Introducing the CLEF 2020 HIPE Shared Task: Named Entity Recognition and Linking on Historical Newspapers

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

[1] Digital Humanities Laboratory, EPFL, Switzerland
[2] Institute of Computational Linguistics, University of Zurich, Switzerland

Abstract. 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. If NE processing tools are increasingly being used in the context of historical documents, performance values are below the ones on contemporary data and are hardly comparable. In this context, this paper introduces the CLEF 2020 Evaluation Lab HIPE (Identifying Historical People, Places and other Entities) on named entity recognition and linking on diachronic historical newspaper material in French, German and English. Our objective 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comparison of NE processing on historical texts, and, in the long run, fostering efficient semantic indexing of historical documents in order to support scholarship on digital cultural heritage collections.

Keywords: named entity processing · text understanding · information extraction · historical 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 [11], named entity-related tasks have undergone major evolutions until now, from entity recognition and classification to entity disambiguation and linking [21, 25]. Besides the general domain of well-written newswire data, named entity (NE) processing is also applied to specific

The final authenticated version is available online at https://doi.org/10.1007/978-3-030-45442-5_68

domains, particularly bio-medical [14, 10], and on more noisy inputs such as speech transcriptions [9] and tweets [26].

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 [1, 17, 16]. 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 [31]. 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' [13]. In this regard, information extraction techniques, and particularly NE recognition and linking, can certainly be regarded as among the first steps.

This paper introduces the CLEF 2020 Evaluation Lab[3] HIPE (Identifying Historical People, Places and other Entities)[4]. With the aim of supporting the development and progress of NE systems on historical documents (Section 2), this lab proposes two tasks, namely named entity recognition and linking, on historical newspapers in French, German and English (Section 3). We additionally report first results on French historical newspapers (Section 4), which comfort the idea of various benefits of such lab for both NLP and DH communities.

2 Motivation and Objectives

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, as per language technologies in general [30], it appears that the application of NE processing on historical texts poses new challenges [7, 23]. 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 [27, 5]. 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) [3, 2]. Further, beside historical VIPs, texts from the past contain

[3] `https://clef2020.clef-initiative.eu/`

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

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 [12]. 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 [22].

Overall, and as demonstrated by Vilain et al. [32], 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 material. This echoes the proposition of Plank [24], 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.

Even though many evaluation campaigns on NE were organized over the last decades[5], only one considered French historical texts [8]. To the best of our knowledge, no NE evaluation campaign ever addressed multilingual, diachronic historical material. In the context of new needs and materials emerging from the humanities, we believe that an evaluation campaign on historical documents is timely and will be beneficial. In addition to the release of a multilingual, historical NE-annotated corpus, the objective of this shared task is threefold: strengthening the robustness of existing approaches on non-standard inputs; enabling performance comparison of NE processing on historical texts; and fostering efficient semantic indexing of historical documents.

3 Overview of the Evaluation Lab

3.1 Task Description

The HIPE shared task puts forward 2 NE processing tasks with sub-tasks of increasing level of difficulty. Participants can submit up to 3 runs per sub-task.

Task 1: Named Entity Recognition and Classification (NERC)

Subtask 1.1 - NERC coarse-grained: this task includes the recognition and classification of entity mentions according to high-level entity types (Person, Location, Organisation, Product and Date).

Subtask 1.2 - NERC fine-grained: this task includes the classification of mentions according to finer-grained entity types, nested entities (up to one level of depth) and the detection of entity mention components (e.g. function, title, name).

Task 2: Named Entity Linking (EL) This task requires the linking of named entity mentions to a unique referent in a knowledge base (a frozen dump of Wikidata) or to a NIL node if the mention does not have a referent.

[5] MUC, ACE, CONNL, KBP, ESTER, HAREM, QUAERO, GERMEVAL, etc.

3.2 Data Sets

Corpus The HIPE corpus is composed of items from the digitized archives of several Swiss, Luxembourgish and American newspapers on a diachronic basis.[6] For each language, articles of 4 different newspapers were sampled on a decade time-bucket basis, according to the time span of the newspaper (longest duration spans ca. 200 years). More precisely, articles were first randomly sampled from each year of the considered decades, with the constraints of having a title and more than 100 characters. Subsequently to this sampling, a manual triage was applied in order to keep journalistic content only and to remove undesirable items such as feuilleton, cross-words, weather tables, time-schedules, obituaries, and what a human could not even read because of OCR noise.

