INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Asking Questions: an Innovative Way to Interact with Oral History Archives

Jan Švec, Martin Bulín, Adam Frémund, Filip Polák

Department of Cybernetics, University of West Bohemia, Pilsen, Czech Republic

[honzas,bulinm,afremund,polakf]@kky.zcu.cz

Abstract

The paper describes our initial effort to use Transformer-based neural networks for understanding and presenting oral history archives. Such archives of interviews often contain large passages of the interviewee's speech. Our approach automatically generates relevant questions, which enrich such monotonous parts and allows the listener to better orient in the interview. The generated questions also allow for finding interesting parts of the interview without changing the original meaning of the testimony. We present our working pipeline consisting of a Wav2Vec speech recognizer, BERT-based punctuation detection, T5 asking questions model and BERT-based semantic continuity model.

Index Terms: Question answering, semantic search, oral history archives

1. Introduction

Oral history archives are a large source of historical knowledge. Different institutions are collecting interviews and testimonies related to major historical topics. Many of the well-known archives are related to Holocaust, for example, the USC Shoah Foundation Visual History Archive or the collection of the US Holocaust Memorial Museum.

The archives are multi-lingual large-scale collections of audio or audiovisual interviews following a similar scenario. For example, the Holocaust witnesses giving testimonies for USC Shoah Foundation completed a 50-page-long questionnaire asking for names, dates, and experiences from before, after, and during the Holocaust and World War II. Interviewers are expected to familiarize themselves with the survivor's history so that relevant questions can be asked. The Foundation Interviewer Guidelines include "helpful hints" for interview questions and suggest attention to chronology over stream-of-consciousness narration. The instructions also propose ways the interviewer might deal with certain subjects [1].

Listening to huge quantities of audio materials is impractical for a common user. On the other side, the testimonies provided by the interviewees were intended to give evidence of historical events, and it is not ethical to change the meaning or cherry-pick single facts from a given interview. Many efforts to provide access to such archives were proposed in recent research works, including speech-to-text technologies [2] and spoken-term detection methods [3]. Other approaches used traditional information retrieval methods [4]. However, the most used method for accessing the archives is a keyword search in automatically / manually created transcripts.

This research was supported by the Czech Science Foundation (GA CR), project No. GA22-27800S, and by the grant of the University of West Bohemia, project No. SGS-2022-017.

This paper proposes an innovative approach that integrates the above-mentioned technologies. The approach generates a new question-answer structure on top of the existing interview transcripts. The questions are automatically generated with related answers for a specific interview fragment. The questions can be indexed and time-aligned with the audio so the user can quickly get to interesting parts of the interview. The questions complement the interviewer and are helpful in passages where only the interviewee speaks. In other words, the questions can be understood as "open-set topics" related to the testimony. It is important to stress that the questions do not change the meaning of the testimony because the related parts of the original interview are presented as an answer. The rich set of such questions also encourages the user to search for the answers. In contrast to the question-answering methods, we use the term *asking questions* for our approach (abbreviated as AQ in contrast to question answering – QA).

2. Asking questions framework

The main motivation of our approach is similar to the recently published Doc2Query [5]. Doc2Query was proposed for information retrieval from textual sources. Since the spoken interviews contain mainly spontaneous speech without any inherent grammatical structure, we had to design methods for this use case. This work focused on English USC Shoah Foundation data, but the approach is not limited to a specific language.

As a speech recognizer, we trained the recent *Wav2Vec 2.0 end-to-end model* with a lower-case n-gram language model estimated from CommonCrawl data. For English, we achieved 12.9% word error rate (for comparison: OpenAI whisper-large model achieved 17.3% WER). In addition, the raw output of the speech recognizer was post-processed using automatic punctuation detection and casing reconstruction [6].

We can define a sliding window context using sentence-like units based on automatic punctuation. We use a *T5-based asking questions (AQ) model* for each context to generate one possible question related to this context. We also generate the answers to the questions because it helps the model generate more specific questions. The T5 AQ model was first fine-tuned using the Stanford Question Answering Dataset (SQUAD) [7] to generate the question and answer from a given textual context. This model was able to ask factual questions, but the interviews often contained utterances related to feelings, emotions, or relations mentioned in the context. We used the ChatGPT prompt to generate a SQUAD-like dataset of natural questions and answers based on spoken interviews. We wanted to secure the privacy of USC Shoah Foundation data, and therefore we used the proxy dataset This American Life [8] containing transcripts of podcast interviews. The T5 AQ model was then fine-tuned for the sec-

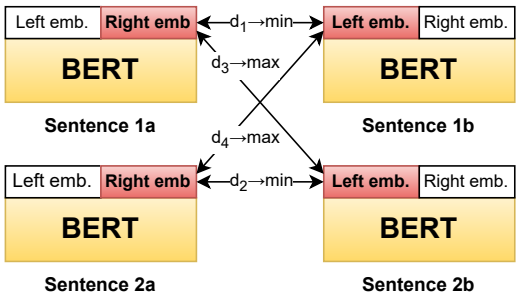

Figure 1: *Semantic continuity model based on BERT and contrastive loss function.*

ond time using these machine-generated "spontaneous speech transcripts."

