Repurposing Holocaust-Related Digital Scholarly Editions to Develop Multilingual Domain-Specific Named Entity Recognition Tools

Maria Dermentzi, Hugo Scheithauer

Repurposing Holocaust-Related Digital Scholarly Editions to Develop Multilingual Domain-Specific Named Entity Recognition Tools

Maria Dermentzi*, **Hugo Scheithauer***
King's College London, Inria Paris, École pratique des hautes études
London, United Kingdom, Paris, France
name.1.surname@kcl.ac.uk, name.surname@inria.fr

Abstract

The European Holocaust Research Infrastructure (EHRI) aims to support Holocaust research by making information about dispersed Holocaust material accessible and interconnected through its services. Creating a tool capable of detecting named entities in texts such as Holocaust testimonies or archival descriptions would make it easier to link more material with relevant identifiers in domain-specific controlled vocabularies, semantically enriching it, and making it more discoverable. With this paper, we release EHRI-NER, a multilingual dataset (Czech, German, English, French, Hungarian, Dutch, Polish, Slovak, Yiddish) for Named Entity Recognition (NER) in Holocaust-related texts. EHRI-NER is built by aggregating all the annotated documents in the EHRI Online Editions and converting them to a format suitable for training NER models. We leverage this dataset to fine-tune the multilingual Transformer-based language model XLM-RoBERTa (XLM-R) to determine whether a single model can be trained to recognize entities across different document types and languages. The results of our experiments show that despite our relatively small dataset, in a multilingual experiment setup, the overall F1 score achieved by XLM-R fine-tuned on multilingual annotations is 81.5%. We argue that this score is sufficiently high to consider the next steps towards deploying this model.

Keywords: Holocaust Testimonies, Named Entity Recognition, Transformers, Multilingual, Transfer Learning, Digital Editions

1. Introduction

Launched in 2010, the European Holocaust Research Infrastructure (EHRI)[1] aims to support Holocaust research by making information about dispersed archival material held by institutions around the world more accessible and interconnected through the EHRI Portal[2] (Blanke et al., 2017). While the EHRI Portal is EHRI's flagship service, the EHRI Consortium is offering a series of additional resources, tools, and services that help researchers and archivists describe, analyze, enrich, and present Holocaust-related material using innovative methods (de Leeuw et al., 2018). Apart from the EHRI Portal, of particular relevance to this paper are the EHRI controlled vocabularies, the EHRI authority sets, and the EHRI Online Editions[3].

As an aggregator of multilingual Holocaust-related archival material from diverse institutions, the EHRI Portal is faced with a significant challenge relating to the fact that this material is often described not only in various languages but also using a variety of methodologies and in-house, language-specific controlled vocabularies that need to be normalized to a shared vocabulary to be smoothly ingested in the EHRI Portal (Erez et al., 2020). For this reason, EHRI has developed custom controlled vocabularies and authority sets mainly derived from already existing ones developed by institutions such as Yad Vashem (YV), the United States Holocaust Memorial Museum (USHMM), Arolsen Archives, etc. (Rodriguez et al., 2016; Erez et al., 2020), covering lists of concentration camps, ghettos, subject headings, personalities and corporate bodies[4]. These vocabularies are primarily used for indexing purposes in the EHRI Portal, allowing for semantic search (Colavizza et al., 2019) through keyword-based browsing and play a crucial role in achieving EHRI's goal of interlinking multilingual and heterogeneous Holocaust collections. They are also used to enhance the EHRI Online Editions and articles in the EHRI Document Blog[5] with more information and references to the EHRI Portal.

However, creating links between resources hosted across different EHRI services and the EHRI vocabularies is a resource-intensive process, usually done manually. Creating a tool capable of detecting named entities (NE) in texts such as Holocaust testimonies or the text in Holocaust-related archival descriptions would make it easier to link more material with relevant identifiers in the EHRI

*Both authors contributed equally to this work.

[1] EHRI project website. Accessed 2/27/2024.

[2] EHRI portal website. Accessed 2/27/2024.

[3] See EHRI controlled vocabularies, EHRI authority sets, and EHRI online editions. Accessed 2/27/2024.

[4] The aforementioned lists and sets are available online. See camps, ghettos, terms, personalities, and corporate bodies. Accessed 2/27/2024.

[5] See EHRI Document Blog. Accessed 2/27/2024.

vocabularies, semantically enriching it and making it more discoverable in the Portal and other EHRI services. The significance that reliable Named Entity Recognition (NER) and entity linking (EL) tools may have for EHRI has been highlighted in previous work (Rodriguez et al., 2012; de Leeuw et al., 2018). Having access to a good NER tool can help with building a reliable EL tool. EHRI partners have previously experimented with the development of such tools (Rodriguez et al., 2012; de Leeuw et al., 2018; Nikolova and Levy, 2018). However, since the publication of the most recent paper related to EHRI and NER (de Leeuw et al., 2018), EHRI's growth in resources and advances in Machine Learning (ML) promise better results compared to earlier experiments. In this paper, we report on recent work towards Holocaust-related NER.

