Acta Informatica Pragensia
2025, Volume 14, Issue 2, 207–214
Special Issue: Theory and Practice of Generative Artificial Intelligence Usage
https://doi.org/10.18267/j.aip.268

PRAGUE UNIVERSITY OF ECONOMICS AND BUSINESS

Article Open Access

Exploring Oral History Archives Using State-of-the-Art Artificial Intelligence Methods

Martin Bulín, Jan Švec, Pavel Ircing, Adam Frémund, Filip Polák

Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia in Pilsen, Pilsen, Czech Republic

Corresponding author: Pavel Ircing (ircing@kky.zcu.cz)

Editorial Record

First submission received:
February 2, 2025

Revision received:
April 23, 2025

Accepted for publication:
May 20, 2025

Special Issue Editors:
David Chudan
Prague University of Economics and Business, Czech Republic

Miroslav Vacura
Prague University of Economics and Business, Czech Republic

This article was accepted for publication by the Special Issue Editors upon evaluation of the reviewers' comments.

How to cite this article:
Bulín, M., Švec, J., Ircing, P., Frémund, A., & Polák, F. (2025). Exploring Oral History Archives Using State-of-the-Art Artificial Intelligence Methods. *Acta Informatica Pragensia*, 14(2), 207–214. https://doi.org/10.18267/j.aip.268

Abstract

Background: The preservation and analysis of spoken data in oral history archives, such as Holocaust testimonies, provide a vast and complex knowledge source. These archives pose unique challenges and opportunities for computational methods, particularly in self-supervised learning and information retrieval.

Objective: This study explores the application of state-of-the-art artificial intelligence (AI) models, particularly transformer-based architectures, to enhance navigation and engagement with large-scale oral history testimonies. The goal is to improve accessibility while preserving the authenticity and integrity of historical records.

Methods: We developed an asking questions framework utilizing a fine-tuned T5 model to generate contextually relevant questions from interview transcripts. To ensure semantic coherence, we introduced a semantic continuity model based on a BERT-like architecture trained with contrastive loss.

Results: The system successfully generated contextually relevant questions from oral history testimonies, enhancing user navigation and engagement. Filtering techniques improved question quality by retaining only semantically coherent outputs, ensuring alignment with the testimony content. The approach demonstrated effectiveness in handling spontaneous, unstructured speech, with a significant improvement in question relevance compared to models trained on structured text. Applied to real-world interview transcripts, the framework balanced enrichment of user experience with preservation of historical authenticity.

Conclusion: By integrating generative AI models with robust retrieval techniques, we enhance the accessibility of oral history archives while maintaining their historical integrity. This research demonstrates how AI-driven approaches can facilitate interactive exploration of vast spoken data repositories, benefiting researchers, historians and the general public.

Index Terms

AI; Oral history archives; Transformer-based models; Machine learning in digital humanities.

1 INTRODUCTION

In the realm of cultural heritage, the preservation and analysis of spoken data, particularly in oral histories such as Holocaust testimonies, represent a rich and complex reservoir of knowledge. This expansive amount of spoken data not only serves as a testament to our collective past but also presents a unique opportunity for advancing computational methods. Particularly, in the context of transfer learning, these extensive datasets could be pivotal in enhancing self-supervised learning algorithms.

This paper delves into the innovative application of pre-trained—mostly generative—models for efficiently navigating and extracting insights from such voluminous archives. The algorithms described in this paper are rather general and well-transferable to other domains, yet we have developed them around one particular oral history archive, the Visual History Archive collected and preserved by the USC Shoah Foundation — The Institute for Visual History and Education[1]. We will refer to this archive as SFI-VHA in this paper. This archive was founded by the director Steven Spielberg, who had established the Survivors of the Shoah Visual History Foundation (VHF) in 1993. The foundation hired a team of field workers who recorded roughly 52 thousand interviews with Holocaust survivors between 1994 and 1999. The interviews were conducted in 32 languages and contained over 116 thousand hours of audio material; roughly half of this material is in English, a significant portion of interviews was also given in Russian and Hebrew and non-negligible amounts in German, Czech and Slovak. In our work presented here, we restricted ourselves to the English portion of the archive.

