A Bilingual Interactive Human Avatar Dialogue System

Dana Abu Ali, Muaz Ahmad, Hayat Al Hassan, Paula Dozsa,
Ming Hu, Jose Varias, Nizar Habash
New York University Abu Dhabi
{daa389,muaz.ahmad,hayat.alhassan,pauladozsa,
ming.hu,jose.varias,nizar.habash}@nyu.edu

Abstract

This demonstration paper presents a bilingual (Arabic-English) interactive human avatar dialogue system. The system is named TOIA (time-offset interaction application), as it simulates face-to-face conversations between humans using digital human avatars recorded in the past. TOIA is a conversational agent, similar to a chat bot, except that it is based on an actual human being and can be used to preserve and tell stories. The system is designed to allow anybody, simply using a laptop, to create an avatar of themselves, thus facilitating cross-cultural and cross-generational sharing of narratives to wider audiences. The system currently supports monolingual and cross-lingual dialogues in Arabic and English, but can be extended to other languages.

1 Introduction

Conversational agents are software programs that are able to conduct conversations with human users (interactors), by interpreting and responding to statements made in ordinary natural language. The components of our system, TOIA, target two types of users: (a) avatar makers, which are the people who wish to create personalized avatars, and (b) interactors, those interacting with the avatar. The system is designed to allow anybody, simply using a laptop, to create an avatar of themselves, thus facilitating cross-cultural and cross-generational sharing of narratives to wider audiences. Through our system, we aim to enable interactor users to conduct conversations with a person (avatar maker) who is not available for conversation in real time with the intention of learning about them. Since face-to-face human interaction is a powerful tool of human understanding, TOIA overcomes the restrictions on time and place that limit this type of interaction, presenting users with a platform for dialogue at their own pace and convenience. Additionally, it allows people from different linguistic backgrounds to communicate by supporting mechanisms for cross-lingual interactions between users and avatars that speak different languages.

2 Related Work

TOIA is inspired by research at the University of Southern California's Institute for Creative Technologies (ICT), such as *SGT Blackwell*, a digitally animated character designed to serve as an army conference information kiosk (Leuski et al., 2006). Users can talk to the character through a microphone, after which their speech is converted into text through an automatic speech recognition (ASR) system. This output is then analyzed by an answer selection module and the appropriate response is selected from the 83 pre-recorded lines that Blackwell can deliver. Another ICT project, based off of video recordings instead of digital media, is *New Dimensions in Testimony* (NDT), a prototype dialogue system allowing users to conduct conversations with Holocaust survivor Pinchas Gutter (Traum et al., 2015a,b; Artstein et al., 2015, 2016). Similarly, users talk to the Gutter Avatar through a microphone; their speech is then converted into text through ASR; a dialogue manager identifies the proper video to play back to simulate a conversational turn. The NDT setup is quite impressive in terms of the amount of resources that went into creating the avatar recording — hours of recording, use of top-of-the-line digital cinema cameras, etc. In TOIA, our goal is to create a system that will enable any avatar maker with a laptop and webcam to create and

Proceedings of the SIGDIAL 2018 Conference, pages 241–244,
Melbourne, Australia, 12-14 July 2018. ©2018 Association for Computational Linguistics

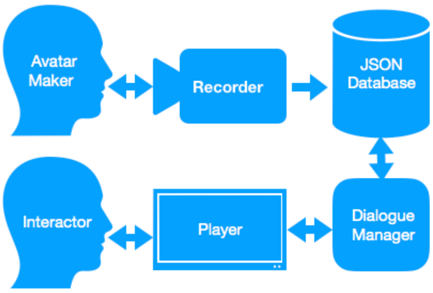


Figure 1: TOIA System Design

publicly share their avatar. We are also interested in enabling the dialogue to happen cross-lingually, where an interactor asks in one language, and the avatar answers in another language (with captions in the interactor's language). In our current system we support Arabic (Ar) and English (En), allowing for all combinations: (Ar-Ar, En-En, Ar-En and En-Ar).

