HoloBERT: Pre-Trained Transformer Model for Historical Narratives

Isuri Anuradha Nanomi Arachchige
Lancaster University, UK
i.nanomiarachchige@lancaster.ac.uk

Le An Ha
Ho Chi Minh City University , Vietnam
anhl@huflit.edu.vn

Ruslan Mitkov
Lancaster University , UK
r.mitkov@lancaster.ac.uk

Abstract

Narratives often consist of spontaneous, unstructured language with features such as disfluencies, colloquialisms, and non-standard syntax. In this paper, we investigate how further pretraining language models enhances the performance in Named Entity Recognition (NER) related to the Holocaust narrative discourse. To evaluate our models, we compare the extracted named entities (NE) against pretrained language models on historical texts and generative AI models such as GPT. Furthermore, we demonstrate practical applications of the recognised NEs by linking them to a knowledge base as structured metadata and representing them in a graph format. With these contributions, our work illustrates how the further-pretrain-and-fine-tune paradigm in Natural Language Processing advances research in Digital Humanities.

1 Introduction

The field of Natural Language Processing(NLP) has witnessed a transformative shift propelled by the emergence of the 'further-pre-train-and-fine-tune' paradigm. This paradigm leverages the capabilities of large language models(LLMs), often built on the foundation of powerful transformer architectures, that have been trained on extensive text collections. Our aim is to explore the hidden capabilities of language models in the domain-specific context of understanding, analysing and preserving the accounts of the Holocaust narratives.

For the pre-training phase, LLM requires comparatively high-performing hardware infrastructure, yet provides robust performance in a wide range of NLP tasks. Holocaust narratives, distinguished by their unique linguistic and contextual characteristics, present both challenges and opportunities for exploring pre-trained language models due to: 1) the unstructured nature of the text, 2) a blend of spoken and written language, and 3) linguistic noise, including grammatical errors, mispronunciations, and interruptions.

Meanwhile, Natural Language Understanding(NLU) tasks have adapted the 'pre-train-and-fine-tune' paradigm and demonstrated significant performance improvements. Notably, the progression of language models has shown that pre-trained and fine-tuned transformer-based models such as BERT are highly effective for a variety of downstream NLP tasks when sufficient data is available. Gururangan et al. (2020) suggested that further pretraining was able to improve the performance in both domain-specific and task-adaptive downstream tasks even in a low-resource setting. However, the above hypothesis has not been particularly experimented with in the historically significant narratives. Therefore, from this study, we propose a further-pretrained language model for Holocaust narratives and evaluate its performance by fine-tuning it for a downstream task, making the following contributions:

- We present the further pretrained model with historically significant textual data [1] and fine-tuned for domain-specific Named Entity Recognition, according to the state of the art of digital humanities.
- We create a blueprint for linking extracted NER tags with Wikipedia IDs to get more information, demonstrating where these models can be applied in real-world scenarios. [2]

2 Related Work

With the field of Digital Humanities growing, the development of BERT-like models trained on 18th-century and 19th-century historical data. Since

[1] https://huggingface.co/Isuri97/
[2] https://github.com/isuri97/Pretraining-holo

Proceedings of Recent Advances in Natural Language Processing,pages 105–110
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-012

there is a limited number of studies considering computational approaches taken for analysing Holocaust narratives, this study explores more on transformer models developed on historical textual corpora. In an early pioneering effort, Hosseini et al. (2021) and Beelen et al. (2021) took a significant step by further training a standard BERT model, utilising English books published between 1760 and 1900, along with data from the Oxford English Dictionaries. The BERT model was pretrained from scratch, using the Eighteenth Century Collections Online (ECCO) dataset, which encompasses a vast collection of over 180,000 titles published during the eighteenth century. Another model, MacBERTh (Manjavacas and Fonteyn, 2021) was also pre-trained from scratch and evaluated for various tasks such as Part-of-Speech (POS) tagging, Named Entity Recognition and Word Sense Disambiguation. The authors of this study further explained that the pre-training approach exceeds the capabilities of the standard BERT.