The research, which is currently crowned by the techniques described in this article, began in 2001, when the MALACH project funded by the US National Science Foundation started[2]. The original idea of the project was to use automatic speech recognition (ASR) for converting the audio track of the video recordings into text and then perform the search using text-based information retrieval (IR) techniques on this textual representation. How the originally clear-cut boundary between ASR and IR engines has been gradually blurred in order to achieve better search results is briefly described in the following section, including references to relevant research papers for interested readers.

In this article, we outline how we employ advanced AI models to improve navigation and interaction with the lengthy monologues typical of oral history archives. Our novel method, which utilizes neural networks based on the transformer architecture, not only makes it easier to find one's way through extended testimonies but also shifts the listening process from a passive activity to an interactive one. By automatically generating questions that are relevant to the context, our system enhances the monologue interviews, helping listeners orient themselves within the story and pinpoint important sections. These questions are crafted to support comprehension without changing the original intent of the testimony, thus upholding the authenticity of the historical account. Furthermore, this approach enables users to engage more actively with the material by asking their own questions, encouraging a dynamic exploration of the rich narratives found in the archives.

1.1 Literature review

First, we would like to briefly review the approaches – and related papers – that were employed by our team since the beginning of the MALACH project. When developing the ASR engines, state-of-the-art approaches were used. Since the first systems date as far back as 2002, the methods — and results — then were naturally quite different from the state-of-the-art approach nowadays. During the course of the project, the ASR system for English (developed at IBM) and Czech (developed in our lab) still used the Gaussian mixture models/hidden Markov model (GMM-HMM) paradigm. The WER (word error rate) of the ASR systems developed within the MALACH project had reached 39.60% for English (Byrne et al., 2004) and 38.57% for Czech (Psutka et al., 2005) by the end of the project in 2006. After finishing the MALACH project, we continued with improving the Czech ASR, predominantly by using speaker adaptation methods, and in 2011, a WER of 27.11% was achieved (Psutka et al., 2011). In 2013, we also developed our first English ASR system for the SFI-VHA dataset. Since then, we have published many papers describing our efforts towards improving the ASR performance for both Czech and English. Our most recent systems based on state-of-the-art transformer-based wav2vec models achieve WER as low as 12.88% for English and 8.43% for Czech (Lehečka et al., 2023).

As for the search (or IR) part, the original idea for search in the SFI-VHA was that the user would specify a complex query (an example can be found in Ircing & Müller, 2006) and the system would process the query and return the a of "starting points" in the video recording where the discussion about the queried topics starts. However, in the very beginning of the research in this area, the oversimplifying approach was taken — the problem of searching speech was reduced to a classic document-oriented retrieval by using only the one-best ASR output and artificially creating "documents" by sliding a fixed-length window across the resulting text stream (Ircing et al., 2008). This approach

[1] See, https://sfi.usc.edu/
[2] See, https://malach.umiacs.umd.edu/

was later dropped, as well as the necessity for specifying complex structured queries — instead, we adopted a Google-like search approach where the user enters only a word or a short phrase and obtains pointers to the audio signal where this query is found (i.e., we moved from document-oriented IR to spoken term detection — STD). In the first versions of the system, the search algorithm was designed somehow empirically (Psutka et al., 2011); later, we gradually started utilizing more sophisticated models such as weighted finite state transducers (Vavruška et al., 2013) and recently of course various neural architectures (Švec et al., 2017; Švec et al., 2018; Švec et al., 2021; Švec et al., 2022a). Since the evaluation metrics for the search performance are not as straightforward and easy to explain as the WER reported above, we refer the reader to the articles cited in this paragraph for further details. Let us just mention here that the performance of the search module has also significantly increased with more sophisticated methods.

Our novel approach, which extends beyond standard spoken term detection (STD), draws significant inspiration from the recently introduced Doc2Query method (Gospodinov et al., 2023). Originally proposed as a query expansion technique for information retrieval from text sources (He & Ounis, 2009), Doc2Query has also proven useful for evaluating machine translation quality (Krubiński et al., 2021). Given that spoken interviews largely consist of spontaneous speech lacking clear grammatical structure, we needed to adapt and develop methods suitable for this context. Along these lines, Yao et al. (2012) presented a strategy for generating conversational agents from text articles. Their approach, which utilizes question generation tools such as Question Transducer and OpenAryhpe, could be adapted to build interactive agents capable of navigating and interpreting complex text documents and reports. Such agents would be able to understand and respond to user queries in a contextually appropriate way, providing a more intuitive means of accessing and comprehending unstructured textual information. Additionally, Wang et al. (2021) introduced the ArchivalQA dataset, which focuses on temporal question answering in news archives. This dataset, with its emphasis on resolving temporal ambiguities and integrating diverse data sources, addresses challenges similar to those encountered when parsing and understanding oral history archives.