3 Overall System Design

TOIA includes four components: (a) a recorder that records the avatar maker's videos, (b) a database that contains the avatar's video responses and other data facilitating the matching of the interactor's questions with the avatar's answers, (c) a player interface through which the interactor is able to interact with the avatar, and (d) a dialogue manager that matches the user's speech to the avatar's answers in any of the four language pairs. Figure 1 illustrates the relationship among the different components.

3.1 Recorder

We implemented a web-based recorder to help avatar makers record personal videos. We used *Node.js* as the runtime environment, *Express* as the web framework, *WebRTC* for live recording support, and *MangoDB* for real-time data updates.

The overall framework is as follows: Avatar makers create scripts consisting of pairs of questions and answers from scratch or customize existing scripts by removing or adding questions as desired. These scripts are uploaded to a *Cloudant* database. Throughout the recording, avatar makers can further update answers, delete questions or add questions on the spot.

The avatar maker selects a specific question prompt in any order and proceeds to record a video response to pair with it. A semi-transparent head location indicator is provided to help the avatar maker create consist videos. The avatar maker can review the recorded video, re-record it or delete it. Figure 2 shows a screen capture of the recorder.

Currently, we have recorded four digital avatars using this interface, three in English and one in Arabic, each consisting of around 300 question-answer pairs.

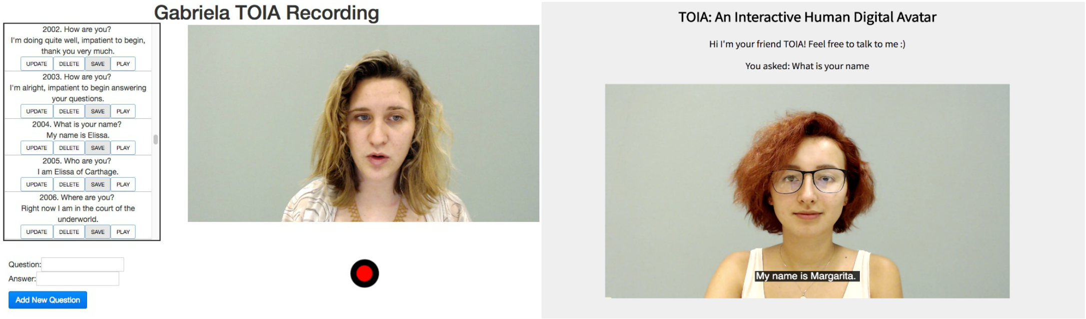
3.2 Database

TOIA's avatar data are stored using two components: (a) a collection of the answer videos saved locally for speedy access; and (b) a JSON database storing question-answer entries. Each question-answer entry in the database has a unique reference id number, the answer video path in file system, as well as a character tag, consisting of the name of the character. Avatar files supporting cross-lingual interaction also include translations of the answer text. We currently use manual translations, but machine translations of the questions and answers may also be included to support cross-lingual dialogue management, allowing for languages beyond Arabic and English. Figure 3 shows three English, and their corresponding Arabic, database entries.

3.3 Player

The main TOIA system interface is the player, through which the interactor user is able to interact with the avatar. Similarly to the recorder, we implemented a web-based user interface using a *Flask* (Grinberg, 2018) web platform utilizing both *Python* and *JavaScript*.

The interactor can initiate a conversation with any of the available avatars by selecting one of them using the first page of the player's interface, and then specifying whether the interaction will be in Arabic or English. The character can easily be switched at any point by returning to the player's main page, after which the interaction session would restart with a different avatar. The player listens to the interactor through a microphone and then passes the collected audio through an ASR system (Google Speech API). The text produced through ASR is then passed on to the dialogue management system. The dialogue management system returns a video file path to be displayed by the player. While the video is playing, the microphone switches to mute mode to avoid feedback. It then starts listening for utterances again once the video ends.