In addition to historical documents available in English, a multi-lingual BERT model has been developed employing French historical documents, following the same approach as MacBERTh. Moreover, another BERT-based model (BERToldo) was developed using Italian books and Wikisource, which was evaluated using Dante Alighieri's works for POS tagging tasks (Palmero Aprosio et al., 2022).

3 Data sources

In this research, we used two distinct datasets with our further pretrained language model. As a special remark Both datasets discussed above are annotated following the BIO schema. These annotations include domain-specific tags such as *B-STREET*, *B-CAMP*, *B-BUILDING*, as well as general tags such as *B-PERSON*, *B-ORG*, *I-LOC*, etc. The Holocaust dataset has more domain-specific tags, which are not included in the HIPE dataset.

3.1 Holocaust testimony dataset

Holocaust testimonies stand as the primary source of information that describes the Holocaust, offering firsthand accounts and personal narratives of those who experienced it. The Holocaust testimonies are highly unstructured by encapsulating biographical, temporal, geographical and emotional information. For the proposed experiment, we employed 2000 digitised Holocaust testimonies collected and preserved by the US Holocaust Memorial Museum, Wiener Library and Fortuneoff video archive. More information about dataset and annotation procedure is described in Anuradha Nanomi Arachchige et al. (2023).

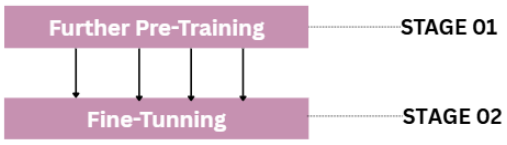

Figure 1: Two Stage training process

3.2 HIPE dataset

For fine-tuning, we utilized the HIPE dataset, specifically the topres19th (18th–19th century British newspapers) and ajmc (historical commentaries) training and development splits. This data was chosen despite its known challenges with OCR errors and lack of gold-standard alignment, as noted in (Ehrmann et al., 2022).

4 Methodology

In this section, we describe the models and experiments used in this study: the further pre-trained language model, the fine-tuning for the NER task and finally the application of NER.

4.1 Further-Pretraining models

From this study, we further pretrained different transformer-based language models. Further Pretraining (known as continued pretraining or domain-adaptive pretraining) is the process of taking a pretrained language model (such as BERT, RoBERTa, or GPT) and continuing to train it on a new dataset that is specific to a certain domain, or task, before fine-tuning it on a specific downstream task (Lee et al., 2023).

Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimised BERT (RoBERTa) are based on stacked transformer layers with a self-attention mechanism (Vaswani et al., 2017). Specifically, BERT optimises a Masked Language Model (MLM) objective (Devlin et al., 2018), while RoBERTa (Liu et al., 2019) refines the architecture further. MLM plays a vital role in BERT by masking a portion of words in the input text and training the model to predict these masked words using the surrounding words.

Table 1: Updated perplexity values after further pretraining models

Model Name	Training objective	Perplexity
HoloBERT	MLM	3.1259
HoloRoBERTa	MLM	3.8178
HoloXLNet	PLM	8.975

Permutation Language Modelling(PLM) is another approach for language modelling used to overcome pretrain-finetune discrepancy. XLNET (Yang et al., 2019) is a PLM-based language model which is trained to estimate the probabilities of various word permutations within sentences, instead of adhering to fixed left-to-right word order by facilitating more effective modelling of global dependencies and long-range relationships within text.

4.2 Experimental settings for further-pretraining

As outlined in the data section, 30% of the data was set aside as the test set to ensure a balanced dataset distribution. In order to assess the language models' performance, we calculated each model's perplexity (PPL). Table 1 reports the updated perplexity scores for the further pre-trained models trained with a masked language modeling (MLM) objective. We evaluate three further pre-trained models derived from BERT, RoBERTa, and XLNet: HoloBERT, HoloRoBERTa, and HoloXLNet, respectively. Each was fine-tuned using its corresponding base model's optimal hyperparameter set[3].

