Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 476–484
Marseille, 11–16 May 2020

The Margarita Dialogue Corpus: A Data Set for Time-Offset Interactions and Unstructured Dialogue Systems

Alberto M. Chierici, Nizar Habash, Margarita Bicec
Computational Approaches to Modeling Language (CAMeL) Lab
New York University Abu Dhabi
{alberto.chierici, nizar.habash, margarita.bee}@nyu.edu

Abstract

Time-Offset Interaction Applications (TOIAs) are systems that simulate face-to-face conversations between humans and digital human avatars recorded in the past. Developing a well-functioning TOIA involves several research areas: artificial intelligence, human-computer interaction, natural language processing, question answering, and dialogue systems. The first challenges are to define a sensible methodology for data collection and to create useful data sets for training the system to retrieve the best answer to a user's question. In this paper, we present three main contributions: a methodology for creating the knowledge base for a TOIA, a dialogue corpus, and baselines for single-turn answer retrieval. We develop the methodology using a two-step strategy. First, we let the avatar maker list pairs by intuition, guessing what possible questions a user may ask to the avatar. Second, we record actual dialogues between random individuals and the avatar-maker. We make the Margarita Dialogue Corpus available to the research community. This corpus comprises the knowledge base in text format, the video clips for each answer, and the annotated dialogues.

Keywords: Corpus Creation, Annotation, Dialogue, Question Answering

1. Introduction

Time-Offset Interaction Applications (TOIAs) are specific types of question answering systems that simulate face-to-face conversations between humans (*interactors or interrogators)* and previously-recorded digital human *avatars* which were created by *avatar makers*. Such conversational avatars are important for preserving personal and cross-generational histories, as well as, for teaching and coaching, among other possible applications. Our work on TOIAs is heavily inspired by Artstein et al. (2015), who first introduced the term *time-offset interaction*.

Developing, streamlining, and making the process of creating TOIAs affordable are important goals. But building such systems is not an easy task because the interaction with the avatars should ideally be as close as possible to a real, human-to-human interaction. This critical feature poses challenges from an engineering point of view (for instance, connecting the video clips flawlessly) and from many other points of view related to dialogue management and natural language processing. We refer to the work by Abu Ali et al. (2018) for a detailed description of the TOIA platform we use for storing and accessing video clips, and for interpreting the interrogator's input.

The first step to build a TOIA involves the creation of its Knowledge Base. This database consists of a set of pairs, and answer video clips. The answer clips include recordings of the avatar's answers, and clarification requests (e.g., "Can you repeat your question?"), as well as filler sequences to play while the system is waiting for the next question (e.g., the avatar adjusting her hair). One interesting problem is how to best construct this KB in the first place. Should avatar makers use their intuition and brainstorm pairs? Or should they record and transcribe real dialogues?

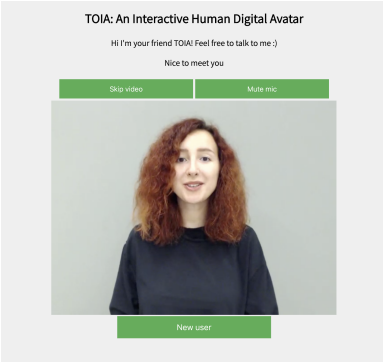

Figure 1: A screenshot from one of the video clips in the Margarita Dialogue Corpus, named after co-author and avatar maker Margarita Bicec.

In this paper, we focus on TOIA data creation. We propose a methodology for creating a 'gold' reference data set for a TOIA, and apply it in the creation of the Margarita Dialogue Corpus. This corpus includes a database of pairs and their corresponding video clips, as well as a number of annotated dialogue transcripts (see Figure 1 and Table 1). We make the methodology and corpus available to researchers interested in streamlining the creation of conversational avatars.