Gabriela TOIA Recording

2002. How are you?
I'm doing quite well, impatient to begin, thank you very much.

2003. How are you?
I'm alright, impatient to begin answering your questions.

2004. What is your name?
My name is Elissa.

2005. Who are you?
I am Elissa of Carthage.

2006. Where are you?
Right now I am in the court of the underworld.

Question:
Answer:

TOIA: An Interactive Human Digital Avatar

Hi I'm your friend TOIA! Feel free to talk to me :)

You asked: What is your name

My name is Margarita.

Figure 2: TOIA Recorder (left) and Player (right)

Entry ID	Language	Avatar	Question	Answer	Video Path
1001	English	Katarina	Who are you?	I am an avatar of an NYUAD student who can answer any questions you may have about the university.	katarina-1.mp4
1002	English	Katarina	What can I talk to you about?	You can to talk me about anything relating to Academics, Admissions, life in Abu Dhabi, and NYUAD in general. Feel free to ask me about my experiences here.	katarina-2.mp4
1032	English	Katarina	Where is the NYU Abu Dhabi campus?	The NYU Abu Dhabi campus is located on Saadiyat Island. Saadiyat is home to a few emerging cultural landmarks such as the Louvre Museum in Abu Dhabi.	katarina-32.mp4
2001	Arabic	Katarina	من أنت؟	أنا أفاتار لطالبة في جامعة نيويورك أبوظبي و يمكنني أن أجيب على أي أسئلة لديك عن الجامعة.	katarina-1.mp4
2002	Arabic	Katarina	عن ماذا يمكنني أن أحدثك؟	يمكنك التحدث معي عن أي شيء له علاقة بالشؤون الأكادمية أو القبول أو الحياة في أبوظبي أو عن جامعة نيويورك أبوظبي بشكل عام. يمكنك سؤالي عن تجربتي الشخصية هنا أيضًا	katarina-2.mp4
2032	Arabic	Katarina	أين يقع حرم جامعة نيويوك أبوظبي؟	يقع حرم جامعة نيويورك أبوظبي في جزيرة السعديات. تعد السعديات موطنا لبعض المعالم الفنية والثقافية الناشئة بما فيها متحف اللوفر أبوظبي.	katarina-32.mp4

Figure 3: TOIA Database example

We designed the interface so that the user's interaction and control could be navigated solely through audio. The speech recognition is done in streaming mode; and the end of a question is determined through silence. This allows for the interaction to feel more natural and as close to a *live* conversation as possible. We have a collection of 'filler' videos with every avatar, that play in a loop while the player is waiting for the question to be completed; the microphone is active only while these videos are playing.

Regardless of the avatar's spoken language, the player accepts utterances in both Arabic and English. The player also displays captions with English subtitles for Arabic speaking avatars, and Arabic subtitles for English speaking avatars. The subtitles are generated based off of the answers in the script, which by design match the video recordings. To support a smoother user experience we display the text processed by the speech recognizer. This allows the user to recognize when the utterance has not been 'heard' correctly, and encourages them to use a clearer or louder voice when interacting with the avatar. The right side of Figure 2 shows the TOIA player interface.

3.4 Dialogue Manager

Once the interactor selects the avatar and interaction language, the dialogue manager loads the data linked to the chosen avatar. A new dialogue session is created in order to save the state of the conversation and ensure a natural flow where repetition and irrelevant answers are minimized.

The input to the dialogue manager is a textual version of the interactor's last utterance with a language id. The output is a path to the video file that is to be played in response.

The dialogue manager matches the interactor's questions with all the questions and answers in the avatar database. In order to facilitate the matching, the text in both the interactor questions and the database entries is preprocessed, removing punctuation and stop words, then expanded into word feature representations for matching purposes. The word representations include: unigram, bigram and trigram sequences, in terms of raw words, their lemmas and their stems. Lemmas abstract over inflectional variants of the word, which is particularly important for Arabic, a morphologically rich language (Habash, 2010).