4.3 Fine-tuning for Domain-Specific NER

To analyse the effectiveness of these further pretrained models, we conducted a downstream domain-specific task, focusing on Named Entity Recognition(NER). We use the general BERT (Bert-base-cased), XLNET models and HmBert, MacBert, dbmz models as benchmarks to evaluate our further pretrained BERT, XLNET and RoBERTA models. Table 2 and Table 3 illustrate the performance of the HIPE Dataset on the NER task compared to existing benchmarks and our best-performing further pretrained BERT model. All the models were experimented with using a batch size of 64, Adam optimiser with a learning rate of 1e-4. They were trained for 3 epochs with linear learning rate warm-up over 10% of the training data. We calculated the weighted average F1 score for B-Tags and I-Tags predicted by the models. Refer to the Table 5 and Table 6 for tag-wise evaluations.

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

Table 2: Weighted Average of the tags in the HIPE dataset

Tags	**HmBERT**	**MACBERT**	**BERT**	**dbmz**	**HoloBERT**
B-tags	0.5135	0.4327	0.7166	0.5124	0.7321
I-tags	0.3800	0.3821	0.6029	0.3509	0.6129

Table 3: Weighted average of tags in Holocaust Dataset

Tags	**HmBERT**	**MACBERT**	**BERT**	**dbmz**	**HoloBERT**
B-Tags	0.62364	0.6212	0.8974	0.9020	0.9026
I-Tags	0.7881	0.7878	0.9025	0.9057	0.9057

After recognising the domain-specific NEs, we assessed the practical applicability of generating knowledge from historical narratives. To achieve that, we perform Named Entity Linking (NEL) to link and associate named entities in text with their corresponding records in a knowledge base via visualising on a graph.

5 Linking with Knowledge Base

NEL assign a unique identity to entities mentioned in the text, corresponding to a knowledge base (Shen et al., 2015). After obtaining the predicted NE tags, we standardised NE tags through preprocessing, ensuring consistency by addressing variations and misspellings. DBpedia [4] is a structured knowledge base extracted from Wikipedia. Utilising DBpedia as our knowledge base, we performed entity linking to map corresponding entities.

$$G \rightarrow (V, E) \tag{1}$$

According to the above graph formulation (G), nodes within the graph symbolise distinct entities (V), while the edges between nodes signify co-occurrence relationships (E). The co-occurrence edges represent connections between entities that appear together in various testimonies, indicating a level of association or co-occurrence among these entities.

From the above identified nodes and edges, we performed extrinsic analysis on the patterns that exist in individual Holocaust narratives, such as entity associations, clusters of related entities, and cross-entity relationships. These insights provide valuable information about the Holocaust as an event. We have removed entities related to dates

[4] https://www.dbpedia.org/

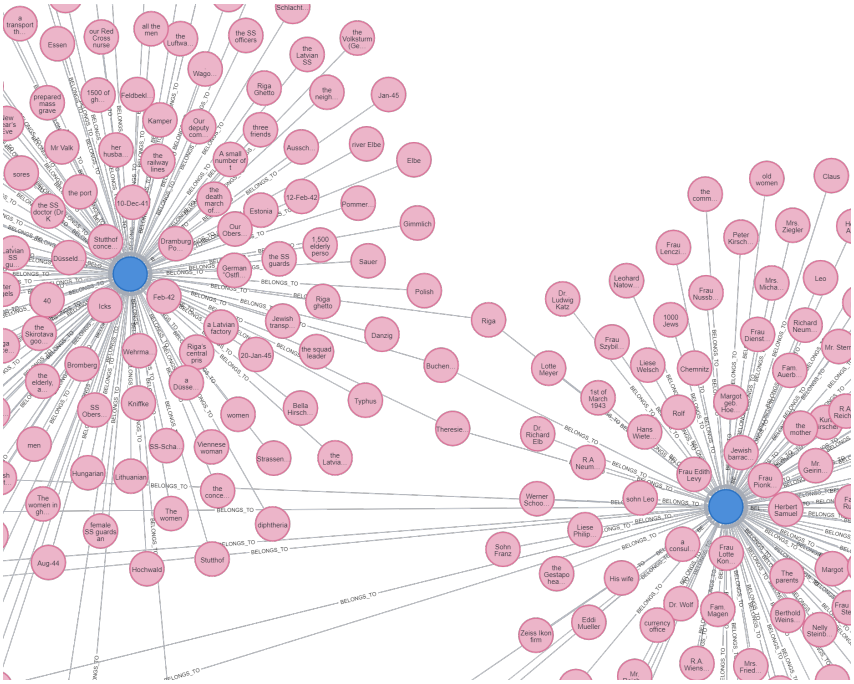

Figure 2: Visual representation of Graph (Blue nodes represent individual narrative, pink nodes represent the entities which **belong to** the narratives.

and times (such as "next day," "tomorrow," "morning," and "8 PM") from the knowledge retrieval process because resolving temporal information is challenging. We are planning to integrate temporal resolution in future with this context.

