Multilingual Automated Subject Indexing: a comparative study of LLMs vs alternative approaches in the context of the EHRI project

Maria Dermentzi [1], Mike Bryant [1, 2], Fabio Rovigo [3], and Herminio García-González [4]

[1]King's College London, UK
[2]NIOD Institute for War, Holocaust and Genocide Studies, NL
[3]Vienna Wiesenthal Institute for Holocaust Studies, AT
[4]Kazerne Dossin, BE

1 Background

Since its launch in 2010, the European Holocaust Research Infrastructure (EHRI) has sought to catalyse transnational Holocaust research by making information about dispersed archival material more interconnected and accessible. Due to the international character of the Holocaust, the displacement of survivors, and the dispersal of material by Allied occupying forces, Holocaust-related documents and objects are scattered across the world. The EHRI Portal[1] (Blanke et al., 2017) attempts to integrate archival metadata from these dispersed sources into a single framework within which fragmented collections can be browsed, searched, connected, and contextualised.

The process of aggregating and integrating archival descriptions from institutions around the world poses considerable institutional, social, and technical challenges. In the context of Holocaust-related archives, institutions take a wide range of approaches to describing their materials (Erez et al., 2020; García-González and Bryant, 2023; Rodriguez et al., 2016). The use of index terms—controlled vocabularies of subject headings, people, organisations, and places—is a cornerstone technique for the discovery and retrieval of archival material. Nonetheless, the hundreds of institutions holding Holocaust-related material share little commonality in the application of index terms or the use of common controlled vocabularies. Where index terms are used, an in-house vocabulary is the most common approach. While a minority of institutions do use general-purpose subject heading thesauri such as the Library of Congress Subject Headings[2], they do not necessarily apply the index terms in consistent or interoperable ways (Erez et al., 2020). EHRI set out to mitigate this lack of a common vocabulary

[1] https://portal.ehri-project.eu/
[2] https://www.loc.gov/aba/publications/FreeLCSH/LCSH44-Main-intro.pdf

for Holocaust-related material through the creation of the *EHRI Terms* vocabulary[3], a hierarchically organised, multilingual set of subject headings, which at the time of writing consists of 913 terms translated in 12 languages[4].

While EHRI has built a robust data integration infrastructure, improving the discoverability and interconnectedness of multilingual collection metadata sourced from multiple institutions remains in focus. The harmonisation of subject headings, using the EHRI Terms vocabulary as an integration point, is one particular area where there is room for improvement, not only because much of the aggregated metadata does not currently align with it but because many archival collections lack access points entirely. At present, 25% of collection-level descriptions have no access points at all[5]. Of those that do have access points, only 30% have subject terms aligned with the EHRI Terms vocabulary[6].

Past work has described how vocabularies in use by a selection of partner institutions were systematically co-referenced to EHRI Terms (Erez et al., 2020). Our paper, however, investigates whether archival descriptions in the EHRI Portal could be made more discoverable and interlinked at scale through automated subject indexing, i.e., the use of machine-based methods, such as computational linguistics and statistics, to perform the subject indexing steps typically performed by human indexers (Golub, 2021). The findings of this investigation are applicable beyond EHRI and its partners, as they could help other institutions address similar challenges in their specific contexts[7].

2 Automated Subject Indexing

Approaches to automated subject indexing vary based on the purpose of the application and the field from which each approach originates (Golub, 2021). The two prevailing approaches are the statistical associative and lexical[8] ones (Suominen and Koskenniemi, 2022; Toepfer and Seifert, 2020). The statistical associative approach involves Multilabel Text Classification (MTC) methods, where a supervised Machine Learning (ML) model is trained on the already indexed texts of a collection. The model learns weights that are used to predict the correct set of terms given the text of a document. Lexical approaches, on the other hand, employ string-matching to match terms in the controlled vocabulary with words in the text of the record's description using similarity measures (Golub, 2021). Recent research has also suggested fusion approaches that combine statistical and lexical methods using ensemble techniques (Suominen and Koskenniemi, 2022; Toepfer and Seifert, 2020). Additionally, the emergence of Large Language Models (LLMs) has enabled a novel approach: transfer learning using zero-shot classification (Zhang et al., 2023). A pre-trained LLM can be used out of the box to predict suitable terms from a list of candidate terms derived from a controlled vocabulary without needing fine-tuning on domain-specific data. Presently, the zero-shot classification approach is commonly used as a data augmentation method to overcome class imbalance or the lack of training examples (Møller