Specifically, we treat the EHRI digital scholarly editions (i.e., EHRI Online Editions) as a dataset for training and evaluating ML-powered NER models. We have converted all available Extensible Markup Language (XML) files from the EHRI Online Editions into a trainable corpus in a format suitable for NER and have leveraged this dataset (See Table 2) to fine-tune a multilingual language model for NER. The resulting model can be used as part of a pipeline whereby, upon inputting some text into a tool that supports our models, potential named entities within the text will be automatically pre-annotated in a way that helps users detect them faster and link them to their associated controlled vocabulary entities. This has the potential to facilitate metadata enrichment of descriptions in the Portal and enhance their discoverability. It would also make it easier for EHRI to develop new Online Editions and unlock new ways for archivists and researchers within the EHRI network to organize, analyze, and present their materials and research data in ways that would otherwise require a lot of tedious work.

Our contributions are: the EHRI-NER dataset, a multilingual NER model for Holocaust-related texts, and experiments studying the multilingual learning and cross-lingual transfer capabilities of Deep Learning NER techniques. In what follows, we describe related work (Section 2) and provide detailed information on the source of our dataset, the EHRI Online Editions (Section 3). Subsequently, we detail how we put together the dataset (Section 4) and how we designed and carried out our fine-tuning experiments (Section 5). We conclude with a summary and future research pathways (Section 6).

2. Related Work

Previously, EHRI experimented with applying off-the-shelf NER tools to the Optical Character Recognition (OCR) output of type-written Holocaust survivor testimonies and newsletters for the crew of H.M.S. Kelly (Rodriguez et al., 2012). Due to the lack of an already available annotated corpus for domain-specific NER tools, Rodriguez et al. (2012) manually annotated the OCRed corpus compiled for their experiments. Given the lack of resources, their experiments remained limited and focused on comparing which of the then-existing NER tools yielded the best results. The maximum total F1 score achieved across all tools and datasets under consideration was 60% (Rodriguez et al., 2012). In 2018, de Leeuw et al. (2018) published another paper detailing EHRI's efforts to offer reliable NER services for the Holocaust domain. They reiterated the lack of suitable corpora and crafted their own gold corpus by crowd-sourcing annotations on transcripts of oral testimonies provided by the USHMM (de Leeuw et al., 2018; Nikolova and Levy, 2018). They used this corpus to develop person and location extraction services. Their methodology included fine-tuning and extending commercial software and they achieved an F1 score of 77% for person extraction. For location extraction, they adapted a proprietary service to tag and disambiguate locations in Holocaust testimonies. The details of these tools are not specified but the authors reported a resulting F1 score of 91% for the disambiguated place-related access points, although it is unclear how the first part of their pipeline (i.e. the tagger) performed. To our knowledge, neither the purpose-built NER datasets nor the EHRI-specific tools developed during earlier work are publicly available today or were formally deployed as EHRI services.

Apart from EHRI-related efforts, there is a broader interest in applying NER tools on Holocaust-related texts (Ezeani et al., 2023; Carter et al., 2022) as well as in developing domain-specific ones. Notable examples include Mattingly's (2021a; 2021b) lessons on Holocaust NER and Nanomi Arachchige et al.'s (2023) paper detailing their work on compiling and annotating an English corpus for Holocaust-related NER, which they used to train and evaluate rules-based and transformer-based (Vaswani et al., 2017) tools. Consistent with other publications (Luthra et al., 2023; Ehrmann et al., 2023), many of the Transformer-based models included in Nanomi Arachchige et al.'s (2023) experiments achieved high F1 scores across most of the entities considered, encouraging us to select a similar architecture for our experiments.

However, since the material processed by EHRI is diverse and multilingual, we wanted to work towards developing a single multilingual NER model that would leverage multilingual learning for cross-lingual transfer (Mueller et al., 2020; Ehrmann et al., 2023; Schweter et al., 2022; Wu et al., 2020). Mul-

tilingual NER in historical documents has seen a growing interest amongst the Digital Humanities (DH), Natural Language Processing (NLP), and cultural heritage communities (Ehrmann et al., 2023). In 2022, Ehrmann et al. (2022) introduced a shared task on NER and EL in multilingual historical documents, encouraging researchers to study approaches that can work well across different contexts and languages. Ehrmann et al. (2022) acknowledge that advances in AI thanks to the Transformer architecture and the increased availability of suitable resources create new opportunities for working towards such solutions. The same is true in the EHRI context, where since the work of Rodriguez et al. (2012) and Nikolova and Levy (2018), EHRI has produced a series of manually annotated digital scholarly editions. Although the original purpose of these editions was not to provide a dataset for training NER models, we argue that they nevertheless constitute a high-quality resource that is suitable to be used in this way. We therefore repurposed them to train multilingual Transformer-based NER models testing the hypothesis that we now have enough resources to develop a single domain-specific tool that can work reliably well across different languages and document types encountered in EHRI collections.

3. EHRI Online Editions

Since 2018, the EHRI Consortium has supported the development and publication of six Holocaust-related digital scholarly editions[6] (EHRI-Consortium, 2021; Frankl and Schellenbacher, 2018; Frankl et al., 2023; Frankl and Schellenbacher, 2023; Frankl et al., 2020; Garscha et al., 2022). Each edition enables digital access to facsimiles and transcripts of thematically related documents held by different EHRI partner institutions through a single web interface and unlocks new ways of presenting and browsing through historical sources using digital tools. Publishing a digital edition is a resource-intensive process. Notwithstanding the extensive archival research needed for selecting the documents, additional steps include transcribing and translating them and, most importantly, annotating words and phrases found within these texts and creating links with entities in controlled vocabularies provided by EHRI and third parties. Currently, this annotation is done manually by or under the supervision of subject matter experts, ensuring a high quality of annotations[7]. We repurposed these resources to convert them into a dataset suitable for training NER models, which we consider as a gold standard.