2 RESEARCH METHODS

We introduced a novel method that generates a new question-answer structure over existing interview transcripts (Švec et al., 2024). These questions are time-aligned and indexed with the audio, enabling users to efficiently navigate to relevant segments of the interview. They serve as a supplement to the original interviewer, particularly enhancing sections where only the interviewee is speaking. Essentially, the questions act as "open-set topics" related to the content of the testimony. Crucially, the meaning of the testimony remains intact, as each question is directly linked to and answered by excerpts from the original interview.

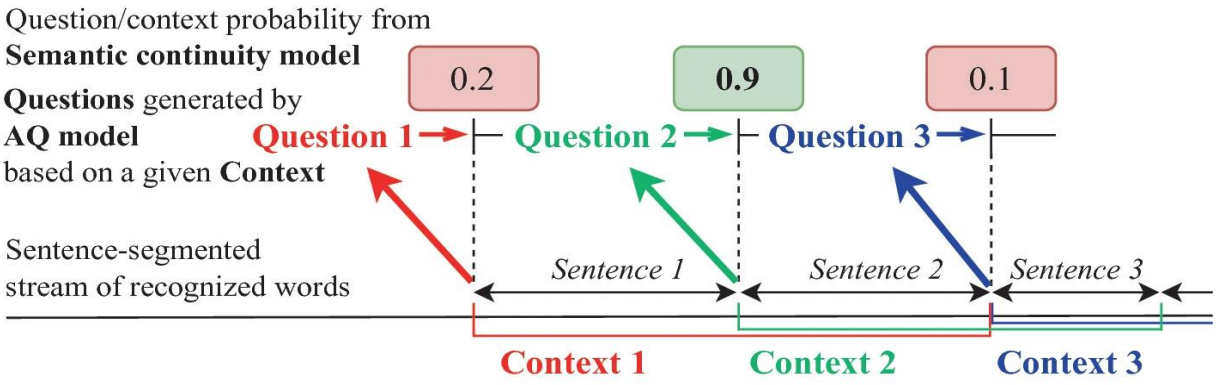

Figure 1. *Processing in the AQ pipeline.*

To generate questions from interview transcripts, we begin by applying a sentence-level sliding window to the sequence of recognized words from the raw audio (Figure 1). This window spans several sentences, providing the contextual input for the question generation process. For each context window, a T5-based **asking questions** (AQ) model produces a candidate question related to the given passage. To encourage the model to generate more targeted questions, we also generate corresponding answers. However, since the T5 model always outputs a question-answer

pair regardless of the informativeness of the context — including cases where the context is trivial or lacks any meaningful information (e.g., mere discourse markers) — we employ a second model to predict the **semantic continuity** (SC) between the generated question and the context. This ensures that only questions which meaningfully precede the context are shown to the user.

2.1 Asking questions (AQ) model

Generating questions for a specific context is essentially the inverse task of question answering. Since this is a text-to-text task, we employ the T5 (text-to-text transfer transformer) model pre-trained by Raffel et al. (2020) on a self-supervised text-restoration task using English Common Crawl web text data.

For fine-tuning, we used the freely available T5-base model pre-trained by Google. Since the T5 architecture is fixed, its performance depends entirely on the quality of the fine-tuning dataset. Given the success of T5 in question answering, as in UnifiedQA (Khashabi et al., 2020), we extended it to the inverse task of question generation. In this setup, the model takes a given context as input and generates a corresponding question. However, during our experiments, we observed that when trained to generate a single question, the model often produced too broad or non-specific questions. To improve question specificity and relevance, we included the required answer in the training target. This approach resulted in higher-quality, topic-specific questions with answers that were directly inferred from the input context.

Our fine-tuning process began with the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2018), a widely used Q&A dataset derived from Wikipedia articles. However, while SQuAD 2.0 effectively supports factual question generation from structured descriptive texts, it does not capture the nuances of spontaneous spoken interviews, such as those found in oral history archives of USC-SFI.