5.1 Results

In our experiments, the further-pretrained BERT model (HoloBERT) demonstrated lower perplexity and yielded good results in both datasets compared to other benchmark models. Table 4 refers to the unique population of named entities we identified using our further-pretrained models. Apart from that, we evaluated the gold set with a generative language model and observed the performance in the domain-specific NER tasks, for which we used the GPT 3.5 turbo model. For the visualisation task, we have used the Neo4j graph database as shown in Figure 2.

We observed the common entities belonging to different narratives and created a knowledge graph to retrieve information on complex queries combining attributes such as year, location, country, etc. However, the results were not the best, as many directions were discovered to optimise the future graph.

We categorise NEs as biographical, temporal, and geographical entities based on our findings. From the identified entity population, it was observed that the majority of Holocaust narratives predominantly contain temporal NEs. Table 5 refers to the F1 score of each tag. However, due to the inherent characteristics of narrative text, many temporal named entities persist with certain ambiguities, such as *yesterday*, *last year*, *next day*, etc. Most of them were illuminated in the results obtained from the GPT 3.5. Further, recognising toponyms in Holocaust narratives proves to be a challenging task, given the dynamic nature of historical locations and naming conventions. According to our observations, we realised there are ambiguities between geopolitical entities, Locations, Camps and Ghettos. Over time, with the development of language models, the names of many concentration camps and ghettos have increasingly been mapped or translated into contemporary geographic locations. As a solution, we integrated additional domain-specific lexicon and external databases such as `https://www.geonames.org/`. Additionally, we observed that the models were biased in labeling capitalised letters as organisational entities (ORG tags). It was challenging for the model to recognise language-specific terms because these models were primarily trained on English-language data. However, our further-pretrained models demonstrated a better ability to identify these language-specific terms than the standard BERT model.

	General			Temporal			Domain Specific				
Models	GPE	LOC	PERSON	ORG	TIME	DATE	CAMP	GHETTO	STREET	WARSHIP	MILITARY
GPT-3.5	510	521	614	261	138	452	173	47	104	2	105
HoloBert	1863	129	1887	1322	1018	3171	196	26	98	1	57

Table 4: Population of unique named entities

Table 5: Results for Holocaust testimonial dataset (F1-Score)

Tags	**HmBERT**	**MACBERT**	**BERT**	**XLNET**	**HoloBERT**
B-CAMP	0.3812	0.3543	0.8954	0.8781	**0.8980**
B-DATE	0.8915	0.8919	0.9371	**0.9381**	0.9363
B-EVENT	0.4390	0.7652	0.7652	**0.7717**	0.7565
B-GHETTO	0.4211	0.4375	**0.9048**	0.8489	0.8529
B-GPE	0.7876	0.7870	**0.9395**	0.9380	0.9377
B-LANGUAGE	0.7116	0.6764	**0.8841**	0.8625	0.8745
B-LAW	0.5301	0.5195	0.7283	0.7553	**0.8045**
B-LOC	0.2677	0.2615	0.6369	0.6237	**0.6458**
B-MILITARY	0.0000	0.0225	0.7579	0.7225	**0.7513**
B-ORG	0.5843	0.5820	0.8457	**0.8525**	0.8450
B-PERSON	0.5666	0.5594	0.9179	0.9193	**0.9198**
B-RIVER	0.4938	0.5679	0.6237	0.6066	**0.7174**
B-SHIP	0.0000	0.0000	0.4000	**0.5714**	0.3333
B-SPOUSAL	0.5263	0.5641	0.8966	0.9153	**0.9153**
B-STREET	0.0683	0.0361	0.9143	0.9019	**0.9145**
B-TIME	0.8798	0.8760	0.8954	**0.8981**	0.8941
I-CAMP	0.3173	0.2680	0.7895	**0.7900**	0.7778
I-DATE	0.9308	0.9320	0.9476	**0.9505**	0.9491
I-EVENT	0.5007	0.4965	0.7785	**0.8048**	0.7994
I-GPE	0.4099	0.4240	0.7293	**0.7387**	0.7080
I-LAW	0.5793	0.6100	0.7360	0.7957	**0.8109**
I-LOC	0.1948	0.2357	**0.5863**	0.5714	0.5707
I-MILITARY	0.0258	0.0144	0.6759	0.6642	**0.6931**
I-ORG	0.6416	0.6442	0.8320	**0.8410**	0.8310
I-PERSON	0.6435	0.6350	0.9124	0.9108	**0.9124**
I-RIVER	0.4471	0.4889	0.6374	0.6415	**0.7071**
I-SPOUSAL	0.5577	0.5739	0.8143	0.8212	**0.8188**
I-STREET	0.1429	0.1569	0.9444	0.8974	**0.9577**
I-TIME	0.9110	0.9076	0.9204	**0.9228**	0.9198

