general and not unexpectedly, we observe that the amount of available training and development data correlates with system performances. French with the largest amount of training data has better results than German, and English is worse than German (see median numbers in Table 5). The one exception is EL only where English, as a well-resourced language, seems to have the necessary tooling to also excel on non-standard, historical text material (cf. Inria results). NERC-Coarse performances show a great diversity but top results are better than expected, specifically for French where they are almost on a par with performances on contemporary texts. Here, six teams have fuzzy F1 scores higher than .8, suggesting good prospects for entity extraction systems on historical texts, when trained with appropriate and sufficient data. Fine-grained NERC with more than 12 classes is obviously more difficult than predicting only 5 categories. However, the performance drop of the best performing system L3i is relatively mild for French, 6.5 percentage points on fuzzy F1, and a little stronger for German (10.7). Finally, the recognition of entity components shows reasonable performances and suggests that knowledge base population and/or biography reconstruction from historical texts is feasible. The same cannot be said of nested entities.

System-based observations. With L3i, the HIPE 2020 campaign has a clear overall winner on NERC coarse and fine, literal and metonymic entities, components as well as EL. The one exception is EL only for English, where Inria's entity-fishing system outperforms L3i. L3i is particularly convincing in terms of F1, as it consistently keeps precision and recall in good balance (even trending toward recall many times). Other systems, e.g. Inria, Ehrmama, or the baseline, typically suffer from a bias towards precision. We assume that actively tackling the problem of OCR noise and hyphenation issues helps to achieve better recall.

Time-based observations. In order to gauge the impact of the article's publication date on system performances, we analyze the variation of F1 scores as a function of time (see Fig. 3). The initial hypothesis here was that the older the article, the more difficult it is to extract and link the mentions it contains. In general, there does not seem to be a strong correlation between the article's publication date and F1 scores. In the specific case of EL, this finding is in line with the uniform distribution of NIL entities across time (see Section 3).

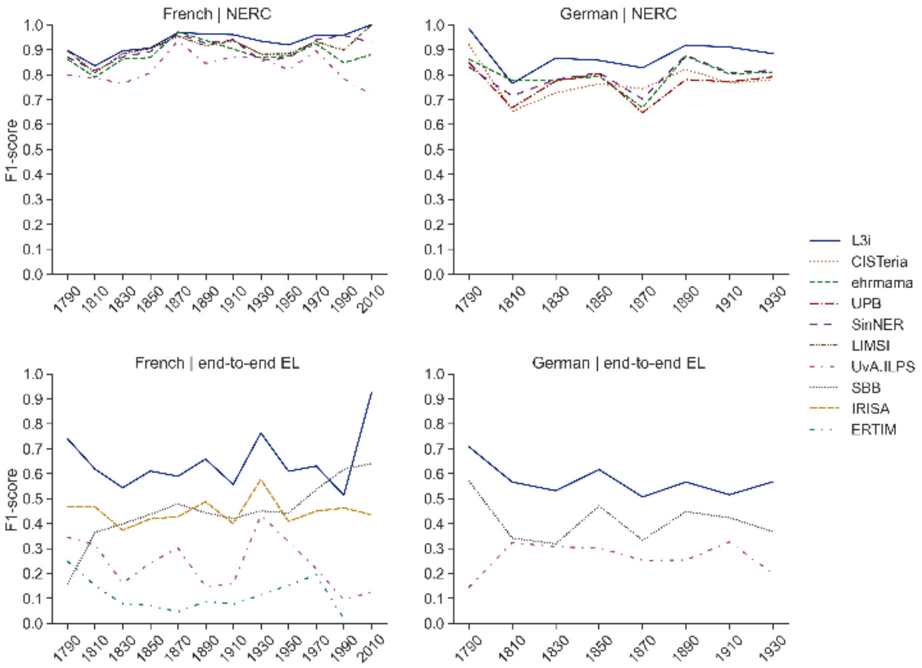
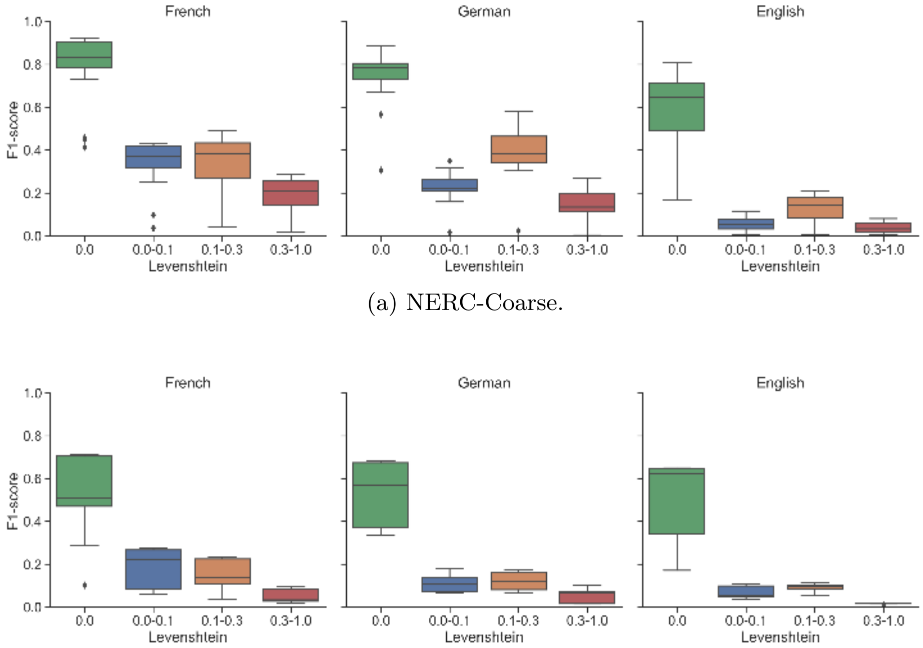


