Towards Semantic Tagging of Segmented Holocaust Narratives

Christopher Brückner and **Pavel Pecina**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{bruckner,pecina}@ufal.mff.cuni.cz

Abstract

With the increasing loss of Holocaust witnesses, it is becoming more and more important to preserve their memories. Items of cultural heritage, including textual data such as diaries or transcripts of video interviews, are abundant. However, large amounts of this data are not annotated, which poses a significant obstacle for domain experts curating digitized information regarding the Holocaust. A solution for this problem is a natural language processing model that links text segments to a rich domain-specific ontology of subject terms to automatically tag documents for further processing. While we have not yet achieved a comprehensive solution, we show that even a simple model fine-tuned on a small dataset of spoken narratives is a promising first step and transfers its capabilities to written testimonies reasonably well.

1 Introduction

In the last decades, great efforts have been made to collect the memories of people who survived the war, in particular those who were directly affected by the Holocaust. One such example is the Visual History Archive (VHA) by the USC Shoah Foundation[1] that comprises more than 55,000 video interviews indexed with ∼68,000 terms, about 12,000 (20%) of which are hierarchically ordered subjects. A different source is the EHRI Online Editions[2], a growing collection of more than 500 digitized documents in at least 12 different languages. These texts are annotated with named entities and document-level subjects but lack segmentation and thus indexing on a segment level.

In fact, the segmentation of such spoken narratives is not trivial: In contrast with structured articles (Koshorek et al., 2018) or scripted video essays (Retkowski and Waibel, 2024), there are no hard boundaries between topics. Instead, they are characterized by smooth transitions between possibly overlapping topics. While this narrative segmentation is a separate challenge not further addressed in this work, it is important to outline that a lack of proper segmentation and topically discriminate text segments negatively affects how well a language model can link these segments to an ontology (Wagner et al., 2022).

The aforementioned VHA is pre-segmented into segments of uniform length rather than by topics, and many of these segments lack annotations. In the following, we ignore this issue and use a small subset of the VHA as training data to evaluate a simple text classification model without re-segmenting the video transcripts. This serves as a first baseline for a future goal: A segmentation and semantic tagging pipeline for domain experts trained on large amounts of Holocaust testimonies.

2 Related Work

While information extraction from historical documents has recently gained traction, most research is focused on Named Entity Recognition (NER). The HIPE-2022 shared task on identifying historical people, places, and other entities (Ehrmann et al., 2022) produced Historical Multilingual BERT (Schweter et al., 2022), a Transformer model pre-trained on 19th century newspapers which has been shown to outperform traditional word embeddings on multiple NER benchmark datasets in the general historical domain. Dermentzi and Scheithauer (2024) fine-tuned XLM Roberta (Conneau et al., 2019) on the EHRI Online Editions and published the first Holocaust-specific NER model for 9 languages which also predicts CAMP and GHETTO labels in addition to standard labels.

Outside NER, the text-to-text Transformer T5 (Raffel et al., 2020) has been fine-tuned on a

[1] https://vha.usc.edu
[2] https://www.ehri-project.eu/ehri-online-editions

synthetic question-answering dataset for conversational speech and applied to machine-generated transcripts of VHA videos (Lehečka et al., 2023; Švec et al., 2024). An online demo is available.[3]

Recent advances in the unsupervised segmentation of manual transcripts of VHA videos are also concerned with the topical classification of the resulting segments. However, they reduce the set of ∼10,000 hierarchical subject terms to 29 flat topics and solve a multi-class rather than a multi-label classification problem, i.e., they only assign one topic to each segment (Wagner et al., 2022).

Rather than ignoring this hierarchical structure, multi-label classification models can use it explicitly and enforce logical constraints on the predictions (Giunchiglia and Lukasiewicz, 2020). Recent such state-of-the-art models specific to text classification encode the hierarchy in graph convolutional networks (Zhou et al., 2020; Chen et al., 2021) or coding trees (Zhu et al., 2023) that use text encoding as inputs, or use a constrained sequence-to-sequence model to predict series of labels along the hierarchy tree (Torba et al., 2024).

3 Dataset

At the time of writing, the full Visual History Archive (VHA) counts 57,058 video testimonies related to the Holocaust and 1,899 more interviews related to other historical events such as the Armenian Genocide. While approximately 10% of these videos are accurately segmented and indexed using a rich hierarchical domain-specific ontology, the majority is segmented uniformly into 1-minute long segments, and only one third of these uniform segments are annotated. The VHA ontology consists of 67,709 hierarchically ordered keywords, about 20% of which are subject terms and the rest being geographical entities.