Next we discuss related work, and the relationship between dialogue problems, data and techniques (Section 2). Section 3 introduces our contribution in generating data sets adequate for the task of building functioning TOIAs, and the data annotation methodology. In Section 4, we present and discuss baseline models to provide an initial evaluation for our work, along with some error analysis. In Section 5, we discuss the most interesting problems encountered in the analysis of the Margarita Dialogue Corpus, such as threshold selection and corpus expansion.

2. Related Work

We start this section with discussing the work that most inspired our efforts. We then contextualize our work within the broader category of dialogue systems. Finally we introduce the most common types of training data sets, techniques and evaluations available for research in dialogues.

The New Dimensions in Testimony project The *New Dimensions in Testimony* project presented by Traum et al. (2015b) inspired us and Abu Ali et al. (2018) to work on TOIAs. Traum et al. (2015b) built a time-offset interaction with a Holocaust survivor, Mr. Pinchas Gutter. We built our TOIA following the same methodology, whereby we prepared a limited (compared to the original project's vast) set of pairs in advance. In other words, we guessed what questions might be asked of the avatar. The Dialogue Manager of Abu Ali et al. (2018) uses a statistical classifier based on traditional techniques from the IR arena. The work by Traum et al. (2015b) develops a proof-of-concept that enables short conversations about Pinchas Gutter's family, his religious views, and the resistance. This work was the source of inspiration for other groups (Nishiyama et al., 2016; Jones et al., 2015) who focused on the experience of interacting with a digital human avatar and the engineering of a fluid user experience for the player, necessarily involving high production costs. Traum et al. (2015a) are able to use speech recognition and language understanding to select sensible responses frequently enough to enable a natural interaction flow. However, while these results are achievable with about 2,000 recorded statements, we want to build a system that allows any person to make their avatar available in a relatively short period of time (i.e., recording significantly fewer than 2,000 statements), using their laptop and webcam. Thus generalizing the process to new users and unpredictable contexts.

Dialogue systems Dialogue systems fall along two central axis pairs: modular (or task-driven) versus end-to-end (or social bots), and structured versus unstructured. Modular systems, like the commercially available Siri, Alexa, Cortana, or Google Assistant, train multiple models to support a set of tasks (Guo and Seltzer, 2012), while end-to-end systems train a single learning algorithm on dialogue data (Serban et al., 2016; Shum et al., 2018). Similarly, structured systems assume a logical representation for the information exchanged in conversation - for example, slot-filling techniques (Williams et al., 2013; Fast et al., 2018) - which unstructured systems do not require (Gao et al., 2019). A dialogue system like a TOIA formulates a new category of chatbots that can be named 'self-narrative bots' (SNB). In terms of NLP tasks, they are a middle ground between social bots and task-driven bots: they may use a combination of structured and unstructured data for training. SNBs will be able to understand an interrogator's question in order to match it to a sensible answer. At the same time, they must be able to engage in a multi-turn conversation, hence they may use reading comprehension of contexts represented by a sequence of pairs rather than a standalone question. For such a system, an approach that is interesting to study would be a combination of algorithms and evaluations that come from the Information Retrieval (IR) context and the sequential approach (i.e., learning the sequence of a conversation).

Question	Answer
"Do you miss the food, the Moldovan cuisine?"	"I used to more then, but now I got so used to just changing my diet depending on where I am at that time. And also I found some place where we can eat Russian food and the dining hall is also making Russian food from time to time... so I don't miss it that much."
"How do you make money with music?"	"By being good at what you do and knowing people. That's how you get a job in the music industry and you grow from there."
"I never asked about your major?"	"I studied music and economics. I'm a music major, economics minor and in music I do mostly composition and sound engineering."

Table 1: Three examples of pairs in the Margarita Dialogue Corpus KB data resource.