Fig. 3: F1-score as a function of time for the 5 best systems for NERC (top) and end-to-end EL (bottom) for the languages French (left) and German (right). The x-axis shows 20-years time buckets (e.g. 1790 = 1790-1809).

Impact of OCR noise. To assess the impact of noisy entities on the task of NERC and EL, we evaluated the system performances on various noise levels (see Fig. 4). The level of noise is defined as the length-normalized Levenshtein distance between the surface form of an entity and its human transcription. There is a remarkable difference between the performances for noisy and non-noisy mentions on both NERC and EL. Already as little noise as 0.1 severely hurts the system's ability to predict an entity and may cut its performance by half. Interestingly, EL also suffers badly from little noise (norm. lev. dist. > 0.0 and < 0.1) even when providing the gold annotations of NERC (EL only, not shown in the plot). Slightly and medium noisy mentions (norm. lev. dist.



(a) NERC-Coarse.

(b) End-to-end EL with the relaxed evaluation regime and a cutoff @3.

Fig. 4: Impact of OCR noise: distribution of performances across systems on entities with different noise level severity for NERC (a) and end-to-end EL (b).

> 0.0 and < 0.3) show a similar impact, while for highly noisy mentions, the performance deteriorates further. We can observe the greatest variation between systems at the medium noise level suggesting that the most robust systems get their competitive advantage when dealing with medium noisiness. On the effect of OCR noise on NERC, [49] claim that OCR errors impact more GPE mentions than persons or dates; in our breakdown of OCR noise impact by type, we can confirm that claim for little noise only (norm. lev. dist. > 0.0 and < 0.1), while this trend turns into the opposite for highly noisy entities.

7 Conclusion and Perspectives

From the perspective of natural language processing, the HIPE evaluation lab provided the opportunity to test the robustness of NERC and EL approaches against challenging historical material and to gain new insights with respect

End-to-end EL	**French**						**German**						**English**					
(a) Literal	Strict			Fuzzy			Strict			Fuzzy			Strict			Fuzzy		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
ERTIM	.150	.084	.108	.150	.084	.108	-	-	-	-	-	-	-	-	-	-	-	-
IRISA	.446	.399	.421	.465	.417	.439	-	-	-	-	-	-	-	-	-	-	-	-
L3I	**.594**	**.602**	**.598**	.613	**.622**	**.617**	.531	**.538**	**.534**	.553	**.561**	**.557**	**.523**	**.539**	**.531**	**.523**	**.539**	**.531**
SBB	**.594**	.310	.407	**.616**	.321	.422	**.540**	.304	.389	**.561**	.315	.403	.257	.097	.141	.257	.097	.141
UVA-ILPS	.352	.195	.251	.353	.196	.252	.245	.272	.258	.255	.283	.268	.249	.375	.300	.249	.375	.300
Baseline	.206	.342	.257	.257	.358	.270	.173	.187	.180	.188	.203	.195	.220	.263	.239	.220	.263	.239
(b) Meton.																		
IRISA	.023	.295	.043	.041	.527	.076	-	-	-	-	-	-	-	-	-	-	-	-
L3I	**.236**	**.402**	**.297**	**.366**	**.625**	**.462**	.324	.508	.396	.384	.602	.469	.172	.200	.185	.172	.200	.185
Baseline	.002	.027	.004	.008	.098	.015	.025	.136	.042	.026	.144	.044	.004	.040	.007	.004	.040	.007