Alongside each article, metadata (journal, date, title, page number, image region coordinates), the corresponding scan(s) and an OCR quality assessment score is provided. 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.

For each task and language—with the exception of English—the corpus is divided into training, dev and test data sets, released in IOB format with hierarchical information. For English, only dev and test sets will be released.

Annotation HIPE *annotation guidelines* [6] are derived from the Quaero annotation guide[7]. Originally designed for the annotation of "extended" named entities (i.e. more than the 3 or 4 traditional entity classes) in French speech transcriptions, Quaero guidelines have furthermore been used on historic press corpora [29]. HIPE slightly recasts and simplifies them, 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, where annotation decisions were tested and difficult cases discussed. Despite these adaptations, HIPE annotated corpora will mostly remain compatible with Quaero guidelines.

The *annotation campaign* is carried out by the task organizers with the support of trilingual collaborators. We use INCEpTION as an annotation tool [15], with the visualisation of image segments alongside OCR transcriptions.[8] Before starting annotating, each annotator is first trained on a mini-reference corpus, where the inter-annotator agreement (IAA) with the gold reference is computed. For each language, a sub-sample of the corpus is annotated by 2 annotators and IAA is computed, before and after an adjudication. Randomly selected articles

[6] From the Swiss National Library, the Luxembourgish National Library, and the Library of Congress, respectively.

[7] See the original Quaero guidelines: `http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf`

[8] HIPE is one of the official INCEpTION project use cases, see `https://inception-project.github.io/use-case-gallery/impresso/`

Type	Freq.	CRF P	CRF R	CRF F_1	Neural P	Neural R	Neural F_1
pers	3,638	75.6	71.0	73.2	84.6	84.7	84.6
func	3,133	69.7	43.7	53.7	76.5	66.8	71.3
loc	2,369	67.8	63.6	65.6	73.1	77.2	75.1
amount	1,716	63.4	53.8	58.2	75.0	71.0	72.9
time	1,675	70.4	59.3	64.4	73.4	71.3	72.3
org	1,412	63.9	44.4	52.4	67.6	52.4	59.1
prod	612	54.1	30.2	38.8	60.7	51.6	55.8
all	14,555	69.4	56.2	62.1	76.2	72.0	74.0

Token OCR	Token correct	IOB Gold	IOB CRF	IOB Neural
T«i*louse	Toulouse	B-loc	O	B-loc
Caa.Qrs	Cahors	B-loc	O	B-loc
o ˚ an	Jean	B-pers	O	B-pers
Chêne	Chêne	I-pers	B-pers	I-pers
©cmp0	Temps	B-prod	O	B-prod
f&tvxps	Temps	B-prod	O	B-prod
f&tïtipB	Temps	B-prod	O	B-prod

Table 1. Results and examples from exploratory experiments on NER for French.

will also be controlled by the adjudicator. Finally, HIPE will provide *complementary resources* in the form of in-domain word-level and character-level embeddings acquired from historical newspaper corpora. In the same vein, participants will be encouraged to share any external resource they might use. HIPE corpus and resources will be released under a CC-BY-SA-NC 4.0 license.

3.3 Evaluation

Named Entity Recognition and Classification (Task 1) will be evaluated in terms of macro and micro Precision, Recall, F-measure, and Slot Error Rate [20]. Two evaluation scenarios will be considered: strict (exact boundary matching) and relaxed (fuzzy boundary matching). Entity linking (Task 2) will be evaluated in terms of Precision, Recall, F-measure taking into account literal and metonymic senses.

4 Exploratory Experiments on NER for Historical French

We made an exploratory study in order to assess whether the massive improvements in neural NER [1, 17] on modern texts carry over to historical material with OCR noise. The data of our experiments is the *Quaero Old Press* (QOP) corpus, 295 OCRed[9] newspaper documents dating from December 1890 annotated according to the *Quaero* guidelines [28], split by us into train (1.45m tokens) and dev/test (each 0.2m). We only consider the outermost entity level (no nested entities or components) and train on the fine-grained subcategories (e.g., *loc.adm.town*) of the 7 main classes.