[6] At the time of writing: 2/26/2024.

[7] More info about this process can be found on the website of each edition.

Each EHRI Online Edition consists of digitized documents originating from various archives that are selected, edited, and annotated by EHRI researchers using the Text Encoding Initiative (TEI) P5 standard (TEI Consortium, 2023), an XML schema, which supports their online publication. Editions enhance the edited documents by contextualizing the information contained within them and linking them to EHRI vocabularies and descriptions, and by visualizing georeferenced entities through interactive maps. Thanks to their encoding in TEI, they are fully searchable and can be filtered using facets such as spatial locations, topics, persons, organizations, and institutions. All documents within an edition have a transcript, either in their original language, a translation, or both, and have access to their facsimile. EHRI Editions are published without a regular schedule and it is possible to update them with new material or improve the already published documents. In the following paragraphs, we present each edition individually.

Begrenzte Flucht Edition The BeGrenzte Flucht (BF) edition (Frankl and Schellenbacher, 2018) gathers documents kept in various Czech and Austrian archives relating to Austrian refugees on the border to Czechoslovakia in the crisis year 1938, including official reports, correspondence, diplomatic notes, newspaper reports, and documents from Jewish aid organizations. The BF edition is in German and the vast majority of documents, if not originally in German, have been translated into German. Transcripts in the original languages of the documents, including Czech, Slovak, and English are also included.

Early Holocaust Testimonies Edition The Early Holocaust Testimony (EHT) edition (Frankl et al., 2020) contains selected and edited testimonies and reports kept in five different archives: the Wiener Holocaust Library in London, Yad Vashem in Jerusalem, the Jewish Historical Institute in Warsaw, the Hungarian Jewish Archives in Budapest, and the Jewish Museum in Prague. All of the documents have an English translation but transcripts of the original documents in Czech, German, Hungarian, Polish, Dutch, and Yiddish are provided.

Diplomatic Reports Edition The Diplomatic Reports (DR) edition (EHRI-Consortium, 2021) gathers documents created by the diplomatic staff of allied countries, opponents, and neutral countries. They all report on the German occupation. They include reports from the diplomatic staff of Denmark, Italy, Japan, Hungary, Slovakia, and the US. All of the documents have been translated into English, regardless of their original language.

Von Wien ins Nirgendwo: Die Nisko-Deportationen 1939 Edition The Von Wien ins Nirgendwo: Die Nisko-Deportationen 1939 (ND) edition (Garscha et al., 2022) was created in cooperation with the Documentation Archive of the Austrian Resistance. It gathers documents on the history and the fate of the Viennese Jewish deportees to Nisko, Poland, in 1939. The source documents are from various archival institutions in different countries and are provided in German.

Documentation Campaign Edition The Documentation Campaign (DC) edition (Frankl et al., 2023) gathers documents held by the Jewish Museum in Prague and by Yad Vashem consisting of Holocaust survivor testimonies and photographs collected within the framework of the so-called "Documentation Campaign" in Prague, one of the earliest postwar projects to document the events of the Shoah, collecting evidence, documents, and witness testimonies. All of the documents have been translated into English but transcripts of the original documents in Czech and German are provided.

Uzavřít Hranice Edition Similar to the BF edition, the Uzavřít Hranice (UH) edition (Frankl and Schellenbacher, 2023) gathers documents kept in various Czech, Austrian, and other archives relating to Austrian refugees on the border to Czechoslovakia in the crisis year 1938. The UH edition is in Czech and the vast majority of documents, if not originally in Czech, have been translated into Czech. Transcripts in the original languages of the documents, including Czech, Slovak, and English are also included.

Since the EHRI Online Editions cover a variety of languages, document types, periods, and thematic and spatial areas of focus, training NER models on this dataset may lead to tools that can generalize better on different types of Holocaust-related documents, compared to training them only on testimony-based corpora like in previous work. This will hopefully make our models more robust and interoperable across different EHRI services.

4. The EHRI-NER Dataset

This section presents EHRI-NER, a multilingual NER dataset derived from the EHRI Online Editions. We fully released EHRI-NER on Hugging Face and GitHub[8].

[8]See the EHRI-NER organization on Hugging Face to access the model and dataset and the EHRI-NER GitHub repository to access the dataset subsets per language.

4.1. Languages and Subsets

We sorted all TEI XML files available from the EHRI Online Editions by language. The resulting EHRI-NER dataset includes nine languages: Czech (cs), German (de), English (en), French (fr), Hungarian (hu), Dutch (nl), Polish (pl), Slovak (sk), and Yiddish (yi). We created a subset for each language since they are not represented in the same proportion.

As presented in Section 3, the dataset includes official reports, correspondences, diplomatic notes, newspaper reports, and testimonies. The creation dates of the documents span from 1936 to 2001.

