FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

STUDENTSKÁ
VĚDECKÁ KONFERENCE 2023

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

Adam Frémund[1], Martin Bulín, Jan Švec, Filip Polák[2]

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, USC Shoah Foundation Visual History Archive or the collection of US Holocaust Memorial Museum.

The archives are basically 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.

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 and spoken-term detection methods. Other approaches used traditional information retrieval methods. However, the most used method for accessing the archives is by using a keyword search in automatically / manually created transcripts.

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 together 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 useful in passages where only the interviewee speaks. 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. 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

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. In addition, the raw output of the speech recognizer was post-processed using automatic punctuation detection and

[1] student of the doctoral study program Applied Sciences and Informatics, field of study Cybernetics, specialization Spoken dialog systems, e-mail: afremund@kky.zcu.cz

[2] University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics, email:bulinm@kky.zcu.cz, honzas@kky.zcu.cz, polakf@kky.zcu.cz

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?

Figure 1: Selected sample of asked questions regarding the testimony of Abraham Bomba publicly available from `https://www.youtube.com/watch?v=1eWo8j6uEow`. The question (Q:) is automatically generated for a given context (Ctx:), score is assigned by the semantic continuity model, and timestamp refers to the above-mentioned interview.

casing reconstruction (Švec (2021)).

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 Standford Question Answering Dataset (SQUAD) (Rajpurkar (2018)) 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. To generate the SQUAD-like dataset based on spoken interviews, we used the ChatGPT prompt to generate natural questions and answers. We wanted to secure the privacy of USC Shoah Foundation data, and therefore we used the proxy dataset This American Life (Mao (2019)) containing transcripts of podcast interviews. The T5 AQ model was then fine-tuned for the second 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.

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.

Acknowledgement

This work was supported by the University of West Bohemia grant, project No. SGS-2022-017

References

Huanru Henry Mao and Shuyang Li and Julian McAuley and Garrison W. Cottrell (2020) Speech Recognition and Multi-Speaker Diarization of Long Conversations, Proc. Interspeech 2020, pp.691–695

Rajpurkar, Pranav and Jia, Robin and Liang, Percy (2018) Know What You Don’t Know: Unanswerable Questions for SQuAD. Melbourne, Australia, Proceedings of ACL 2018, pp. 784–789.

Švec, Jan and Lehečka, Jan and Šmídl, Luboš and Ircing, Pavel (2021) Transformer-Based Automatic Punctuation Prediction and Word Casing Reconstruction of the ASR Output. Cham, Springer International Publishing, Text, Speech, and Dialogue, pp.86–94.