Modeling NER as a sequence labeling problem and applying Bi-LSTM networks is state of the art [1, 4, 17, 19]. Our experiments follow [1] in using character-based contextual string embeddings as input word representations, allowing to "better handle rare and misspelled words as well as model subword structures such as prefixes and endings". These contextualized word embeddings rely on neural forward and backward character-level language models that have

[9] [28] reports a character error rate of 5.09% and a word error rate of 36.59%.

been trained by us on a large collection (500m tokens) of late 19th and early 20th centuries Swiss-French newspapers. In accordance to the literature, a Bi-LSTM NER model with an on-top CRF layer (Bi-LSTM-CRF) works best for our data.

As a baseline system, which will also be provided for the shared task, we train a traditional CRF sequence classifier [18] using basic spelling features such as a token's character prefix and suffix, the casing of the initial character and the presence of punctuation marks and digits. The baseline classifier shows fairly modest overall performance scores of 69.4% recall, 56.2% precision and 62.1 F_1 (see Table 1).

Trained and evaluated on the QOP data, the neural model relying on contextual string embeddings clearly outperforms the baseline classifier. As shown in Table 1, the Bi-LSTM-CRF model achieves better F_1 for all of the 7 entity types and surpasses the feature-based classifier by nearly 12 points F_1. Examples in Table 1 evidence that the CRF model frequently struggles with entities containing miss-recognized special characters and/or punctuation marks. In many such cases, the Bi-LSTM-CRF classifier is capable of assigning the correct label. These results indicate that the new neural methods are ready to enable substantial progress in NER on noisy historical texts.

5 Conclusion

From the perspective of natural language processing (NLP), the HIPE evaluation lab provides the opportunity to test the robustness of existing NERC and EL approaches against challenging historical material and to gain new insights with respect to domain and language adaptation. From the perspective of digital humanities, the lab's outcomes 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 will contribute to advance the state of the art in semantic indexing of historical material, within the specific domain of historical newspaper processing, as in e.g. the '*impresso* - Media Monitoring of the Past' project[10] and, more generally, within the domain of text understanding of historical material, as in the Time Machine Europe project[11] which ambitions the application of AI technologies on cultural heritage data.

[10] `https://impresso-project.ch`
[11] `https://www.timemachine.eu`

Acknowledgements

This CLEF evaluation lab is part of the research activities of the project "*impresso* – Media Monitoring of the Past", for which authors gratefully acknowledge the financial support of the Swiss National Science Foundation under grant number CRSII5_173719. We would also like to thank C. Watter, G. Schneider and A. Flückiger for their invaluable help with the construction of the data sets, as well as R. Eckart de Castillo, C. Neudecker, S. Rosset and D. Smith for their support and guidance as part of the lab's advisory board.

Bibliography

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 1638–1649, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics.

[2] Marcel Bollmann. 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)*, pages 3885–3898, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Lars Borin, Dimitrios Kokkinakis, and Leif-Jöran Olsson. 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 (LaT-eCH 2007)*, pages 1–8, 2007.

[4] Jason P. C. Chiu and Eric Nichols. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics (TACL)*, 4:357–370, 2016.

[5] M. Dinarelli and S. Rosset. Tree-Structured Named Entity Recognition on OCR Data: Analysis, Processing and Results. In 2012, editor, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, Istanbul, Turkey, May 2012*. European Language Resources Association (ELRA), 2012. ISBN 978-2-9517408-7-7.

[6] Ehrmann, Watter, Romanello, Clematide, and Flückiger. *Impresso Named Entity Annotation Guidelines*, January 2020. URL `https://doi.org/10.5281/zenodo.3604227`.

[7] Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. Diachronic Evaluation of NER Systems on Old Newspapers. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016))*, pages 97–107. Bochumer Linguistische Arbeitsberichte, 2016. URL `https://infoscience.epfl.ch/record/221391?ln=en`.

[8] O. Galibert, S. Rosset, C. Grouin, P. Zweigenbaum, and L. Quintard. 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*, pages 3126–3131, Istanbul, Turkey, 2012.