4.2. From TEI XML to the IOB Format

To build the subsets, we created a Python script to parse the TEI XML documents and convert them to the CoNLL Inside-Outside-Beginning (IOB) format (Sang and De Meulder, 2003), which is typical for NER datasets (Ehrmann et al., 2016)[9].

The BF, UH, DC, and EHT editions all include translations of some of their original transcribed documents. To avoid contaminating our validation and test sets, we filtered them out. Additionally, both the BF and the UH editions contain some documents that overlap. We also filtered these out to avoid having duplicates in our dataset.

4.3. Entity Classes

Given that the primary purpose of this work is to enhance the services and facilitate the work of EHRI stakeholders, we used a custom typology of entity classes that corresponds better to how we envision deploying this tool in the EHRI environment, extending the CoNLL typology (Sang and De Meulder, 2003) to include classes such as camps and ghettos, which correspond to custom EHRI vocabularies used when annotating Holocaust materials to produce new EHRI Editions. However, our typology is coarser compared to more fine-grained typologies found in similar work (Nanomi Arachchige et al., 2023). We extracted all TEI elements `<persName>`, `<placeName>`, `<orgName>`, and `<date>` from the selected TEI XML files. The `<placeName>` element sometimes includes an attribute @type to indicate whether it is referencing a concentration camp or a ghetto. We distinguish between `<placeName>`, `<placeName type="camp">`, and `<placeName type="ghetto">` to include fine-grain camp and ghetto entities in addition to the coarse-grain location entity. The conversion table is presented in Table 1.

EHRI TEI XML files also contain the `<term>` entity, used for annotating various subjects related to

[9]Our script is available on GitHub.

the Holocaust and for linking them with their associated entries in the EHRI vocabulary of terms[10]. However, we have chosen to consider these instances as non-entity tokens, as their broad coverage of themes, their variability, and lack of semantic regularity in how they are used in annotations make them unsuitable in a token classification context. Had we included them in our typology, we hypothesize that the NER models would tag a disproportionate number of tokens as terms, rendering the output noisy and confusing. Instead, EHRI is working on a different solution for extracting subject metadata, which is outside the scope of this paper.

The EHRI-NER dataset includes a total of 505758 tokens, with 5351 person entities, 9399 location entities, 1867 organization entities, 2237 date entities, 528 ghetto entities, and 1229 camp entities. The distribution of tokens and entity classes is detailed in Table 2.

TEI XML Element	Entity Class
`<persName>Helene Hirsch</persName>`	Person
`<placeName>Berlin</placeName>`	Location
`<orgName>Gestapo</orgName>`	Organization
`<date when="1937-10">Oct. 1937</date>`	Date
`<placeName type="camp">Auschwitz</placeName>`	Camp
`<placeName type="ghetto">getcie</placeName>`	Ghetto

Table 1: Conversion table for TEI XML Elements and Entity Classes.

4.4. Data Format and Preprocessing

We chose to convert TEI annotations and non-entity tokens into the CoNLL IOB format, as presented in Sang and De Meulder (2003) (see Table 3). The IOB format ensures that our dataset is interoperable with common NER tools. Each token and its annotation have been put on a separate line and there is an empty line after each sentence, as shown in the following example:

```
Von O
Gross B-CAMP
- I-CAMP
Rosen I-CAMP
Bahntransport O
nach O
Buchenwald B-CAMP
. O
```

Each language subset has been tokenized at the sentence and word levels. We used SpaCy (Honnibal et al., 2020) and its multi-language pipeline to process each subset[11].

[10]See the EHRI Terms database. Accessed 2/27/2024.

[11]See the multi-language pipelines available on SpaCy website. Accessed 2/27/2024.

5. Experimental Setup

We conducted two experiments to determine whether our dataset was sufficiently large for fine-tuning a reliable NER model that could be used in a real-life setting, e.g. speeding up named entity annotation when curating a new EHRI Online Edition. We also leveraged the multilingual aspect of our dataset to test XLM-RoBERTa (XLM-R) (Conneau et al., 2020) in a low-resource setting, as our dataset is significantly smaller than, for instance, the CONLL2003 NER dataset used for evaluating this model on a token classification task. In this section, we describe the model that we used for fine-tuning, the experiments we conducted on the dataset, and their results.

5.1. Model

We chose to experiment with the multilingual Transformer-based masked language model `XLM-RoBERTa-large` (Conneau et al., 2020) as it demonstrates high efficacy in multilingual settings and strong cross-lingual transfer capabilities, especially on token classification tasks, without sacrificing per-language performance[12]. According to Nanomi Arachchige et al. (2023), this model outperforms the multilingual hmBERT (Schweter et al., 2022) model which was pre-trained on German, French, Swedish, Finnish, and English historical newspapers (thus not pre-trained in all of the languages present in our dataset). It is important to note that XLM-R has seen all languages represented in the EHRI-NER dataset during its pre-training.

The same fine-tuning parameters were kept for all our experiments. The learning rate is set at $3e^{-5}$, the number of epochs for training at 3 to avoid overfitting, the weight decay at 0.01, and the train and evaluation batch size at 16.