To bridge this gap, we utilized OpenAI's ChatGPT (GPT-3.5 Turbo), the most advanced available model at the time of dataset creation. We provided ChatGPT with passages from the interviews sampled from a proxy dataset and instructed it to generate relevant question-answer pairs based on the given context. The following system prompt was used to guide ChatGPT in generating appropriate questions and answers (see the example below):

- **System prompt:** *You are a helpful assistant. Your task is to generate factual questions based on a provided interview context. You should aim to generate 1 to 3 general questions that can be truthfully and reasonably answered from the given context. In case the answer is not available in the context or is not mentioned in the interview, label it as '<not-known>'. For each generated question, please provide a straightforward answer based on the given context.*
- **Context:** *I had enough money left over to basically pay for food and whatnot on my bicycle journey across America. And the path that I had to visit all my brothers and sisters was not a direct route, going from San Francisco to New York. I actually had to go up to Idaho and back down to Texas and then back up through Indiana. So it was a 5,000 mile trip. The day which, coincidentally, was exactly six months from when I had this assignment, was October 31. It was Halloween. And so the plan would be that I would ride back home, so that I would come back to die on the day after Halloween.*
- **Output:** *How long was the bicycle journey across America? – The bicycle journey across America was 5,000 miles long. – When did the interviewee plan to come back home? – The interviewee planned to come back home on the day after Halloween. – Did the interviewee have enough money to pay for food during the journey? – Yes, the interviewee had enough money left over to pay for food and whatnot during the journey.*

Since USC-SFI Holocaust testimonies contain sensitive personal information, we could not use them directly for training data generation. Instead, we sought a comparable dataset that captures spontaneous spoken conversations. We used the This American Life Podcast (TALP) (Mao et al., 2020) dialog transcript dataset as a proxy for real interview transcripts. This dataset consists of transcribed spoken conversations from various episodes of TALP, making it a suitable alternative for training a model on conversational speech.

This synthetic dataset, consisting of over 15,000 passages with multiple questions per passage, was then used to fine-tune our AQ model. By training on both datasets, we observed a significant improvement in the specificity and fluency of generated questions, particularly for spontaneous spoken content.

2.2 Semantic continuity model

The semantic continuity (SC) model is a BERT-based neural architecture designed to measure semantic coherence between consecutive text fragments. Unlike traditional unidirectional sentence embeddings, the SC model produces two distinct representations: a left-embedding and a right-embedding. These embeddings capture the semantic similarity across sentence boundaries, enabling the model to assess whether one fragment naturally follows another.

In other words, the left-embedding represents the starting point of a fragment, while the right-embedding captures its potential continuation, allowing the model to determine whether one sentence naturally follows another. Inspired by Sentence-BERT, the model operates in a shared embedding space, where a trainable similarity metric determines the degree of semantic continuity.

The SC model first encodes input text using BERT, transforming a sequence of tokens into a series of hidden representations. Left- and right-embeddings are generated by applying two independent fully connected layers with GELU activation to the BERT output, forming two processing branches. Then – for each branch – a global average pooling layer followed by the normalization layer aggregates token-level representations into a single sentence-level vector. The branches result in two embeddings: the right-embedding capturing the semantic characteristics of the preceding context and the left-embedding representing the anticipated continuation.

To train the SC model, text segments are sampled from a training dataset and split at sentence boundaries, producing pairs of consecutive segments. The model learns to align the right-embedding of the first segment with the left-embedding of the following segment when they form a coherent sequence. A dot-product similarity function, combined with trainable scaling parameters, quantifies the semantic continuity between embeddings. The trained model is able to distinguish valid continuations from unrelated text segments.

The SC model is trained using a contrastive loss function that increases similarity for correct continuation segment pairs while reducing it for mismatched segment pairs, reinforcing its ability to distinguish between coherent and incoherent text segment pairs. To further refine the model predictions, a binary cross-entropy loss calibrates the probabilistic outputs, ensuring that the similarity scores represent the likelihood of one segment following another.

In the experiments, we trained the SC model on a combined dataset that included both structured and unstructured text sources: the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2018) and the USC Shoah Foundation Interview corpus (USC-SFI). This diverse training set allowed the model to learn semantic continuity across different text types, improving its ability to generalize to various linguistic patterns and discourse structures.

