INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

Multimodal Dialog with the MALACH Audiovisual Archive

Adam Chýlek, Luboš Šmídl, Jan Švec

New Technologies for the Information Society (NTIS)
Faculty of Applied Sciences, University of West Bohemia, Czech Republic
chylek@ntis.zcu.cz, smidl@ntis.zcu.cz, honzas@ntis.zcu.cz

Abstract

In this paper, we present a multimodal dialog system capable of information retrieval from the large audiovisual archive MALACH of Holocaust testimonies. The users can use spoken natural language queries to search the archive. A graphical user interface allows the users to quickly view footage with the answers and explore their context. The dialog was deployed in two languages - English and Czech. The system uses automatic speech recognition and natural language processing for knowledge base construction and for processing of the user's input.
Index Terms: dialog systems, human-computer interaction, information retrieval

1. Introduction

The MALACH archive of Holocaust testimonies[1] is made out of thousands of hours of video footage. It contains interviews and personal testimonies of Holocaust survivors and witnesses. The archive is maintained by the USC Shoah Foundation[2]. Its collection was initiated by the Shoah Visual History Foundation founded by Steven Spielberg.

A large portion of the archive is not manually transcribed. The footage is annotated only with hand-picked keywords. This limits the capability to search the archive.

Our research group has focused in the past years on automatic transcription of the audio using automatic speech recognition (ASR) [1, 2]. With the transcriptions available, we were able to develop and deploy a full-text search system [3] that can be used by experts and by the general public in the Malach Centre for Visual History and Jewish Museum in Prague.

We now build upon this modular system to make the archives even more accessible in the form of a human-machine dialog, allowing the user to ask natural language queries and retrieve information from the archive that would be more difficult to obtain with the textual input[3].

The system was created for the English and Czech portions of the archive and it can be easily extended to other languages.

2. System Outline

The system can be divided into two parts - constructing a **knowledge base (KB)** and building a **dialog system**. The **KB** is constructed from the output of an ASR system in an offline process. Subsequently, we use a semantic entity detection (SED, [4]) approach to extract the knowledge from the footage. The semantic entities are defined by their word forms in expert-made context-free grammars. We have created grammars for semantic entities that we consider to be useful the most to the dialog system in our domain: cities, countries, dates, names, proper nouns, family members, life events (e.g. birth, death, injury), education and other geographical places (e.g. camp names). Because of that, extending the algorithm to other languages is simply a matter of training the ASR on the language-specific data and then localizing the grammars.

The **dialog system** uses agent-based dialog manager and it is easily extensible.We are using an in-house developed SpeechCloud framework that provides access to a realtime ASR, SED and text-to-speech system (TTS). It also serves as an event dispatcher between the GUI and the dialog manager. In the dialog system, we use the same SED approach for natural language understanding as in the KB construction. We use the same entities and their grammars and we only extend them by a small number of dialog-related entities (e.g. keywords to repeat the system's utterance or play the footage).

The users make **queries in natural language**. The system processes the output of the ASR and SED into dialog state updates. The dialog agents then react to these changes. If it is possible to make a KB search request based on the information contained in the dialog state, the KB agent selects an appropriate template for the KB queries, fills the necessary information and executes them. The information the agent retrieves from the KB contains the word forms of the answers, the footage that they are mentioned in and the timestamps when they are mentioned. If the user's query did not contain any known entities or keywords, we fall back to a full-text search using the ASR's 1-best hypothesis (with predefined stopwords removed). This fallback search will be presented to the user seamlessly as if we were searching our knowledge base instead of the full-text index.

3. Application Overview

The application's interface is a dynamic web page. Upon navigating to the welcome screen, the user is greeted with a simple search bar. After the recognition engine is initialized, an experienced user can start making queries right away. For inexperienced user, the system has an agent that produces increasingly more detailed instructions on how to use the system.

The recognition is started manually by clicking or tapping a button. Using the same button the recognition can be stopped. The recognition can stay active for the whole dialog as we are using voice activity detection to recognize the end of the user's query. We are also automatically pausing the recognition when the system is speaking (via TTS) in order not to recognize the system's utterance as the user's input.

If the system recognizes a **search query** in the user's utterance, the dialog agents parse the natural language input into a query to our KB and execute the search. The system then uses

[1] https://malach.umiacs.umd.edu/

[2] https://sfi.usc.edu/

[3] This paper describes mainly the interface and the principles of the system. We invite you to look at https://youtu.be/baluRdW4FzI for an example of the dialog.