5.2. Experiments

Experiment 1: We fine-tuned XLM-R on all subsets (cs, de, en, fr, hu, nl, pl, sk, yi) to evaluate the overall performance of the model on a multilingual level. Instead of relying on a simple shuffle, and to ensure that all languages are represented in the train, validation, and test set, we first split each subset into train (80%), validation (10%), and test (10%) sets, using a seed of 42 for result reproducibility[13]. Each split subset is then concatenated and the final dataset is used for fine-tuning. Our objective was to acquire a single fine-tuned model

[12]See `XLM-RoBERTa-large` model card on Hugging Face website. Accessed 02/27/2024.

[13]The mentioned seed used for splitting the subsets was used for all the experiments.

ISO code	Language	Tokens	PERS	LOC	ORG	DATE	GHETTO	CAMP
cs	Czech	106392	1415	2627	359	741	**212**	**502**
de	German	**218570**	**2516**	**3592**	**871**	**950**	202	396
en	English	58 405	363	1015	225	287	52	77
fr	French	2273	3	39	8	4	0	5
hu	Hungarian	24686	157	304	148	97	2	114
nl	Dutch	1991	17	25	33	7	0	2
pl	Polish	18385	221	328	54	126	17	51
sk	Slovak	3550	30	158	11	21	0	0
yi	Yiddish	71506	629	1311	158	4	43	82
/	All	505758	5351	9399	1867	2237	528	1229

Table 2: EHRI-NER Dataset: tokens and entity classes distribution.

Entity	Example	Annotation
Person	Kurt Lichtenstern	B-PERS, I-PERS
Location	Moravská Ostrava	B-LOC, I-LOC
Organization	Pártfogó iroda	B-ORG, I-ORG
Date	1941 roku	B-DATE, I-DATE
Ghetto	getta łódzkiego	B-GHETTO, I-GHETTO
Camp	Auschwitz camp	B-CAMP, I-CAMP

Table 3: Entity types illustrated with examples and IOB tagging.

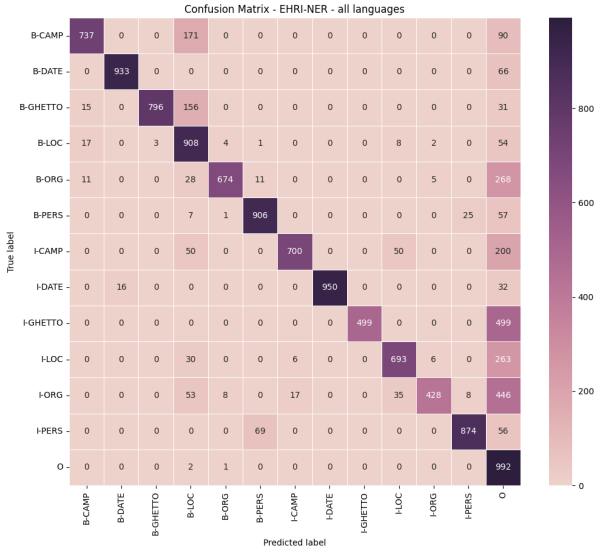

Figure 1: Matrix confusion for predicted classes in the test set, when fine-tuning XLM-R on **all languages** (experiment 1, Section 5.2). The confusion matrix was normalized using a scaling factor of 1000.

with reliably good performance across most if not all languages, suitable primarily as part of an editorial pipeline that streamlines the creation of new digital scholarly editions related to the Holocaust.

Experiment 2: To assess the cross-lingual capabilities of XLM-R in a low-resource setting, we fine-tuned it three more times—each time leaving out one language subset which was reserved for testing. Our chosen target languages were nl (experiment 2.1) and yi (experiment 2.2) as they represent some of the smallest subsets, while still containing enough examples for meaningful evaluation. For each target language, we fine-tuned XLM-R on every other subset split into train and validation (80% / 20%) and used the entire subset of the target language as the test set. This experiment sought to simulate a scenario where we would need to use our fine-tuned model to pre-annotate documents from a Holocaust domain but in a language not seen by our model during fine-tuning.

The fine-tuning processes were repeated three times for each experiment, we then computed the average of each of the three runs to obtain a reliable evaluation.

5.3. Evaluation

Experiment 1 yielded a consistent and satisfying overall performance across the validation and the test sets, with an overall F1 score of 81.3% for the former and 81.5% for the latter (Table 4), achieving higher scores compared to earlier EHRI NER work, surpassing Rodriguez et al.'s 2012 maximum total F1 score of 60% while additionally tagging domain-specific entities and exceeding the F1 score of 77% reported for the person tagger (de Leeuw et al., 2018). Domain-specific entities (Camp, Ghetto) are also consistently classified by the model. Only the Organization entity demonstrated poor F1 scores probably caused by the relatively low number of examples (1867 in total). This behavior has been previously observed by Rodriguez et al. (2012). The overall evaluations for cs, de, en, hu, pl, and yi test sets showed that the performance of the fine-tuned model is corollary to the number of examples in the training set. However, even though we see a

decrease in the F1 scores depending on the size of the subset (minimum 73.2% overall F1 score for the hu test set), we still consider the performance of the fine-tuned model strong considering the relatively small size of some of the subsets.