We are mainly interested in four languages: English, German, Dutch, and Czech. These comprise 31,012 testimonies, i.e., little more than half of all the Holocaust-related testimonies of the VHA. The composition of this subset is described in Table 1. Out of the full ontology of 67,709 unique keywords, these four languages use 34,348 (50.73%). Approximately half of these are subject terms, indicating that most unused keywords are geographical locations while abstract subjects are used more consistently across all languages. On average, the testimonies are 163.38 minutes in length, i.e., they comprise 164 segments, about 55 of which are annotated with 84.69 ± 39.00 different keywords.

Language	Testimonies	Fraction
English	28,457	91.76%
Dutch	1,077	3.47%
German	917	2.96%
Czech	561	1.81%
Total	31,012	100.00%

Table 1: The composition of the Holocaust-related subset of the Visual History Archive in four languages.

Although the ontology has a maximum depth of 8, the majority of used keywords (57.42%) are located on the fifth level of the hierarchy while keywords are rare if they are too general or too specific. This results in a significantly skewed distribution with an immense variance: Individual keywords are used between 1 and 81,472 times across all available testimonies, the arithmetic mean is 118.39 ± 1070.20, and the median count is only 4. This excludes higher-level keywords that are never explicitly used. On average, the annotated segments are indexed with 2.56 ± 1.91 keywords, although there are some outliers with up to 31 annotations. The most common keywords and their rapidly decreasing counts are summarized in Table 2. It is not surprising that these are centered on family while more specific keywords related to individual experiences will follow later in the ranking.

Keyword	Count
extended family menbers	81,742
interviewee photographs (stills)	63,061
family life	43,757
loved ones' fates	33,568
working life	31,304

Table 2: The five most used keywords and their counts across all the Holocaust-related testimonies in English, Dutch, German, and Czech in the VHA.

This is a lot of data: 27 TB of MP4 video interviews, reduced to 4.6 GB of plain text using domain-specific automatic speech recognition models (Lehečka et al., 2023), 2.1 GB of annotations, and 56 MB for the keyword hierarchy with definitions and synonyms.

However, while this will be extensively used for a domain-specific semantic tagging model, most of it was not yet available at the time of the experiments described in Section 4. These experiments

[3] https://malach-aq.kky.zcu.cz

use a tiny set of manual transcriptions that were used as training data for the mentioned ASR model. This set is monolingual and covers merely 0.78% of all English data. In total, it contains only 2,115 annotated 1-minute segments indexed with 1,063 different subject terms which is little training data for this amount of labels.

4 Experiments

Four different Transformer models are first trained and evaluated on annotated segments of the small English VHA subset described in Section 3. The models are trained on a multi-label classification task using the flattened subject hierarchy as labels, i.e., they do not make use of any structural information and do not predict geographical entities which are, in conjunction with people's names, better handled by named entity recognition models.

To mitigate the segmentation problem, the input text is a window encompassing one previous and one subsequent text segment in addition to the target segment. In case of a topical overlap, i.e., at least one of the additional segments is annotated, the sets of labels are merged. The training splits are randomized using 20% of the data as the test set and 20% of the remaining data as the development set. The candidate models are the following:

1. BERT (Devlin et al., 2018)
   Since the training dataset is English only, the uncased version of BERT is used as an English-focused baseline.

2. Multilingual BERT (mBERT)
   Since four languages are targeted, the cased version of mBERT is evaluated to compare the performance of a multilingual variant on English data with monolingual BERT.

3. Historical Multilingual BERT (hmBERT) (Schweter et al., 2022)
   hmBERT has been pre-trained on 19th century Europeana[4] newspapers (German, French, Finnish, Swedish) and British Library[5] books published between 1510 and 1900.

4. LaBSE (Feng et al., 2022)
   Language-agnostic BERT Sentence Embedding is a multilingual sentence Transformer (Reimers and Gurevych, 2019) and thus better suited to encode the meanings of sentences.

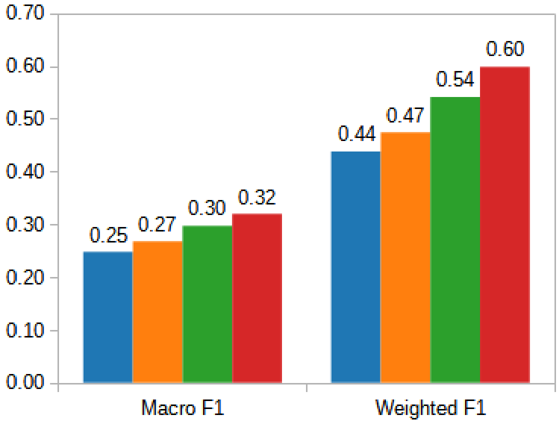

Figure 1: Macro and weighted classification scores of four different Transformer models. From left to right: BERT, mBERT, hmBERT, LaBSE.