Dialogue data, techniques, and evaluation Social bots are expected to entertain the user hence and are often evaluated by the number of turns they can make in conversations (Khatri et al., 2018).[1] They are designed to address two primary Natural Language Processing (NLP) challenges: understanding the user question, and generating a sensible answer. From the IR standpoint, they must also search into a large enough KB to retrieve the right answer(s) across many different topics. These KBs usually come from different contexts such as Twitter or Wikipedia pages (Ritter et al., 2010; Wilcock, 2012). Some dialogue management systems use question answering techniques, which usually address a reading comprehension task. For instance work on the SQuAD dataset (Rajpurkar et al., 2016), the Ubuntu dialogue corpus (Lowe et al., 2015), and bAbI[2] (Weston et al., 2015) are designed to perform time reasoning and inductive logic (Kumar et al., 2016). Techniques include recurrent neural network models (RNN) such as sequence2sequence, word embeddings, and LSTMs. It is worth mentioning that word or sentence pre-trained embeddings alone are a simple tool to produce powerful results. Recent examples are Google's BERT (Devlin et al., 2018), Facebook's InferSent (Conneau et al., 2017), and OpenAI's GPT (Radford et al., 2018). While large data sets are available and the recent success of deep learning techniques suit big data, we make available a smaller data set aimed at the practical implementation of a TOIA available for the every-day user. Finally, it is worth pointing out that the evaluation of dialogue systems is an area of research where - to the best of our knowledge - no established or robust methodology seems to exist yet for two main inter-related reasons: automatic metrics do not correlate well with human judgments, and human judgments are difficult to measure (Li et al., 2019).

[1] https://developer.amazon.com/alexaprize/challenges/current-challenge/rules

[2] https://research.fb.com/downloads/babi/

Category Class	Examples	Frequency in KB	(in %)
Meta-interactions	Hinting to user what to ask about or providing diversions.	62	7%
Pleasantries & Short Answers	Greetings, yes/no answers, compliments.	159	18%
Personal Information	Family, country, past, future, love, etc.	362	41%
New York University Abu Dhabi	University life, admissions, courses, and life in the United Arab Emirates.	309	35%
	Total Frequency	**892**	

Table 2: Summary of the categories in the Margarita Dialogue Corpus Knowledge Base (KB) defined by the avatar maker.

Statistics	Knowledge Base	Dialogues (All)	Dialogues Train	Dialogues Test	Dialogues EDU Mode	Dialogues PER Mode
# dialogues	NA	20	10	10	10	10
# q-a pairs (in total)	892	659	340	319	296	363
# unique questions	758	NA	NA	NA	NA	NA
# unique answers	431	NA	NA	NA	NA	NA
# annotated answers	NA	888	472	416	421	467
# no-answers	NA	49	0	49	25	32
(in %)	*(NA)*	*(15%)*	*(0%)*	*(15%)*	*(8%)*	*(9%)*
# words (in total)	20,303	40,557	20,230	20,327	20,084	20,473
Min. # turns per dialogue	NA	22	22	26	22	24
Avg. # turns per dialogue	NA	33	34	32	30	36
Avg. # words per question	7.75	14.5	14.5	14.6	16.3	13.1
Avg. # words per answer	15.0	47.0	45.0	49.1	51.5	43.3

Table 3: Summary statistics on the two main data sets in the Margarita Dialogue Corpus: Knowledge Base (KB) and Dialogues. Statistics for the dialogues are also shown for the train portion vs. test and university mode (EDU) vs. personal mode (PER).

3. Data Acquisition and Annotation

We propose that the best way to create the KB of a TOIA involves two methodologically different steps. First, the avatar maker can brainstorm several question-answer pairs. Second, the avatar maker records real dialogues with different people. In this way, questions that the avatar maker may have not brainstormed, and yet do happen in real conversations can be covered. Such examples include introductions and greetings question-answer pairs like "Hi-Hello", "How are you?-I'm fine, thank you", "Goodbye-Bye bye!".

Though our TOIA architecture allows for creating avatars quickly and with low production costs, we appreciate that recording conversations might not be the most convenient option for streamlining the avatar creation process. We propose this methodology as a starting point to have enough question-answer pairs to cover common conversation topics that the avatar maker may have not thought about in the prior brainstorming. Moreover, recording real dialogue gives insight into other people's reactions and picks up on topics that depend on the interrogator's background. Starting from a user experience perspective, what would be the ideal world experience for the interrogator? She or he should be able to question an avatar in a fluent conversation that would mirror the real experience of getting to know a stranger in a 10 to 15 minute interaction.