The confusion matrix for predicted classes in the test set (Fig. 1) shows instances where the fine-tuned model occasionally misclassifies entities as non-entity tokens, I-GHETTO being the most confused entity. The fine-tuned model occasionally encounters challenges in extracting multi-tokens entities, such as I-CAMP, I-LOC, and I-ORG, which are sometimes confused with the beginning of an entity. Moreover, it tends to misclassify B-GHETTO and B-CAMP as B-LOC, which is not surprising given that they are semantically close and there are cases where even an expert would hesitate to pick a single label. Indeed, sometimes an entity such as the camp/ghetto "Theresienstadt" could be assigned any of these classes without introducing errors[14].

Overall, we argue that these scores are high enough to at least pre-annotate Holocaust-related textual documents when developing a new EHRI Online Edition or when wanting to enrich an archival description with access points that an archivist can verify. Additionally, as long as the new unseen texts to be fed into the model belong to a similar domain and period, we can assume that the scores will remain relatively consistent across all nine languages used for fine-tuning.

We released the fine-tuned XLM-R model on Hugging Face[15].

Experiment 2 revealed that we can leverage the cross-lingual capabilities of XLM-R depending to some extent on how much data it has seen about a specific language during its pre-training and on how many examples the training dataset has.

Experiment 2.1 showed unexpectedly high performance, about 94% overall F1 score, in one of the runs on the Dutch subset. However, it decreased in the second run to around 80% F1 score. After the third run, the overall F1 score of 84.6% proved that the fine-tuned model achieved satisfying performance, except for the classification of Organization entities (see Table 5), and despite not being evaluated on the Ghetto entity due to lack of examples. The confusion matrix shows that the fine-tuned model has trouble extracting multi-token entities, as noted in experiment 1 (Fig. 2).

Experiment 2.2 on Yiddish yielded poor performance, with an overall F1 score of 46.5% (Table 6). Only Person and Location entities showed an F1

[14]See more about the function of Theresienstadt here. Accessed 2/27/2024.

[15]See the EHRI-NER fine-tuned XLM-R model on Hugging Face.

Figure 2: Evaluation of XLM-R on the **nl subset, when fine-tuned on all languages except nl**, by entity type (experiment 2.1, Section 5.2).

score of above or equal to 50%. The Organization entity and the domain-specific entities Date, Camp, and Ghettos are all under 10% F1 score, the latter having an F1 score of 0. As depicted in Fig. 3, the model mainly misclassified entities as non-entity tokens, which is a common problem in NER (Luthra et al., 2023).

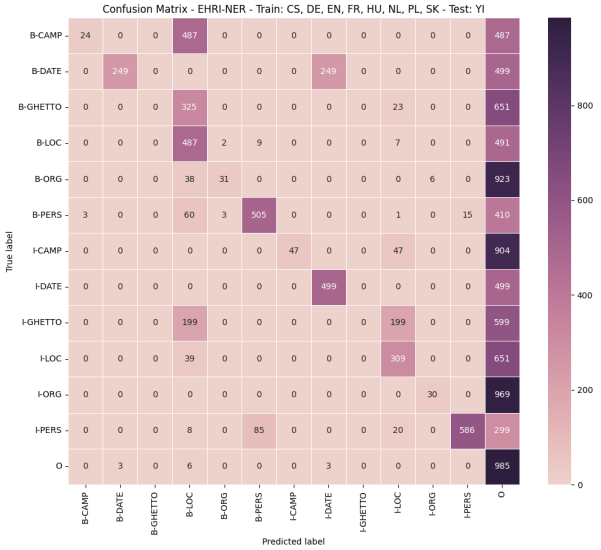

Figure 3: Matrix confusion for predicted classes in the **yi subset**, when fine-tuning XLM-R on **all languages except yi** (experiment 2.2, Section 5.2). The confusion matrix was normalized using a scaling factor of 1000.

The fluctuations in performance are probably related to the pre-training of XLM-R. As reported in Conneau et al. (2020), the model was pre-trained

	Validation Set				Test Set			
Entity	**Acc.** (%)	**Prec.** (%)	**Rec.** (%)	**F1** (%)	**Acc.** (%)	**Prec.** (%)	**Rec.** (%)	**F1** (%)
Person	/	**85**	90.3	87.5	/	83.8	**88.7**	**86.2**
Location	/	78.1	86	81.8	/	78.1	87.3	82.5
Organization	/	62.3	56.8	59.4	/	61.9	60.7	61.3
Date	/	81.5	**92.9**	**86.8**	/	81.1	90.3	85.4
Camp	/	76.4	68.7	72.3	/	73	72.7	72.8
Ghetto	/	75.2	75.2	75.2	/	**87.1**	80.7	83.7
Overall	98	78.9	83.9	81.3	98	78.6	84.7	81.5
Overall - CS test set	/	/	/	/	98.3	82.5	87.1	84.7
Overall - DE test set	/	/	/	/	98.6	78	86.6	82.1
Overall - EN test set	/	/	/	/	98	75.4	84.4	79.6
Overall - HU test set	/	/	/	/	98.5	71.9	74.6	73.2
Overall - PL test set	/	/	/	/	97.2	73.3	77.7	75.5
Overall - YI test set	/	/	/	/	98.5	75.6	78.8	77.2

Table 4: Evaluation of fine-tuned XLM-R on EHRI-NER on **all languages**, by entity type (experiment 1, Section 5.2), and specific overall evaluation on cs, de, en, hu, pl, and yi test sets. fr, nl, and sk test sets were omitted because of a lack of examples.