[3] `https://portal.ehri-project.eu/vocabularies/ehri_terms`

[4] `English,Hebrew,Italian,Dutch,Russian,Ukrainian,Czech,Hungarian,French,Polish,Serbo-Croatian,andGerman.Consultedon24thJan2024.`

[5] Accessed 29th Jan 2024: `https://portal.ehri-project.eu/api/datasets/E136TY2zwL`

[6] Accessed 29th Jan 2024: `https://portal.ehri-project.eu/api/datasets/CtaJXtPuZA`

[7] This is especially the case given EHRI's use of archival standards, e.g., ISAD(G), and the fact that the tools used by the authors are open-source, meaning that others can reuse and adapt them for their use cases.

[8] Lexical approaches are also known as string-matching or rule-based approaches (Golub, 2021).

et al., 2023; Van Nooten and Daelemans, 2023).

3 Methodology

Our paper offers a comparative evaluation of each of the aforementioned techniques (i.e., statistical, lexical, fusion, and transfer learning), focusing particularly on the potential role of LLMs in developing reliable automated subject indexing tools. Our experiments start with baseline LLMs before we invest more resources into larger-scale experiments with state-of-the-art models. In particular, we treat the metadata on the EHRI Portal-ingested archival descriptions (including their associated subject terms that are matched with terms in the EHRI Terms vocabulary) as training material for ML algorithms. We then fine-tune an open-source LLM, *BERT-base Multilingual Cased* (Devlin et al., 2019), for domain-specific MTC. We compare this fine-tuned model with predictions made by another open-source multilingual LLM-based text classifier, *mDeBERTa-v3-base-mnli-xnli* (Laurer et al., 2023), which we use in a zero-shot setup. We investigate how these newer LLM-based methods fare when compared to other types of tools by employing Annif (Suominen, 2019; Suominen et al., 2022, 2023; Suominen and Koskenniemi, 2022). Annif comprises a framework that can integrate a variety of approaches to automated subject indexing: from classical information retrieval methods, such as TF-IDF (Spärck Jones, 1972), to classification algorithms, such as Parabel (Prabhu et al., 2018) and fusion approaches that employ multiple classifiers in an ensemble.

4 Results

We evaluate these approaches quantitatively and qualitatively based on their performance on a small but representative sample of 167 archival descriptions. We find that Annif's Neural Network Ensemble (NN Ensemble) classifier achieves the highest F1 document average and F1 subject average scores[9]. The fine-tuned BERT model achieves the highest micro average F1 score and is the least prone to erroneous or neutral predictions. Given that this is a fairly lightweight 109M-parameter LLM, this result leads us to hypothesise that given a more balanced dataset (i.e., where rare subject terms would be adequately represented), we could fine-tune a multilingual Transformer-based model that would be sufficiently reliable. However, we observe that the models trained or fine-tuned on EHRI's metadata—mirroring the label distribution in the dataset they were trained on—tend towards broader, less specific terms.

Nevertheless, our qualitative evaluation suggests that the potential of the zero-shot model is not fully captured by the quantitative evaluation, where it achieves very low scores. Quite consistently, the output of the zero-shot model included labels that were deemed accurate. This indicates that the zero-shot model can be more creative and useful in matching archival descriptions with more fine-grained terms that are not adequately represented in our dataset, provided that it serves as an assistive tool, with a human expert verifying its output to counter the much larger probability of false positive terms.