We selected twenty subjects as interrogators, making sure they did not know the avatar maker in person, and we instructed each of them to engage in a 15-minute conversation with the avatar maker. We instructed the avatar maker to avoid asking questions back to the interrogator, although this naturally happened in a few dialogue turns. When interrogators were not sure what to ask, they could ask the avatar maker what questions or topics she could talk about. Part of these recordings (named 'train set' in later sections) were used to define the KB, and part of the recordings were used as held-out test samples (named 'test set' in later sections) to evaluate the baseline models. Moreover, we wanted data to be both 'on-topic' and 'wild', to study two different avatar interactions: making an avatar who can act as an information kiosk for a university, and an avatar who can talk about herself. We forced half of the conversation to be about one topic, called the 'university mode' (or 'EDU'): the avatar maker, a student at New York University Abu Dhabi (NYUAD), could only answer questions about the campus and academic life. The second half of the dialogues did not have a set topic, and we call this the 'personal mode' (or 'PER'): we asked the interrogator to get to know the avatar maker as one would do when meeting a person for the first time. In this upcoming section, we dive into the specific data collection and annotation methodologies.

Knowledge Base and Dialogues We initially let the avatar maker brainstorm question-answer pairs 'out of context', meaning that the questions and answers were not part of a dialogue flow between two individuals. She defined 241 pairs of questions and answers. We then recorded 20 dialogues with real people, 10 about the university (EDU mode), and 10 about the avatar maker (PER mode). We randomly selected 10 conversations (5 in university mode and 5 in personal mode) and used them as the 'train' set, and we used the other 10 conversations as the 'test' set. We

then asked the avatar maker to check which questions in the recorded conversations were not present in the initial question-answer pairs brainstormed 'out of context'. The new questions were then added to form the avatar's KB, and the avatar maker recorded videos answering these questions. As part of this consolidation work, some questions with similar meanings were kept in the KB and were assigned the same answer. For instance, "I like your... / I love your... / You are so... / Nice... / That's interesting!" are paired with the same answer, "Awww thanks!" in the KB. The resulting KB has 892 question-answer pairs and 431 unique answers that are also available in video clips. The avatar maker recorded also videos like "Can you ask me something else?", "Can you repeat that?", "Could you elaborate?" and so forth. The system would play them when the ranking model does not produce a result within a certain level of confidence. We categorized these answers as 'unsure' in the KB.

We summarize the statistics of the two data sets, the 'Knowledge Base' and the 'Dialogues', in Table 3.

Annotation After the data was collected, the avatar maker took the role of the annotator. Hereafter, we will use the words 'annotator' and 'avatar maker' interchangeably. She engaged in an exercise similar to a 'Wizard of Oz' setting. In a typical Wizard of Oz setup, the interrogator would query the system on a screen, and behind the scenes, there would be a real person - the wizard - selecting an appropriate answer to play to the interrogator. We performed the exercise in a 'post hoc' fashion, which may be less prone to error because it removes the time pressure to play an answer immediately. The avatar maker impersonated the wizard, and she had at her disposal all the answers recorded in the Knowledge Base. She then paired each question appearing in the dialogues data (i.e., from the new conversations recorded) to the first-best answers available in the Knowledge Base, the second-best answer, the third, until the sixth-best answer.

The purpose of this exercise is to build two 'oracle' data sets. The 'train' portion of the dialogues will have an answer for each question guaranteed because the interrogators' recorded questions were used to consolidate the KB, as described in the paragraph above. The 'test' portion of the dialogues will have some questions without an answer in case none of the answers available in the KB are a good fit for a given question. We say both the 'train' and 'test' sets are 'oracle' data because it is as if the annotator already knew all the questions that were asked by interrogators, and could identify first if there was an answer in the KB, and if so, what were the best answers for any given question from the KB, excluding the actual answers the avatar maker gave in the recording session.