All of these models are fine-tuned for 100 epochs using a linearly decreasing initial learning rate of 2e-5 and then rolled back to the checkpoint with the best F_1 score which is usually at around 98 epochs. The model weights are optimized using Adam without weight decay and binary cross-entropy loss for multi-label classification, which can be interpreted as training independent binary classifiers for all 1,063 present subject terms on top of the language model weights. These labels are binarized, i.e., each annotation is represented by a sparse 1,063-dimensional binary vector where a 1 signifies that the input text is annotated with the corresponding indexed keyword from the VHA ontology.

Results are shown in Figure 1. The macro-averaged F_1 scores do not take the number of subject term occurrences into account and are rather low, which was expected considering the little training data (2,115 segments) compared with the number of predicted labels (1,063). However, the weighted averages being approximately twice as high as the macro averages indicates that frequent subjects are handled more reasonably well and we expect improvements when adding more training data with more occurrences of all subject terms.

Surprisingly, mBERT outperforms English BERT on the English-only data. This is possibly due to a significant amount of non-English entities such as people and locations present in historical text. This is further emphasized by hmBERT, even though the period of its training data predates the period of interest. Finally, LaBSE achieves the best results despite being domain-agnostic, likely due to it being focused on encoding whole sentences which can help identify more abstract sub-

[4] http://www.europeana-newspapers.eu
[5] https://labs.biblios.tech

1	ghetto living conditions
2	ghetto-related aid giving
3	sustenance provision
4	ghetto forced labor
5	camp forced labor
6	ghetto housing conditions
7	anti-Jewish roundups
8	beatings
9	Poland 1941 (June 21) - 1945 (May 7)
10	loved ones' fates
11	Poland 1941 (June 21) - 1944 (July 21)
12	ghettoization
13	deportation awareness
14	German soldiers
15	ghetto selections

Table 3: The top 15 keywords predicted by LaBSE for an English testimony excerpt published by DEGOB.

1	ghetto-related aid giving
2	sustenance provision
3	ghetto forced labor
4	loved ones' fates
5	beatings
6	ghetto living conditions
7	anti-Jewish measures
8	Poland 1939 (Sept 1) - 1945 (May 7)
9	anti-Jewish roundups
10	Hungary 1944
11	German soldiers
12	deportation awareness
13	ghettoization
14	shootings
15	Romania 1944

Table 4: The top 15 keywords predicted by LaBSE for a Hungarian testimony excerpt published by DEGOB.

jects. This model is available on Hugging Face[6] but requires the VHA ontology to disambiguate the labels as it only predicts keyword IDs.

Finally, the fine-tuned model is manually evaluated on a testimony published by the Hungarian National Committee for Attending Deportees (DEGOB) in English[7] and Hungarian[8], the latter one being the original text in a language not present in our training data. Since the model can only process input texts no longer than 512 tokens, only the first 26 sentences are used. In this excerpt, a Jewish survivor from Munkács, Hungary (today Mukachevo, Ukraine) describes their ghetto experiences in May 1944 before being deported by train to Auschwitz, including beatings by German soldiers, forced labor for men, and the existence of a soup kitchen in the ghetto. The top 15 predicted keywords ranked by their likelihoods as predicted by the model and limited to exhibit diversity, are listed in Table 3 and Table 4 for the English and Hungarian texts, respectively.

Clearly, only a few of the keywords extracted from the English translation are wrong: There is no mention of aid giving, the forced labor does not happen in a camp, the fate of the parents is not mentioned, and Poland (more specifically: Auschwitz) is only mentioned after the excerpt. The biggest problem is not shown in the table: In practice, the model would classify none of the keywords as true. Their prediction scores range from 0.2 to 0.4 and thus all lie below a standard threshold of 0.5. Evidently, even a simple Transformer is able to learn and identify these abstract concepts but requires more training data to do so reliably.