EL only	**French**						**German**						**English**					
	Strict			Fuzzy			Strict			Fuzzy			Strict			Fuzzy		
(a) Literal	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
INRIA	.585	**.650**	.616	.604	**.670**	.635	-	-	-	-	-	-	**.633**	**.685**	**.658**	**.633**	**.685**	**.658**
IRISA	.475	.473	.474	.492	.491	.492	-	-	-	-	-	-	-	-	-	-	-	-
L3I	.640	.638	**.639**	.660	.657	**.659**	.581	**.582**	**.582**	.601	**.602**	**.602**	.593	.593	.593	.593	.593	.593
SBB	**.677**	.371	.480	**.699**	.383	.495	**.615**	.349	.445	**.636**	.361	.461	.344	.119	.177	.344	.119	.177
UVA.ILPS	-	-	-	-	-	-	-	-	-	-	-	-	.607	.580	.593	.607	.580	.593
Baseline	.502	.495	.498	.516	.508	.512	.420	.416	.418	.440	.435	.437	.506	.506	.506	.506	.506	.506
(b) Meto.																		
IRISA	.025	.357	.047	.041	.580	.076	-	-	-	-	-	-	-	-	-	-	-	-
L3I	**.303**	**.446**	**.361**	**.461**	**.679**	**.549**	**.443**	**.627**	**.519**	**.515**	**.729**	**.604**	**.286**	.480	**.358**	**.286**	.480	**.358**
UVA.ILPS	-	-	-	-	-	-	-	-	-	-	-	-	.031	.058	.031	.031	.058	.031
Baseline	.213	.312	.254	.323	.473	.384	.265	.373	.310	.331	.466	.387	.219	.280	.246	.219	.280	.246

Table 8: Results for end-to-end EL (top) and EL only (bottom) with P, R and F1 @1.

to domain and language adaptation. With regard to NERC, results show that it is possible to design systems capable of dealing with historical and noisy inputs, whose performances compete with those obtained on contemporary texts. Entity linking, as well as the processing of metonymy and nested entities remain challenging aspects of historical NE processing (the latter two probably due to the limited amount of annotated material).

From the perspective of digital humanities, the lab's outcomes will help DH practitioners in mapping state-of-the-art solutions for NE processing on 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 digitized 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.

Overall, HIPE has contributed to advance the state of the art in semantic indexing of historical newspapers and, more generally, of historical material. As future work, we intend to explore the several directions for a potential second edition of HIPE: expanding the language spectrum, strengthening the already covered languages by providing more training data, considering other types of historical documents, and exploring to what extent the improvements shown in HIPE can be transferred to similar tasks in other domains, or to linking problems that require knowledge bases other than Wikidata.

Acknowledgements

This HIPE evaluation lab would not have been possible without the interest and commitment of many. We express our warmest thanks to: the Swiss newspapers *NZZ* and *Le Temps*, and the Swiss and Luxembourg national libraries for sharing part of their data in the frame of the *impresso* project; Camille Watter and Gerold Schneider for their commitment and hard work with the construction of the data set; the INCEpTION project team for its valuable and efficient support with the annotation tool; Richard Eckart de Castillo, Clemens Neudecker, Sophie Rosset and David Smith for their encouragement and guidance as part of the HIPE advisory board; and, finally, the 13 teams who embarked in this first HIPE edition, for their patience and scientific involvement. HIPE is part of the research activities of the project "*impresso* – Media Monitoring of the Past", for which we also gratefully acknowledge the financial support of the Swiss National Science Foundation under grant number CR-SII5_173719.