The annotator also subjectively categorized her answers, and we make them available in the data. She defined 66 categories (68, excluding the 'unsure' category, and the 'filler' category which we describe in the next paragraph). About one-third of the categories account for 80% of the KB, showing a typical power-law behavior. The top 10 categories, accounting for more than half of the KB, are about music, pleasantries, opinions, compliments, applying to NYUAD, languages, NYUAD in general, positive memories, travel, and short answers. These categories reflect well common conversation topics for the two contexts we defined (information about the university and meeting a person for the first time) as well as the personality of the annotator/avatar-maker. We show high-level statistics related to these categories in Table 2. Given the subjective, unstructured way these categories were defined, further work would be needed to study their applicability. For instance, they might be grouped into macro-topics in a way similar to that Table 2.

It is worth pointing out that the KB contains two categories that we eliminated for creating the baselines. The 'filler' category corresponds to videos where the avatar is making gestures to fill video space between an answer and the next question the interrogator will ask. So this category is useful only for a user experience perspective when using the TOIA.

The data has been verified by the first two co-authors besides the avatar maker (third co-author). The corpus is named the 'Margarita Dialogue Corpus' after the avatar's first name, and it is available to download at NYUAD CAMeL Lab's Resource page.[3]

4. Baseline Models and Benchmarks

We created simple baselines by using an information retrieval methodology. We report metrics that are relevant for single-turn interactions. The data set though is well suited for multi-turn dialogues and we are developing further work on multi-turn metrics, including human evaluations. We used three different models to convert sentences into sentence vectors, then we computed the distance between an interrogator's question-vector and all the vector-representations of all the questions present in the KB. The distance gives us a ranking function for every answer in the KB: the closer the question relative to a given answer in the KB for an interrogator's question in the sentence vector space, the higher the rank of the answer in the KB as a potential reply for the interrogator's question. Now, there are two tasks we need to get right. First, the model should be able to decide whether or not, for any question posed to the system, there exists an answer in the KB. Then it must identify what the best answer is. In our proposed baselines, we tackled the two questions simultaneously by thresholding on the train set, i.e., we came up with a heuristic to decide if the value of the distance (or similarity) metric is high enough to ascertain whether or not an interrogator's question is indeed similar to a question within the KB. If all the questions in the KB have their distance metric below the selected threshold, the system decides to output a 'no-answer' message. If there are questions in the KB with distance metrics above the threshold, we rank their corresponding answers as candidates for the interrogator's question using the value of the distance metric itself. In the future we plan to use more sophisticated approaches, including machine learning models on the two separate problems.

[3] `http://resources.camel-lab.com/`. Go to 'Corpora', then to 'Margarita Dialogue Corpus'.

4.1. Evaluation metric

Inspired by Lowe et al. (2015), Schatzmann et al. (2005) and Bleu (Papineni et al., 2002) (which uses multiple references), we evaluated the baseline models using a multi-reference Recall@k metric. We tasked the baseline models to select the k most likely responses, and the metric accounts for the true answer lying within the top k candidate responses. In practice, for real-world TOIAs, only the Recall@1 metric would be relevant. Although our database provides a rank of choices, we consider them to be equal and leave the ordering to further work (i.e., making models that give to the 'first-best answer' a higher ranking than the 'third-best answer').