[9] For more details on the metrics used, see `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html` and `https://github.com/NatLibFi/Annif/blob/main/annif/eval.py`

5 Conclusion

In conclusion, although some of the tools examined were more favourably evaluated compared to the ground truth terms based on our qualitative evaluation, the increased likelihood of incorrect output means that they can only be considered as part of a semi-automated indexing pipeline requiring expert oversight. If EHRI were to deploy a term-suggesting tool at present, the NN Ensemble model would be an optimal choice given its sufficiently high scores. Furthermore, as a welcome corollary, this model would have a lower environmental impact. However, the ability of the zero-shot model to correctly predict fine-grained labels confirms that it can be considered for data augmentation purposes to balance our training dataset. Once a more balanced dataset has been created, fine-tuning an LLM could be the way towards developing a tool that would be more reliable than those currently available. Overall, this research charts a promising course towards the development of ML-powered, multilingual tools that would help archival institutions streamline and scale subject indexing, semantically enriching their metadata according to FAIR Data Principles (Wilkinson et al., 2016). It would thereby be easier for EHRI to ingest new access points, improving contextualisation and interoperability within the overall landscape of Holocaust-related archival material (García-González and Bryant, 2023).

References

Blanke, T., Bryant, M., Frankl, M., Kristel, C., Speck, R., Daelen, V. V., and Horik, R. V. (2017). The European Holocaust Research Infrastructure Portal. *Journal on Computing and Cultural Heritage*, 10(1):1:1–1:18.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Erez, S. A., Blanke, T., Bryant, M., Rodriguez, K., Speck, R., and Daelen, V. V. (2020). Record linking in the EHRI portal. *Records Management Journal*, 30(3):363–378.

García-González, H. and Bryant, M. (2023). The Holocaust Archival Material Knowledge Graph. In Payne, T. R., Presutti, V., Qi, G., Poveda-Villalón, M., Stoilos, G., Hollink, L., Kaoudi, Z., Cheng, G., and Li, J., editors, *The Semantic Web – ISWC 2023*, Lecture Notes in Computer Science, pages 362–379, Cham. Springer Nature Switzerland.

Golub, K. (2021). Automated Subject Indexing: An Overview. *Cataloging & Classification Quarterly*, 59(8):702–719.

Laurer, M., Atteveldt, W. v., Casas, A., and Welbers, K. (2023). Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis*, pages 1–17.

Møller, A. G., Dalsgaard, J. A., Pera, A., and Aiello, L. M. (2023). Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks.

Prabhu, Y., Kag, A., Harsola, S., Agrawal, R., and Varma, M. (2018). Parabel: Partitioned Label Trees for Extreme Classification with Application to Dynamic Search Advertising. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, pages 993–1002, Lyon, France. ACM Press.

Rodriguez, K. J., Alexiev, V., Brazzo, L., Riondet, C., Gherman, Y., and Speck, R. (2016). EHRI-2 - D.11.2 Road Map Domain Vocabularies. Deliverable GA no. 654164. Issue: GA no. 654164.

Spärck Jones, K. (1972). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*, 28(1):11–21. Publisher: MCB UP Ltd.

Suominen, O. (2019). Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 29(1):1–25.

Suominen, O., Inkinen, J., and Lehtinen, M. (2022). Annif and Finto AI: Developing and Implementing Automated Subject Indexing. *JLIS.it*, 13(1):265–282.

Suominen, O., Inkinen, J., Virolainen, T., Fürneisen, M., Kinoshita, B. P., Veldhoen, S., Sjöberg, M., Zumstein, P., Neatherway, R., and Lehtinen, M. (2023). Annif.

Suominen, O. and Koskenniemi, I. (2022). Annif Analyzer Shootout: Comparing text lemmatization methods for automated subject indexing. *The Code4Lib Journal*, (54).

Toepfer, M. and Seifert, C. (2020). Fusion architectures for automatic subject indexing under concept drift. *International Journal on Digital Libraries*, 21(2):169–189.

Van Nooten, J. and Daelemans, W. (2023). Improving Dutch Vaccine Hesitancy Monitoring via Multi-Label Data Augmentation with GPT-3.5. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 251–270, Toronto, Canada. Association for Computational Linguistics.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

Zhang, S., Wu, M., and Zhang, X. (2023). Utilising a Large Language Model to Annotate Subject Metadata: A Case Study in an Australian National Research Data Catalogue. arXiv:2310.11318 [cs].