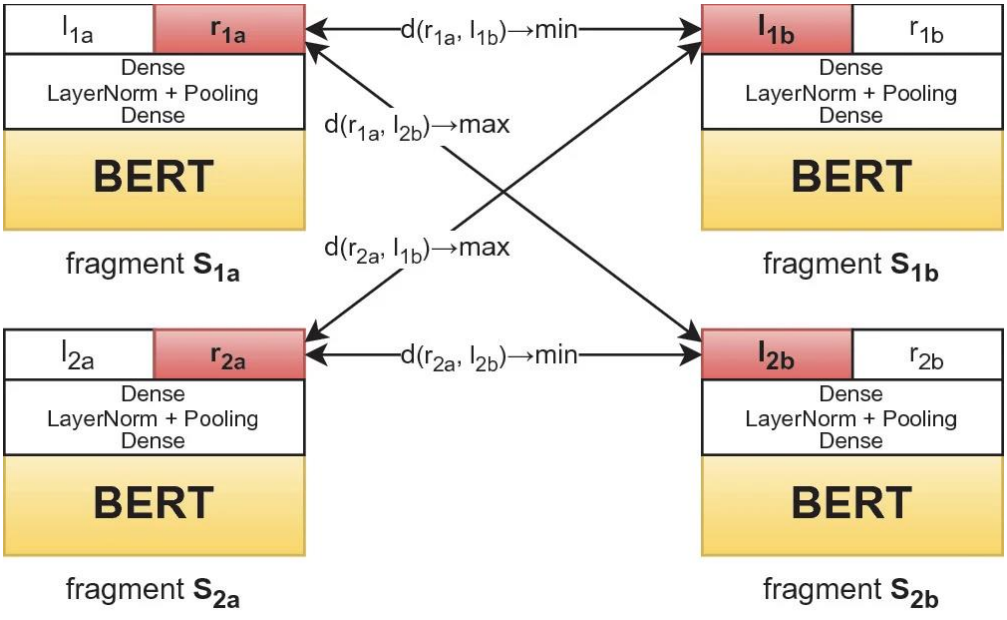

Figure 2. *Semantic continuity model based on BERT and contrastive loss function.*

The final training objective was a weighted sum of the contrastive loss and binary cross-entropy loss, with carefully tuned weights to balance similarity learning and probabilistic calibration. To ensure stable convergence, we applied a linear learning rate schedule, decaying from 1e-4 to zero over 500k steps. All the model parameters, including the BERT weights and trainable similarity coefficients, were updated during training, ensuring full adaptation to the task.

The trained SC model demonstrated robustness across multiple text domains, effectively capturing semantic continuity in both structured (SQuAD) and unstructured (USC-SFI) datasets. Training on a diverse dataset enhanced its generalization ability, enabling the model to distinguish meaningful text transitions across different contexts.

3 RESULTS

In our previous work (Švec et al., 2024), we compared the performance of several versions of the SC model on the SQuAD and USC-SFI datasets separately. Our findings demonstrated that the SC model trained on a combined dataset (union of SQuAD and USC-SFI data) outperformed the model trained solely on USC-SFI when evaluated on the USC-SFI test set. Moreover, it matched the performance of the specialized SQuAD-trained model on the SQuAD test set. By incorporating both structured, fact-based (SQuAD) and conversational (USC-SFI) data, the SC model learned to handle diverse linguistic styles, leading to improved generalization. Specifically, the model correctly identified the appropriate question continuation from a set of 10 randomly sampled continuations with an accuracy of 78.1% on the SQuAD dataset and 71.0% on the USC-SFI dataset.

We then compared two different AQ models: AQ1, fine-tuned on the SQuAD dataset (QA dataset) and AQ2, trained on the TALP synthetic QA dataset (ChatGPT-generated dataset). Since direct evaluation of AQ models on the USC-SFI dataset is inherently challenging due to the absence of ground-truth question-context pairs, we adopted an indirect evaluation strategy. Specifically, we assessed the performance of both AQ models using the well-established SQuAD dataset as a reference. To ensure that only semantically coherent questions were retained, we applied the SC model as a filtering mechanism, excluding question-context pairs with a probability score below 95%. This step retained only high-confidence, contextually relevant questions, essential for preserving the integrity of oral history testimonies. By eliminating low-probability question-context pairs, we ensure that the generated questions align well with the testimony content without introducing misleading or irrelevant prompts.