4.2. Models

TF-IDF The first model uses the term frequency-inverse document frequency (TF-IDF) statistics. This quantity should capture how important a given word is to some document, which in our case is the question (Ramos and others, 2003). TF-IDF is a technique that is often used in document classification. The 'term-frequency' is the count of the number of times a word appears in a given document, and the 'inverse document frequency' is a multiplier that penalizes how often the word appears elsewhere in the overall collection of documents (the corpus). The statistic is defined as

$$TFIDF(w, d, D) = f(w, d) \cdot log\frac{N}{|\{d \in D : w \in d\}|},$$

where $f(w, d)$ is the term frequency of the word w into document d, N is the total number of documents, and the denominator represents the number of documents in which the word w is present. Questions and answers are transformed into TF-IDF vectors, returning k answers corresponding to the top k cosine similarities between a test questions in the test set and the questions in the training set.
Given the size of this data set, yet trying to leverage state-of-the-art results achieved by deep learning in other question answering contexts, we make an attempt to leverage pre-trained models that generalize well in other contexts. We chose InferSent by Conneau et al. (2017) and BERT by Devlin et al. (2018) because of their generalization power and versatility towards NLP tasks.

InferSent We investigated the InferSent pre-trained word embeddings proposed by Conneau et al. (2017) as a second baseline. Similarly to TF-IDF, we developed the matching technique by checking the cosine similarity of word vectors between the validation or test sets questions and the training set answers.

BERT The Bidirectional Encoder Representations from Transformers (BERT) method was one of the breakthroughs in NLP in late 2018 (Devlin et al., 2018). BERT is a method of pre-training language representations and it is useful for transfer learning tasks. It can be used to extract high-quality language features from any text data, as well as to fine-tune the model on a specific task like classification, entity recognition, or question answering. In particular, we use BERT to extract word and sentence embedding vectors, again calculating cosine similarities. It is worth pointing out that BERT offers an advantage over models like Word2Vec because while each word has a fixed representation under Word2Vec regardless of the context within which the word appears, BERT produces word representations that are dynamically informed by the words around them.

Statistics	KB'	Train Dialogues	Test Dialogues
# q-a pairs (in total)	776	401	319
# unique questions	698	NA	NA
# unique answers	398	NA	NA
# no-answers	NA	61	57
(in %)	*(NA)*	*(15%)*	*(18%)*

Table 4: Summary statistics after sampling non-answers for the train set: down-sampled KB (KB') and up-sampled train dialogues. The train set remained the same with minor changes due to answers no longer present in the KB'.

We implement the three vector representations of sentences described in the previous sections, namely TF-IDF, InferSent, and BERT, without pre-processing tokens. We then compute the similarity to questions in the KB to establish whether we have an answer for a given question as well as ranking the answers to retrieve. We considered the similarity between a new question and every answer in the KB as an alternative model, but the results are so much weaker than question-similarity that it is not worth reporting them.

4.3. Confidence threshold selection

Threshold	TF-IDF	InferSent	BERT
0.05	0.277	0.234	0.269
0.1	0.277	0.234	0.269
0.15	0.277	0.234	0.269
0.2	0.277	0.234	0.269
0.25	0.277	0.234	0.269
0.3	0.277	0.234	0.269
0.35	0.282	0.234	0.269
0.4	0.287	0.234	0.269
0.45	0.307	0.234	0.269
0.5	0.317	0.234	0.269
0.55	0.342	0.234	0.269
0.6	0.352	0.239	0.269
0.65	0.357	0.242	0.269
0.7	0.362	0.252	0.269
0.75	0.362	0.267	0.272
0.8	**0.369**	0.282	0.294
0.85	0.357	0.289	0.307
0.9	0.339	**0.302**	**0.332**
0.95	0.312	0.284	0.319

Table 5: Recall@1 statistics for each baseline evaluated on the train set. The metrics corresponding to the thresholds that were automatically selected are bolded.