Entity	**Acc.** (%)	**Prec.** (%)	**Rec.** (%)	**F1** (%)
Person	/	100	96	97.9
Location	/	83.2	96	89
Organization	/	76.1	61.6	67.5
Date	/	100	100	100
Camp	/	100	100	100
Ghetto	/	/	/	/
Overall	98.7	86.4	82.9	84.6

Table 5: Evaluation of XLM-R on the **nl subset, when fine-tuned on all languages except nl**, by entity type (experiment 2.1, Section 5.2).

Entity	**Acc.** (%)	**Prec.** (%)	**Rec.** (%)	**F1** (%)
Person	/	68.9	53.1	59.9
Location	/	48.4	52.2	50
Organization	/	21.3	04.8	07.6
Date	/	00.7	41.6	01.3
Camp	/	28.4	02	03.7
Ghetto	/	0	0	0
Overall	96.6	47.2	46.2	46.5

Table 6: Evaluation of XLM-R on the **yi subset, when fine-tuned on all languages except yi**, by entity type (experiment 2.2, Section 5.2).

on 5025M tokens for Dutch, but merely 34M tokens for Yiddish. Therefore, we can hypothesize that the performance of XLM-R on the Yiddish subset is likely due to the limitations in its representation of this language after pre-training. This may have impacted the fine-tuning of the model and its cross-lingual capabilities for a token classification task on a small subset, such as the Yiddish subset, whereas the fine-tuning on the Dutch subset, despite being smaller, achieved a good performance. Other work on the zero-shot language transfer capabilities of multilingual Transformer models supports this hypothesis (Lauscher et al., 2020). Since the authors do not understand Yiddish, a comprehensive error analysis was not possible. However, it is worth noting that the challenges observed, as shown in experiment 1, can be mitigated when fine-tuning XLM-R on all subsets.

This experiment also confirms the hypothesis we made when discussing the lack of examples for the Organization entity and its consequence on the results in experiment 1.

6. Conclusion

In this work, we released EHRI-NER, a multilingual dataset for NER in Holocaust-related textual documents, built from the numerous TEI XML files made available across all EHRI Online Editions. We also evaluated the multilingual and cross-lingual capabilities of XLM-R by fine-tuning it on our dataset and proved that it can perform well when using relatively small domain-specific datasets. We also provided a baseline for future evaluations of NER systems on the dataset. Our future objective for the dataset is to include the translations mentioned in Subsection 4.2 while filtering them out from the train set. They indeed represent a sizable portion of data that would increase the number of examples in our dataset and could potentially lead to an increase in the fine-tuned model's performance.

For future work, we would like to experiment on

multilingual named entity disambiguation, which would allow us to automatically link recognized entities with IDs in the EHRI vocabularies mentioned in the introduction (1). Another idea for future work could be to source similar annotated datasets and merge them with EHRI-NER. As a next step, we are planning to invite EHRI partners to evaluate our model qualitatively as part of their work and provide feedback. Based on that feedback, we can improve our model and deploy it as part of EHRI's cataloging and editorial pipelines. Another interesting course for future work would be to create a stable annotation typology for Holocaust documents with the help of experts. Finally, we hope to be able to provide a more complete baseline by experimenting with more multilingual Large Language Models (LLMs), including state-of-the-art LLMs for zero-shot and few-shot NER.

7. Acknowledgments

This research was conducted within the framework of the EHRI-3 project, which is funded by the European Commission under the call H2020-INFRAIA-2018–2020, with grant agreement ID 871111 and DOI 10.3030/871111.

We would like to acknowledge and thank Mike Bryant (King's College London), Lydia Nishimwe (Inria Paris, ALMAnaCH), and Armel Randy Zebaze (Inria Paris, ALMAnaCH) for their guidance and helpful discussions. The work described herein was made possible thanks to the previous work of the editors and contributors of the EHRI Online Editions, including the annotators, the people who produced digital facsimiles of the original archival material, and those who created the transcripts and translations. Finally, we would like to thank the open-source communities for generously sharing tools and know-how.

8. Bibliographical References

Tobias Blanke, Michael Bryant, Michal Frankl, Conny Kristel, Reto Speck, Veerle Vanden Daelen, and René Van Horik. 2017. The European Holocaust Research Infrastructure Portal. *Journal on Computing and Cultural Heritage*, 10(1):1–18.

Kirsten Strigel Carter, Abby Gondek, William Underwood, Teddy Randby, and Richard Marciano. 2022. Using AI and ML to optimize information discovery in under-utilized, Holocaust-related records. *AI & SOCIETY*.

Giovanni Colavizza, Maud Ehrmann, and Fabio Bortoluzzi. 2019. Index-Driven Digitization and Indexation of Historical Archives. *Frontiers in Digital Humanities*, 6.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. ArXiv:1911.02116 [cs].

TEI Consortium. 2023. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Publisher: Zenodo.

Daan de Leeuw, Mike Bryant, Michal Frankl, Ivelina Nikolova, and Vladimir Alexiev. 2018. Digital Methods in Holocaust Studies: The European Holocaust Research Infrastructure. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pages 58–66.