For English, we use the Natural Language Tool

Kit (NLTK) (Bird et al., 2009). For Arabic, we use the Standard Arabic Morphological Analyzer (SAMA) (Graff et al., 2009). Due to Arabic's orthographic ambiguity, SAMA produces multiple analyses (lemmas and stems) per word. We select a single analysis using the Yamama system's maximum liklihood models for Arabic lemmas (Khalifa et al., 2016).

We further add lemma synonyms to increase the possibility of matching. For English, we used NLTK Synset support (Bird et al., 2009). For Arabic, we created synthetic synsets by clustering Arabic lemmas with the same English glosses. Only for English, the database of questions and answers is also enriched through automatically generated questions, based off of the answers, to increase the probability of finding an appropriate answer for an interactor's query.

The matching process is optimized for speed using a number of hash maps to allow the fast generation of an answer ranked list. The ranking uses the number of matches in the various dimensions mentioned above, term-frequency inverse-document-frequency weights, as well as a history of whether a particular video had been played already in the current session. The more matches between the interactor's utterance and the question-answer pair, the more the likelihood that pair's entry will be selected. In the case of multiple tied entries, the one whose answer video was played the least in the current session is chosen. The playing count of the chosen entry is updated.

4 Preliminary Evaluation

We performed a user study with ten Arabic speakers and ten English speakers, each of whom chatted with three avatars (two English, one Arabic). The metrics we recorded were accuracy, understanding, interaction pace, timely response and conversation flow. On average across all metrics and users, we received a score of 3.6 out of 5. 85% of users enjoyed interacting with avatars, and 60% said they would like to interact with other avatars.

5 Demo Plan

In the demo of our work, we will present the four avatars we have created, each of which can be spoken to in English and Arabic. We will also present users with the ability to test the recorder by creating their own list of questions and answers, and recording a set of videos.

6 Conclusion and Future Work

We presented a bilingual (Arabic-English) interactive human avatar dialogue system that simulates face-to-face conversations between humans using previously recorded digital human avatars.

In the future, we plan to work on a detailed user study to evaluate the performance of the various components in our system. Consistent with our motivating mission, we also plan to make the recorder and player available online to allow users anywhere to use it. We look forward to maximizing its usability so that any person can start sharing their life stories at their own pace, from their point of view, and in the comfort of their home.

References

R. Artstein, A. Gainer, K. Georgila, A. Leuski, A. Shapiro, and D. Traum. 2016. New dimensions in testimony demonstration. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

R. Artstein, A. Leuski, H. Maio, T. Mor-Barak, C. Gordon, and D. Traum. 2015. How many utterances are needed to support time-offset interaction? In *Proceedings of the International Florida Artificial Intelligence Research Society Conference*.

S. Bird, E. Klein, and E. Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

D. Graff, M. Maamouri, B. Bouziri, S. Krouna, S. Kulick, and T. Buckwalter. 2009. Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium.

M. Grinberg. 2018. *Flask web development: developing web applications with python*. O'Reilly Media, Inc.

N. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

S. Khalifa, N. Zalmout, and N. Habash. 2016. Yamama: Yet another multi-dialect Arabic morphological analyzer. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

A. Leuski, R. Patel, and D. Traum. 2006. Building Effective Question Answering Characters. In *Proceedings of the SIGDIAL Workshop on Discourse and Dialogue*.

D. Traum, K. Georgila, R. Artstein, and A. Leuski. 2015a. Evaluating spoken dialogue processing for time-offset interaction. In *Proceedings of SIGDIAL*.

D. Traum, A. Jones, K. Hays, H. Maio, O. Alexander, R. Artstein, P. Debevec, A. Gainer, K. Georgila, K. Haase, et al. 2015b. New Dimensions in Testimony: Digitally preserving a Holocaust survivor's interactive storytelling. In *International Conference on Interactive Digital Storytelling*. Springer.