When applying the SC model-based filtering to the questions generated by AQ1 and AQ2, we observed that AQ2 consistently retained a higher percentage of question-context pairs compared to AQ1. This suggests that AQ2 generated questions with stronger semantic alignment to their contexts, making it more suitable for oral history testimonies, where maintaining conversational flow and coherence is critical. Using the SC model as the filter, AQ2 achieved a retention rate of 56.39%, surpassing AQ1 (retention rate of 54.75%) and closely approaching the benchmark set by the SQuAD reference dataset (retention rate of 59.31%). This highlights the advantage of training on synthetic, conversational data, which better reflects the spontaneous, unstructured nature of oral history interviews compared to structured, fact-based datasets such as SQuAD.

For the final AQ framework, we used SC model-based filtering applied to the AQ2 model. We implemented this framework on the USC-SFI dataset to generate contextually relevant questions for oral history testimonies. The model processed 15.6 hours of automatically generated interview transcripts (ASR word error rate: 12.88%) (Lehečka et al., 2023), producing an initial 8,027 candidate questions. After applying the 95% probability threshold in the SC model, only 477 high-confidence questions (5.94%) were retained. This strict filtering ensured that only semantically relevant questions were included, eliminating misleading or vague outputs. With an average of one question every two minutes of testimony, the system balances between enriching user navigation and maintaining the authenticity of the testimonies.

4 DISCUSSION

Our ongoing efforts make use of retrieval-augmented generation (RAG) techniques and vector databases to develop conversational, chatbot-like interfaces. These systems allow users to interact seamlessly with extensive datasets, combining semantic matching, advanced filtering and intent classification within a unified environment powered by specialized AI agents. The asking questions framework, whose principles are described above, generates contextually relevant and factually accurate questions for arbitrary text passages. By indexing these questions in

vector databases, we further enhance the semantic matching capabilities of RAG systems. To further enhance retrieval performance, we deploy large language models (LLMs) on our in-house servers, where we experiment with fine-tuning and utilizing them to assist in pre-generating training data. Early testing on 150 hours of transcribed speech data demonstrates the effectiveness of this approach. These advancements underpin innovative, voice-enabled, web-based interfaces, enabling conversational engagement with archives.

Let us also stress that in our approach to enhancing the accessibility of oral history archives using advanced technology, we rigorously adhere to the principle of not altering the original testimony. This commitment ensures that the integrity and authenticity of the historical records are maintained. Our system, employing ASR and transformer-based neural networks, does not summarize, modify or reinterpret the original content of the interviews. Instead, it generates questions relevant to the content of each interview segment, to which the interviewee's responses, in their original form and voice, are presented. This method guarantees that the authentic voice of the interviewee, an invaluable asset in oral history, remains undisturbed and true to its original form.

5 CONCLUSION

Our work represents a seamless integration of cutting-edge AI technologies – including generative models – with historical preservation. By combining fine-tuned LLMs, advanced retrieval techniques and user-friendly web interfaces, we bridge the gap between complex archives and accessibility. These innovations empower researchers, historians and the public to interact with historical data meaningfully and efficiently, transforming the way archives are explored and utilized. Since the developed techniques support dialogue-like interaction with a non-human entity (in this case, an archive), the principles and findings of our research can also be utilized in other domains, e.g., in human-robot interaction.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding: This work was funded by the Czech Science Foundation (GA ČR), project number GA22-27800S.

Conflict of Interests: The authors declare no conflict of interest.

Author Contributions: M.B.: Conceptualization, Methodology, Software. J.Š.: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Resources. P.I..: Writing – original draft, Writing – review & editing, Project administration. A.F.: Data curation. F.P.: Validation.

Statement on the Use of Artificial Intelligence Tools: The authors declare that they didn't use artificial intelligence tools for generating the actual text of the article. However, ChatGPT is used in the research itself – in concordance with the topic of this special issue (Theory and Practice of Generative Artificial Intelligence Usage). The details of employing the AI tools are given in the text of the article.

Data Availability: The data that support the findings of this study are available from the corresponding author.