Interestingly, even though only English data was available during the fine-tuning, using a multilingual pre-trained Transformer as its basis appears to be sufficient to give the model multilingual capabilities. While the order of the keywords extracted from the Hungarian testimony is slightly different and Romania 1944 has been incorrectly added to the list, the keywords are generally similar and Hungary 1944 and shootings have been additionally identified as correct subject terms. This is likely due to LaBSE having been trained on pairs of equivalent sentences in different languages, and cannot be generally assumed for the other three models that achieved lower scores on the test data.

Furthermore, most of the ghetto-related keywords which have the same parent "ghetto experiences" which already describes the testimony well and would have probably been predicted as a true label by the model, had this annotation been used instead of the less frequent and more fine-grained labels. This also raises the question what granularity is actually desired, and the answer to this question depends on the user. In the end, a likely erroneous prediction can often be resolved to a correct more general one: In the given hierarchy, there is always a path from the best general keyword to the best granular one. In addition to using more training data, this concept can be used to devise

[6] https://huggingface.co/ChrisBridges/labse-malach-multilabel

[7] http://degob.org/index.php?showjk=2

[8] http://degob.hu/index.php?showjk=2

more complex model architectures and hopefully achieve better performance on data with a much greater set of labels, as described in Section 3.

5 Conclusion

We presented one of the largest available annotated Holocaust-related datasets, the Visual History Archive, and showed that even a simple multi-label classification model trained on little data can produce promising results despite achieving rather low F_1 score due to the rarity of many fine-grained labels. There are different ways to address this issue:

- Using substantially more data
- Re-segmenting the data to align the segments with the annotations semantically instead of using uniform-length segments
- Using a more complex model that incorporates the hierarchical structure of the label ontology in its predictions.

We will explore these options in the near future and believe that we are on the right track to providing an efficient semantic tagging tool for domain experts.

While we cannot publish the whole described dataset due to licensing issues, we further plan to prepare a representative subset for open research and advance computational approaches in Holocaust research and similar domains related to the digital humanities and digital cultural heritage.

Limitations

This study describes preliminary research results based on a small amount of data and focuses on future prospects. The summarized dataset cannot be published in its entirety which means that the exact results cannot be reproduced. However, the presented semantic tagging model is openly available and representative datasets can be published in the future to ensure as much reliability and reproducibility as possible.

Acknowledgments

This research was partially supported by SVV project number 260 698 and Horizon Europe grant agreement number 101061016. We thank the University of Southern California and the University of West Bohemia for their cooperation and providing the data, and the reviewer for the valuable feedback.

References

Haibin Chen, Qianli Ma, Zhenxi Lin, and Jiangyue Yan. 2021. Hierarchy-aware label semantics matching network for hierarchical text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4370–4379, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Maria Dermentzi and Hugo Scheithauer. 2024. Repurposing holocaust-related digital scholarly editions to develop multilingual domain-specific named entity recognition tools. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 18–28, Torino, Italia. ELRA and ICCL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Eleonora Giunchiglia and Thomas Lukasiewicz. 2020. Coherent Hierarchical Multi-Label Classification Networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9662–9673. Curran Associates, Inc. MCM, MCLoss.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.

Jan Lehečka, Jan Švec, Josef V. Psutka, and Pavel Ircing. 2023. Transformer-based Speech Recognition Models for Oral History Archives in English, German, and Czech. In *Proc. INTERSPEECH 2023*, pages 201–205.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Fabian Retkowski and Alexander Waibel. 2024. From text segmentation to smart chaptering: A novel benchmark for structuring video transcriptions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 406–419, St. Julian's, Malta. Association for Computational Linguistics.

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

Fatos Torba, Christophe Gravier, Charlotte Laclau, Abderrhammen Kammoun, and Julien Subercaze. 2024. A Study on Hierarchical Text Classification as a Seq2seq Task. In *Advances in Information Retrieval*, pages 287–296, Cham. Springer Nature Switzerland. T5 + Constrainer.

Eitan Wagner, Renana Keydar, Amit Pinchevski, and Omri Abend. 2022. Topical segmentation of spoken narratives: A test case on holocaust survivor testimonies. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6809–6821, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. 2020. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117, Online. Association for Computational Linguistics.

He Zhu, Chong Zhang, Junjie Huang, Junran Wu, and Ke Xu. 2023. HiTIN: Hierarchy-aware tree isomorphism network for hierarchical text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7809–7821, Toronto, Canada. Association for Computational Linguistics.

Jan Švec, Martin Bulín, Adam Frémund, and Filip Polák. 2024. Asking Questions Framework for Oral History Archives. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval*, volume 14610, pages 167–180. Springer Nature Switzerland, Cham.