Bibliography

[1] Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics (demonstrations). pp. 54–59. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://www.aclweb.org/anthology/N19-4010

[2] Akbik, A., Blythe, D., Vollgraf, R.: Contextual String Embeddings for Sequence Labeling. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1638–1649. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), http://www.aclweb.org/anthology/C18-1139

[3] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017), https://www.aclweb.org/anthology/Q17-1010

[4] Bollmann, M.: A Large-Scale Comparison of Historical Text Normalization Systems. 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. 3885–3898. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1389

[5] Borin, L., Kokkinakis, D., Olsson, L.J.: Naming the past: Named entity and animacy recognition in 19th century Swedish literature. In: Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007). pp. 1–8 (2007)

[6] Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.): CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings (2000)

[7] Chiron, G., Doucet, A., Coustaty, M., Visani, M., Moreux, J.P.: Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information. In: Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries. pp. 249–252. JCDL '17, IEEE Press, Piscataway, NJ, USA (2017), http://dl.acm.org/citation.cfm?id=3200334.3200364

[8] Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics **4**, 357–370 (2016). https://doi.org/10.1162/tacl_a_00104

[9] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805

[10] Dupont, Y., Dinarelli, M., Tellier, I., Lautier, C.: Structured Named Entity Recognition by Cascading CRFs. In: Intelligent Text Processing and Computational Linguistics (CICling) (2017)

[11] 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 (2016), https://infoscience.epfl.ch/record/221391?ln=en

[12] 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.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)

[13] Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: HIPE - Shared Task Participation Guidelines (v1.1) (2020). https://doi.org/10.5281/zenodo.3677171

[14] Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Impresso Named Entity Annotation Guidelines (Jan 2020). https://doi.org/10.5281/zenodo.3604227

[15] El Vaigh, C.B., Goasdoué, F., Gravier, G., Sébillot, P.: Using Knowledge Base Semantics in Context-Aware Entity Linking. In: Proceedings of the ACM Symposium on Document Engineering 2019. pp. 1–10. DocEng '19, Association for Computing Machinery, Berlin, Germany (Sep 2019). https://doi.org/10.1145/3342558.3345393

[16] Galibert, O., Rosset, S., Grouin, C., Zweigenbaum, P., Quintard, L.: Extended Named Entity Annotation on OCRed Documents : From Corpus Constitution to Evaluation Campaign. In: Proceedings of the Eighth conference on International Language Resources and Evaluation. pp. 3126–3131. Istanbul, Turkey (2012)

[17] Ganea, O.E., Hofmann, T.: Deep Joint Entity Disambiguation with Local Neural Attention. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2619–2629 (2017)

[18] Grishman, R., Sundheim, B.: Design of the MUC-6 evaluation. In: Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland (1995)

[19] Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: EMNLP (2011)

[20] Hooland, S.V., Wilde, M.D., Verborgh, R., Steiner, T., Walle, R.V.d.: Exploring entity recognition and disambiguation for cultural heritage collections. Digital Scholarship in the Humanities **30**(2), 262–279 (2015). https://doi.org/10.1093/llc/fqt067

[21] van Hulst, J.M., Hasibi, F., Dercksen, K., Balog, K., de Vries, A.P.: REL: An entity linker standing on the shoulders of giants. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20, ACM (2020)

[22] Kaplan, F., di Lenardo, I.: Big Data of the Past. Frontiers in Digital Humanities **4** (2017). https://doi.org/10.3389/fdigh.2017.00012

[23] Klie, J.C., Bugert, M., Boullosa, B., de Castilho, R.E., Gurevych, I.: The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. pp. 5–9 (2018)

[24] Kolitsas, N., Ganea, O.E., Hofmann, T.: End-to-End Neural Entity Linking. In: Proceedings of the 22nd Conference on Computational Natural Language Learning. pp. 519–529. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/K18-1050

[25] Krippendorff, K.: Content analysis: An introduction to its methodology. Sage publications (1980)

[26] Labusch, K., Neudecker, C., Zellhöfer, D.: BERT for Named Entity Recognition in Contemporary and Historic German. In: Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers. pp. 1–9. German Society for Computational Linguistics & Language Technology, Erlangen, Germany (2019)

[27] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural Architectures for Named Entity Recognition. arXiv:1603.01360 [cs] (Mar 2016), `http://arxiv.org/abs/1603.01360`

[28] Lavergne, T., Cappé, O., Yvon, F.: Practical very large scale CRFs. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 504–513. Association for Computational Linguistics (2010)