EHRI-Consortium. 2021. *Diplomatic Reports - Digital Edition*. EHRI Online Editions. European Holocaust Research Infrastructure project (EHRI).

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):27:1–27:47.

Maud Ehrmann, Damien Nouvel, and Sophie Rosset. 2016. Named Entity Resources - Overview and Outlook. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3349–3356, Portorož, Slovenia. European Language Resources Association (ELRA).

Maud Ehrmann, Matteo Romanello, Antoine Doucet, and Simon Clematide. 2022. Introducing the HIPE 2022 Shared Task: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 347–354, Cham. Springer International Publishing.

Sigal Arie Erez, Tobias Blanke, Mike Bryant, Kepa Rodriguez, Reto Speck, and Veerle Vanden Daelen. 2020. Record linking in the EHRI portal. *Records Management Journal*, 30(3):363–378. Num Pages: 16 Place: Bradford, United Kingdom Publisher: Emerald Group Publishing Limited.

Ignatius Ezeani, Paul Rayson, Ian Gregory, Erum Haris, Anthony Cohn, John Stell, Tim Cole, Joanna Taylor, David Bodenhamer, Neil Devadasan, Erik Steiner, Zephyr Frank, and Jackie Olson. 2023. Towards an Extensible Framework for Understanding Spatial Narratives. In *Proceedings of the 7th ACM SIGSPATIAL International*

Workshop on Geospatial Humanities, pages 1–10, Hamburg Germany. ACM.

Michal Frankl and Wolfgang Schellenbacher, editors. 2018. *BeGrenzte Flucht: Die österreichischen Flüchtlinge an der Grenze zur Tschechoslowakei im Krisenjahr 1938 - Digital Edition*. EHRI Online Editions. European Holocaust Research Infrastructure project (EHRI).

Michal Frankl and Wolfgang Schellenbacher, editors. 2023. *Uzavřít hranice! - Digital Edition*. EHRI Online Editions. European Holocaust Research Infrastructure project (EHRI).

Michal Frankl, Magdalena Sedlická, Hana Dauš, and Wolfgang Schellenbacher, editors. 2023. *Documentation Campaign - Digital Edition*. EHRI Online Editions. European Holocaust Research Infrastructure project (EHRI). Partner Institution: Masaryk Institute and Archives of the Czech Academy of Sciences.

Michal Frankl, Magdalena Sedlická, Wolfgang Schellenbacher, Daniela Bartáková, Michał Czajka, Jessica Green, Kat Hubschmann, Gábor Kádár, Yehudit Levin, Daphna Sehayek, Christine Schmidt, Zoltán Vagi, and Marta Wojas, editors. 2020. *Early Holocaust Testimony - Digital Edition*. EHRI Online Editions. European Holocaust Research Infrastructure project (EHRI), 2020.

Winfried Garscha, Claudia Kuretsidis-Haider, and Wolfgang Schellenbacher, editors. 2022. *VON WIEN INS NIRGENDWO: DIE NISKO-DEPORTATIONEN 1939*. EHRI Online Editions. Funded by: Nationalfonds der Republik Österreich für Opfer des Nationalsozialisten, Zukunftsfonds der Republik Österreich, Bundesministerium für Soziales, Gesundheit, Pflege und Konsumentenschutz.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Mrinalini Luthra, Konstantin Todorov, Charles Jeurgens, and Giovanni Colavizza. 2023. Unsilencing colonial archives via automated entity recognition. *Journal of Documentation*, ahead-of-print(ahead-of-print).

W.J.B. Mattingly. 2021a. Holocaust Named Entity Recognition.

W.J.B. Mattingly. 2021b. wjbmattingly/holocaust_ner_lessons. Original-date: 2021-01-04T18:25:11Z.

David Mueller, Nicholas Andrews, and Mark Dredze. 2020. Sources of Transfer in Multilingual Named Entity Recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8093–8104, Online. Association for Computational Linguistics.

Isuri Anuradha Nanomi Arachchige, Le Ha, Ruslan Mitkov, and Johannes-Dieter Steinert. 2023. Enhancing Named Entity Recognition for Holocaust Testimonies through Pseudo Labelling and Transformer-based Models. In *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, pages 85–90, San Jose CA USA. ACM.

Ivelina Nikolova and Michael Levy. 2018. Using Named Entity Recognition to Enhance Access to a Museum Catalog – Document Blog.

Kepa J Rodriguez, Vladimir Alexiev, Laura Brazzo, Charles Riondet, Yael Gherman, and Reto Speck. 2016. EHRI-2 - D.11.2 Road Map Domain Vocabularies. Deliverable GA no. 654164. Issue: GA no. 654164.

Kepa Joseba Rodriguez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. 2012. Comparison of named entity recognition tools for raw OCR text. pages 410–414.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. ArXiv:cs/0306050.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmBERT: Historical Multilingual Language Models for Named Entity Recognition. In *CEUR Workshop Proceedings*, Bologna, Italy. arXiv. ArXiv:2205.15575 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F. Karlsson, Biqing Huang, and Chin-Yew Lin. 2020. Enhanced Meta-Learning for Cross-Lingual Named Entity Recognition with Minimal

Resources. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9274–9281. Number: 05.