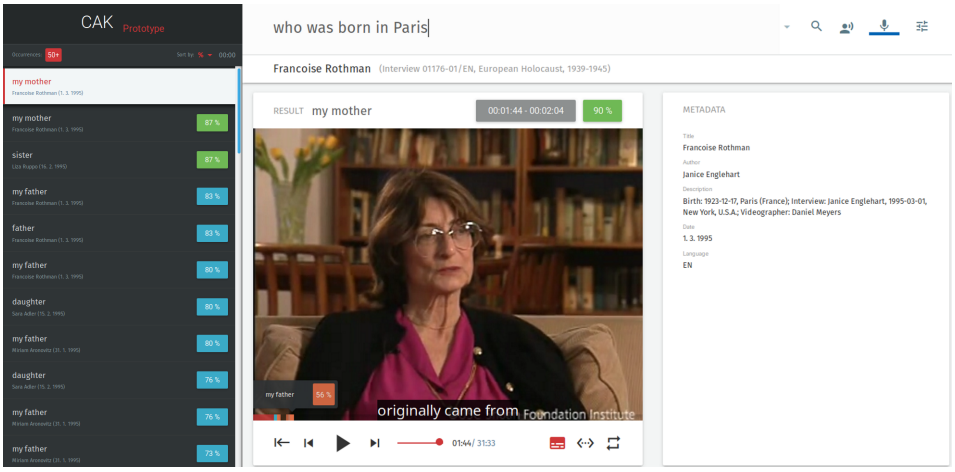

Figure 1: *The graphical user interface with answers to a natural language query "Who was born in Paris". The top part shows a search bar with buttons that control the recognition and the search. On the left, there is a list of the answers and the corresponding speakers that uttered them. In the middle, there is a video player with footage for an answer with the highest score. The video is automatically rewound to a moment before the answer "my mother" was uttered. A progress bar for the player shows that other answers were found in the footage as well (e.g. "my father"). On the right, there are the metadata for the footage.*

the TTS to present the answers to the user in natural language and also to display a list of all the records it found with their corresponding video footage on the web page as in Figure 1. The user can then choose to play the footage and **explore the context** of the answer. When the search takes a long time to finish, we make short announcements to show that the search is still ongoing. We also allow the user to cancel the search.

The dialog flow is designed for an **exploration** of the knowledge base. For example, the user wants to know whether the witness mentions Auschwitz and the witness never mentions it. We will not only inform the user that the witness never mentioned Auschwitz but we will also look if anyone else mentioned it and suggest the user to view their footage.

The dialog also allows the user to explore the whole footage by creating a summary of the entities the witness mentioned in the footage. Furthermore, the dialog agents keep a reference to the current footage's speaker to allow **follow-up questions**, like "When was her sister born?"

The web page with results contains a video player that can play the footage related to the answer. Its progress bar highlights other positions in the footage where we found the answer mentioned. This allows the user to quickly go through all the mentions and their **context**. The player also displays subtitles that are created from the ASR's automated transcription of the audio. Basic controls for the video playback are also available.

We have built the dialog and the KB in two languages - **English** and **Czech**. The GUI allows the user to simply switch the language of the dialog (and the corresponding KB) on the web page. This also simplifies the deployment of such system - we only specify which languages should be available to each deployed instance.

The user can even make the natural language queries using a text input box. Thanks to the modularity of the SpeechCloud platform we simply bypass the ASR module and provide the text from the input box directly as the input to the SED module.

4. Conclusion

We have extended a search interface for large audiovisual archives to allow natural language queries and a dialog with the system. The users are able to retrieve information from the archive that would be difficult to obtain otherwise. We have maintained the original modular structure of the system by using agent-based dialog manager and by using the modular SpeechCloud platform. The system can be used on other large archives (e.g. of TV broadcast) without modifications. Further research will focus on enhancing the KB by deriving knowledge from external sources (web pages and other knowledge bases about the Holocaust). We will also focus on improving the methods that derive the knowledge from the transcripts.

5. Acknowledgments

This research was supported by the Technology Agency of the Czech Republic, project No. TN01000024 and by the European Regional Development Fund under the project Robotics for Industry 4.0 (reg. no. CZ.02.1.01/0.0/0.0/15_003/0000470). The computing infrastructure was provided by the grant of Ministry of Education, Youth and Sports of the Czech Republic project No. LO1506.

6. References

[1] J. Psutka, J. Švec, J. V. Psutka, J. Vaněk, A. Pražák, L. Šmídl, and P. Ircing, "System for fast lexical and phonetic spoken term detection in a Czech cultural heritage archive," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2011, no. 1, p. 10, 2011.

[2] J. Švec, J. V. Psutka, J. Trmal, L. Šmídl, P. Ircing, and J. Sedmidubsky, "On the Use of Grapheme Models for Searching in Large Spoken Archives," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, 2018, pp. 6259–6263.

[3] P. Stanislav, J. Švec, and P. Ircing, "An engine for online video search in large archives of the holocaust testimonies," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept, 2016, pp. 2352–2353.

[4] J. Švec, P. Ircing, and L. Šmídl, "Semantic entity detection from multiple ASR hypotheses within the WFST framework," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings*, 2013, pp. 84–89.