[29] Linhares Pontes, E., Hamdi, A., Sidere, N., Doucet, A.: Impact of OCR Quality on Named Entity Linking. In: Jatowt, A., Maeda, A., Syn, S.Y. (eds.) Digital Libraries at the Crossroads of Digital Information for the Future. pp. 102–115. Lecture Notes in Computer Science, Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-34058-2_11

[30] 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)

[31] Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., Éric Villemonte de la Clergerie, Seddah, D., Sagot, B.: Camembert: a tasty french language model (2019)

[32] May, P.: German ELMo Model (2019), `https://github.com/t-systems-on-site-services-gmbh/german-elmo-model`

[33] Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)

[34] Neudecker, C., Antonacopoulos, A.: Making Europe's Historical Newspapers Searchable. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS). pp. 405–410. IEEE, Santorini, Greece (Apr 2016). https://doi.org/10.1109/DAS.2016.83

[35] Nguyen, D.B., Hoffart, J., Theobald, M., Weikum, G.: Aida-light: High-throughput named-entity disambiguation. In: LDOW (2014)

[36] Nouvel, D., Antoine, J.Y., Friburger, N.: Pattern Mining for Named Entity Recognition. LNCS/LNAI Series **8387i (post-proceedings LTC 2011)** (2014)

[37] Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007), http://www.chokkan.org/software/crfsuite/

[38] Ortiz Suárez, P.J., Dupont, Y., Muller, B., Romary, L., Sagot, B.: Establishing a new state-of-the-art for French named entity recognition. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 4631–4638. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.569

[39] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. vol. 14, pp. 1532–43 (2014)

[40] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep Contextualized Word Representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://doi.org/10.18653/v1/N18-1202

[41] Piotrowski, M.: Natural language processing for historical texts. Synthesis Lectures on Human Language Technologies **5**(2), 1–157 (2012)

[42] Plank, B.: What to do about non-standard (or non-canonical) language in NLP. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)). Bochumer Linguistische Arbeitsberichte (2016)

[43] Rao, D., McNamee, P., Dredze, M.: Entity linking: Finding extracted entities in a knowledge base. In: Multi-source, multilingual information extraction and summarization, pp. 93–115. Springer (2013)

[44] Rosset, S., Grouin, C., Fort, K., Galibert, O., Kahn, J., Zweigenbaum, P.: Structured named entities in two distinct press corpora: Contemporary broadcast news and old newspapers. In: Proceedings of the 6th Linguistic Annotation Workshop. pp. 40–48. Association for Computational Linguistics (2012)

[45] Rosset, Sophie, Grouin, Cyril, Zweigenbaum, Pierre: Entités nommées structurées : guide d'annotation Quaero. NOTES et DOCUMENTS 2011-04, LIMSI-CNRS (2011)

[46] Smith, D.A., Cordell, R.: A Research Agenda for Historical and Multilingual Optical Character Recognition. Tech. rep. (2018), http://hdl.handle.net/2047/D20297452

[47] Sporleder, C.: Natural Language Processing for Cultural Heritage Domains. Language and Linguistics Compass **4**(9), 750–768 (2010). https://doi.org/10.1111/j.1749-818X.2010.00230.x

[48] van Strien, D., Beelen, K., Ardanuy, M.C., Hosseini, K., McGillivray, B., Colavizza, G.: Assessing the impact of OCR quality on downstream NLP tasks. In: ICAART 2020 - Proceedings of the 12th International Conference on Agents and Artificial Intelligence. SCITEPRESS - Science and Technology Publications (Jan 2020). https://doi.org/10.17863/CAM.52068

[49] van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B., Colavizza, G.: Assessing the Impact of OCR Quality on Downstream NLP Tasks. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence. pp. 484–496. SCITEPRESS

- Science and Technology Publications, Valletta, Malta (2020). https://doi.org/10.5220/0009169004840496

[50] Terras, M.: The Rise of Digitization. In: Rikowski, R. (ed.) Digitisation Perspectives, pp. 3–20. SensePublishers, Rotterdam (2011). https://doi.org/10.1007/978-94-6091-299-3_1

[51] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017), http://arxiv.org/abs/1706.03762

[52] Vilain, M., Su, J., Lubar, S.: Entity Extraction is a Boring Solved Problem: Or is It? In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers. pp. 181–184. NAACL-Short '07, Association for Computational Linguistics (2007), http://dl.acm.org/citation.cfm?id=1614108.1614154, event-place: Rochester, New York