Because the T5 model always generates the question-answer pair for a given context – even if the context does not contain any meaningful information (e.g., it is only a discourse marker) – we used a second model trained to classify the semantical continuity of the question-context pair. This way, only the questions which naturally precede the context can be presented to the user. The *semantic continuity model* is a BERT architecture producing two sentence-level embeddings: left- and right-embedding (Fig. 1). The idea of training objective is similar to Sentence-BERT [9]. We randomly sample two fragments of texts from the training textual data - *Sentence1* and *Sentence2*. Each fragment is then split on the sentence boundary into four fragments (*1a*, *1b*, *2a*, *2b*). Therefore, the fragment *b* is a semantic continuation of fragment *a*. The network is trained to minimize distances between corresponding continuation pairs ($d_1, d_2 \rightarrow \min$, Fig. 1) and to maximize distances of two mismatching pairs ($d_3, d_4 \rightarrow \max$). The distances are based on the cosine similarity of the right-embedding vector and the following left-embedding vector. Finally, we use a sigmoid with trainable parameters to calibrate the distances into a [0; 1] score interval.

To summarize the processing steps: we first recognize the audio. Then the punctuation is inserted into the textual transcript, and word casing is restored. For each sliding context window, the questions are generated using the T5 AQ model, and the relevance score to the context is predicted using the semantic continuity model.

3. Example output

We used the described pipeline to automatically process testimonies from the YouTube channel of USC Shoah Foundation[1]. We designed a simplistic web user interface that displays the generated questions and the extracted answers. After clicking the question, the video playback starts from the given timestamp. A selected sample of asked questions, corresponding recognized contexts, and automatically extracted answers are shown in Fig. 2.

4. Conclusion & Future work

The demonstration shows our initial effort in generating contextual questions for the oral history interviews that help the user better understand the testimonies. Our future work should define the full experimental setup of this task, especially evaluation protocols, metrics, and datasets. Although the asking questions framework works completely offline, the evaluation procedure of spoken dialog systems could inspire our future work.

[1] `https://www.youtube.com/USCShoahFoundation`

Q: **Was the wedding large?** (score: 1.0, timestamp: 0:49:46)
Ctx: Yes, we had a ceremony. We had a small wedding. Just the family. Were you able to have any kind of celebration after the war? No, we had no celebration. Where where could we celebrate?

Q: **What is Pesach?** (score: 0.98, timestamp: 0:15:23)
Ctx: How about Pesach? Do you remember anything about Pesach? Oh, Pesach was a special holiday for Jewish people, not only myself, but for all the Jewish people. What was the for you?

Q: **What was the speaker's experience with the organization?** (score: 0.0, timestamp: 1:42:47)
Ctx: That's what they hit over the head, what happened to you afterwards? What happened to me? After that? We had an illegal organization between our own people who was involved in this illegal organization.

Figure 2: *Selected samples of asked questions regarding the testimony of Abraham Bomba publicly available from* `https://www.youtube.com/watch?v=1eWo8j6uEow`. *The questions (Q:) are automatically generated for a given context (Ctx:), score is assigned by the semantic continuity model, and timestamp refers to the above-mentioned interview. The first two questions show highly scored semantic continuity; the remaining has a poor continuity score.*

5. References

[1] "USC Shoah Foundation Oral History with Abraham Bomba — Experiencing History: Holocaust Sources in Context," accessed: 2023-04-12. [Online]. Available: https://perspectives.ushmm.org/item/usc-shoah-foundation-oral-history-with-abraham-bomba

[2] M. Picheny, Z. Tüske, B. Kingsbury, K. Audhkhasi, X. Cui, and G. Saon, "Challenging the Boundaries of Speech Recognition: The MALACH Corpus," in *Proc. Interspeech 2019*, 2019, pp. 326–330.

[3] J. Švec, L. Šmídl, J. V. Psutka, and A. Pražák, "Spoken Term Detection and Relevance Score Estimation Using Dot-Product of Pronunciation Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 4398–4402.

[4] P. Pecina, P. Hoffmannová, G. J. F. Jones, Y. Zhang, and D. W. Oard, "Overview of the clef-2007 cross-language speech retrieval track," in *Advances in Multilingual and Multimodal Information Retrieval*. Berlin, Heidelberg: Springer, 2008, pp. 674–686.

[5] M. Gospodinov, S. MacAvaney, and C. Macdonald, "Doc2query–: When less is more," in *Advances in Information Retrieval*. Cham: Springer Nature Switzerland, 2023, pp. 414–422.

[6] J. Švec, J. Lehečka, L. Šmídl, and P. Ircing, "Transformer-based automatic punctuation prediction and word casing reconstruction of the asr output," in *Text, Speech, and Dialogue*. Cham: Springer International Publishing, 2021, pp. 86–94.

[7] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," in *Proceedings of ACL 2018*. Melbourne, Australia: ACL, Jul. 2018, pp. 784–789.

[8] H. H. Mao, S. Li, J. McAuley, and G. W. Cottrell, "Speech Recognition and Multi-Speaker Diarization of Long Conversations," in *Proc. Interspeech 2020*, 2020, pp. 691–695.

[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 EMNLP-IJCNLP*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992.