Table 6: Results for HIPE dataset

Tags	**HmBERT**	**MACBERT**	**BERT**	**dbmz**	**HoloBert**
B-BUILDING	**0.3256**	0.0667	0.2000	0.2857	0.2857
B-LOC	0.6059	0.5031	**0.7932**	0.6004	0.7752
B-PERSON	0.5192	0.3656	0.8321	0.5302	**0.8321**
B-SCOPE	0.4603	0.6408	0.6300	0.5128	**0.6602**
B-STREET	**0.5581**	0.4667	0.0000	0.5263	0.4516
B-WORK	0.3065	0.1887	**0.7320**	0.2679	0.7308
I-BUILDING	0.4918	0.3636	0.4857	0.4545	**0.5352**
I-LOC	0.2105	0.1778	0.3210	0.2857	**0.4000**
I-PERSON	0.0000	0.0000	**0.7500**	0.0000	0.7451
I-SCOPE	0.5417	0.5950	**0.6275**	0.4638	0.5837
I-WORK	0.3490	0.3497	0.7385	0.3129	**0.7660**

6 Discussion and Conclusion

The presented study explores further pretraining models tailored for specific domains, investigating their performance across various downstream tasks in NLP. While further pretraining methods have been applied across different domains, this research uniquely focuses on evaluating the effectiveness of methods when applied to Historical narratives. Moreover, we assess the applicability and performance of these models in the context of historical linguistic content with spoken data. In the preceding sections, as the main contribution, we have presented pre-trained and fine-tuned language

models with their performances for downstream tasks of NER to a diverse set of corpora. As the next step, we developed a sample application where HoloBert was employed. We tried to identify the patterns in individual documents by mapping the information into a graph.

Considering the current limitations in accessing future research opportunities, the model described in this study could be extended to accommodate multilingual settings. In order to construct a more complex graph, the next phase of this research will involve integrating relationships between named entities. This expansion aims to enhance the complexity of the graph, contributing to a more comprehensive understanding of the underlying structures within the data. Moreover, future research will improve the more domain-specific entities and categorising the concepts with respect to the Holocaust, such as Aryanisation, Death Marches, etc.

References

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.

Kaspar Beelen, Federico Nanni, Mariona Coll Ardanuy, Kasra Hosseini, Giorgia Tolfo, and Barbara McGillivray. 2021. When time makes sense: A historically-aware approach to targeted sense disambiguation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2751–2761.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Extended overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, volume 3180. CEUR-WS.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy. 2021. Neural language models for nineteenth-century english. *arXiv preprint arXiv:2105.11321*.

Seanie Lee, Minki Kang, Juho Lee, Sung Ju Hwang, and Kenji Kawaguchi. 2023. Self-distillation for further pre-training of transformers.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Enrique Manjavacas and Lauren Fonteyn. 2021. Macberth: Development and evaluation of a historically pre-trained language model for english (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36.

Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2022. Bertoldo, the historical bert for italian. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.