[9] Olivier Galibert, Jeremy Leixa, Gilles Adda, Khalid Choukri, and Guillaume Gravier. The etape speech processing evaluation. In *LREC*, pages 3995–3999. Citeseer, 2014.

[10] Rodrigo Rafael Villarreal Goulart, Vera Lúcia Strube de Lima, and Clarissa Castellã Xavier. A systematic review of named entity recognition in biomedical texts. *Journal of the Brazilian Computer Society*, 17(2):103–116, June 2011. ISSN 1678-4804. https://doi.org/10.1007/s13173-011-0031-9. URL `https://doi.org/10.1007/s13173-011-0031-9`.

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

[12] S. Van Hooland, M. De Wilde, R. Verborgh, T. Steiner, and R. Van de Walle. Exploring entity recognition and disambiguation for cultural heritage collections.

Digital Scholarship in the Humanities, 30(2):262–279, 2015. ISSN 2055-7671. https://doi.org/10.1093/llc/fqt067.

[13] Frédéric Kaplan and Isabella di Lenardo. Big Data of the Past. *Frontiers in Digital Humanities*, 4, 2017. ISSN 2297-2668. https://doi.org/10.3389/fdigh.2017.00012. URL `https://www.frontiersin.org/articles/10.3389/fdigh.2017.00012/full`.

[14] J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182, 2003.

[15] Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, 2018.

[16] Kai Labusch, Clemens Neudecker, and David Zellhöfer. 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*, pages 1–9, Erlangen, Germany, 2019. German Society for Computational Linguistics & Language Technology.

[17] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June 2016. Association for Computational Linguistics.

[18] Thomas Lavergne, Olivier Cappé, and François Yvon. Practical very large scale CRFs. In Jan Hajič, Sandra Carberry, and Stephen Clark andJoakim Nivre, editors, *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[19] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany, August 2016. Association for Computational Linguistics.

[20] John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. Performance measures for information extraction. In *In Proceedings of DARPA Broadcast News Workshop*, pages 249–252, 1999.

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

[22] Clemens Neudecker and Apostolos Antonacopoulos. Making Europe's Historical Newspapers Searchable. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 405–410. IEEE, 2016.

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

[24] Barbara Plank. 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.

[25] Delip Rao, Paul McNamee, and Mark Dredze. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*, pages 93–115. Springer, 2013.

[26] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, 2011.

[27] K. J. Rodriquez, M. Bryant, T. Blanke, and M. Luszczynska. Comparison of named entity recognition tools for raw OCR text. In Jeremy Jancsary, editor, *Proceedings of KONVENS 2012*, pages 410–414. ÖGAI, September 2012. URL http://www.oegai.at/konvens2012/proceedings/60_rodriquez12w/.

[28] Sophie Rosset, Cyril Grouin, Karën Fort, Olivier Galibert, Juliette Kahn, and Pierre Zweigenbaum. Structured named entities in two distinct press corpora: Contemporary broadcast news and old newspapers. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 40–48, Jeju, Republic of Korea, July 2012. Association for Computational Linguistics.

[29] Sophie Rosset, Grouin, Cyril, Fort, Karen, Galibert, Olivier, Kahn, Juliette, and Zweigenbaum, Pierre. Structured Named Entities in two distinct press corpora: Contemporary Broadcast News and Old Newspapers. In *6th Linguistics Annotation Workshop (The LAW VI)*, pages 40–48, Jeju, South Korea, July 2012.

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

[31] Melissa M Terras. The Rise of Digitization. In Ruth Rikowski, editor, *Digitisation Perspectives*, pages 3–20. SensePublishers, Rotterdam, 2011. ISBN 978-94-6091-299-3. https://doi.org/10.1007/978-94-6091-299-3_1. URL http://dx.doi.org/10.1007/978-94-6091-299-3_1http://www.emeraldinsight.com.ezproxy.lancs.ac.uk/doi/full/10.1108/OIR-06-2015-0193.

[32] M. Vilain, J. Su, and S. Lubar. 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*, NAACL-Short '07, pages 181–184. Association for Computational Linguistics, 2007. URL http://dl.acm.org/citation.cfm?id=1614108.1614154. event-place: Rochester, New York.