REFERENCES

Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajic, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T. & Zhu, W. J. (2004). Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing*, 12(4), 420-435. https://doi.org/10.1109/TSA.2004.828702

Gospodinov, M., MacAvaney, S., & Macdonald, C. (2023). Doc2Query–: when less is more. In *European Conference on Information Retrieval*, (pp. 414–422). Springer. https://doi.org/10.1007/978-3-031-28238-6_31

He, B., & Ounis, I. (2009). Studying query expansion effectiveness. In *European conference on information retrieval*, (pp. 611–619). Springer. https://doi.org/10.1007/978-3-642-00958-7_57

Ircing, P., & Müller, L. (2006). Benefit of proper language processing for Czech speech retrieval in the CL-SR task at CLEF 2006. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, (pp. 759–765). Springer. https://doi.org/10.1007/978-3-540-74999-8_95

Ircing, P., Psutka, J., & Vavruška, J. (2008). What can and cannot be found in Czech spontaneous speech using document-oriented IR methods—UWB at CLEF 2007 CL-SR track. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, (pp. 712–718). Springer. https://doi.org/10.1007/978-3-540-85760-0_90

Khashabi, D., Min, S., Khot, T., Sabharwal, A., Tafjord, O., Clark, P., & Hajishirzi, H. (2020). UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, (pp. 1896–1907). ACL. https://doi.org/10.18653/v1/2020.findings-emnlp.171

Krubiński, M., Ghadery, E., Moens, M. F., & Pecina, P. (2021). Just ask! evaluating machine translation by asking and answering questions. In *Proceedings of the Sixth Conference on Machine Translation*, (pp. 495–506). ACL.

Lehečka, J., Švec, J., Psutka, J. V., & Ircing, P. (2023). Transformer-based speech recognition models for oral history archives in English, German, and Czech. In *Proceedings of the Interspeech 2023*, (pp. 201–205). ISCA. https://doi.org/10.21437/Interspeech.2023-872

Mao, H. H., Li, S., McAuley, J., & Cottrell, G. (2020). Speech Recognition and Multi-Speaker Diarization of Long Conversations. In *Proceedings of the Interspeech 2020,* (pp.691–695). ISCA. https://doi.org/10.21437/Interspeech.2020-3039

Psutka, J., Ircing, P., Psutka, J. V., Hajič, J., Byrne, W. J., & Mírovský, J. (2005). Automatic transcription of Czech, Russian, and Slovak spontaneous speech in the MALACH project. In *Proceedings of the Interspeech 2005,* (pp. 1349–1352). ISCA. https://doi.org/10.21437/Interspeech.2005-489

Psutka, J., Švec, J., Psutka, J. V., Vaněk, J., Pražák, A., Šmídl, L., & Ircing, P. (2011). System for fast lexical and phonetic spoken term detection in a Czech cultural heritage archive. *EURASIP Journal on Audio Speech and Music Processing*, 2011(1), Article 10. https://doi.org/10.1186/1687-4722-2011-10

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21, 1–67.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (pp. 784–789). ACM. https://doi.org/10.18653/v1/P18-2124

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 EMNLP-IJCNLP*, (pp. 3982–3992). ACL. https://doi.org/10.18653/v1/D19-1410

Švec, J., Lehečka, J., Šmídl, L. (2022a) Deep LSTM Spoken Term Detection using Wav2Vec 2.0 Recognizer. In *Proceedings of the Interspeech 2022*, (pp.1886–1890). ISCA. https://doi.org/10.21437/Interspeech.2022-10409

Švec, J., Neduchal, P., & Hrúz, M. (2022b). Multi-modal communication system for mobile robot. *IFAC-PapersOnLine*, 55(4), 133–138. https://doi.org/10.1016/j.ifacol.2022.06.022

Švec, J., Bulín, M., Frémund, A., & Polák, F. (2024). Asking questions framework for oral history archives. In *European Conference on Information Retrieval*, (pp. 167–180). Springer. https://doi.org/10.1007/978-3-031-56063-7_11

Vavruška, J., Švec, J., & Ircing, P. (2013). Phonetic spoken term detection in large audio archive using the WFST framework. In *International Conference on Text, Speech and Dialogue,* (pp. 402–409). Springer. https://doi.org/10.1007/978-3-642-40585-3_51

Wang, J., Jatowt, A., & Yoshikawa, M. (2022). Archivalqa: A large-scale benchmark dataset for open-domain question answering over historical news collections. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 3025–3035). ACM. https://doi.org/10.1145/3477495.3531734

Yao, X., Tosch, E., Chen, G., Nouri, E., Artstein, R., Leuski, A., Sagae, K. & Traum, D. (2012). Creating conversational characters using question generation tools. *Dialogue & Discourse*, 3(2), 125–146. https://doi.org/10.5087/dad.2012.206

Acta Informatica Pragensia is published by the Prague University of Economics and Business, Czech Republic | eISSN: 1805-4951