A particular challenge seems to be the setting of a confidence threshold for the ranking function (in all models, the cosine similarity between word vectors) to decide if in low similarity cases the system should give the top-ranked answers a non-answer. The presence of non-answers in the test set influences the performance metrics. 15% of the questions in the test set do not have an appropriate answer

Threshold	# Correct Answer	# Correct Non-Answer	TPR-ans	TPR-non-ans	Recall@1
0.05	108	0	0.269	0.000	0.269
...	...	...	...	...	...
0.7	108	0	0.269	0.000	0.269
0.75	108	1	0.339	0.018	0.272
0.8	107	11	0.335	0.193	0.294
0.85	103	20	0.323	0.351	0.307
0.9	**91**	**42**	**0.285**	**0.737**	**0.332**
0.95	72	56	0.226	0.982	0.319

Table 6: Thresholding considerations for the BERT model. The table shows, for each threshold level imposed to cosine similarities, the number of correct answers predicted, the number of correct non-answers predicted by the model, hence the true positive rate for answers (TPR-ans), the true positive rate for non-answers (TPR-non-ans) and the Recall@1 metric. The automatic choice of threshold for this setup is highlighted in bold face.

selected by the annotator. Given that we use the 'train' set to select the confidence threshold, we need to make some adjustments because the train set does not have any 'non-answer'. We up-sample questions by picking questions in the KB whose answers were not selected by the annotator in the train portion of the dialogues. We then remove the corresponding question-answer pairs from the KB. We show the resulting, 'adjusted' KB (KB'), train set and test set statistics in Table 4. We use a simple heuristic that accounts for the trade-off between answers' true positives rate (TPR-ans) and non-answers' true positive rate (TPR-non-ans): We give more weight to the TPR-ans because there are more answers than non-answers in the annotated data sets. Moreover, it is easier for the model to predict a non-answer by merely picking a high threshold, whereas the task to select the right answer (when there is one) is more complex and more interesting task from an NLP standpoint. A high recall - like the Recall@1 metrics seen in Table 5 for high threshold values - may be misinterpreted as a good result when all the model is doing is achieving maximum accuracy on the 15% of the test set's examples that have no answers. See, for example, a look 'under the hood' of the Recall@1 metrics for the BERT model at different threshold levels in Table 6.

4.4. Results

	TF-IDF	InferSent	BERT
Similarity Threshold	0.8	0.9	0.9
Recall@1	0.194	0.169	**0.201**
Recall@2	**0.207**	0.169	**0.207**
Recall@5	**0.210**	0.169	**0.210**
Recall@10	0.210	0.169	**0.213**
Recall@20	0.210	0.169	**0.213**

Table 7: Results for each baseline on the test set. For each model and threshold selection, the Recall@k metric is shown for different levels of k on the test set.

The results (see Table 7) show that these baselines are limited to achieving an optimal user experience. It is somewhat surprising to see that a pre-trained model like InferSent, that generalizes well in other NLP tasks, does not improve the results of the more traditional TF-IDF statistics. This level of performance might be due to the size of the problem, and perhaps further parameter tuning might lead to better results. Although BERT shows the best results, the metrics are only slightly better than the TF-IDF model.

	TF-IDF	InferSent	BERT
Recall@1	0.103	0.046	**0.111**
Recall@2	**0.168**	0.084	0.153
Recall@5	**0.240**	0.126	0.221
Recall@10	0.282	0.164	**0.302**
Recall@20	0.363	0.210	**0.420**

Table 8: Recall@k only for answers (i.e., ignoring questions in the test set that did not have and answer in the KB) for each baseline. For each model Recall@k metric is shown for different levels of k on the test set.

To gauge a better performance differentiation between the baselines, we simplified the problem by focusing only on the ability of the distance metrics to pick up the right answer within the first, second, first-five, first-ten, and first-twenty ranked responses in the KB independently of the threshold. In other words, by forgetting about being able to give a non-answer, how well do we rank the answers? Table 8 shows these results by considering only the Recall@k for answers rather than the former Recall@k that takes into account both answers and non-answers. We can see that BERT embeddings do a much better job than InferSent embeddings and, although not far from the TF-IDF model's performance, they improve the TF-IDF performance materially when looking at the TPR in the first-twenty ranked responses.

4.5. Error analysis

Table 9 shows a few meaningful examples from the error analysis, where we can make some hypotheses on how the baselines work and point out the challenges associated with the threshold selections.

In the first example, all baselines pick up the right question in the KB (hence the right answer). However, the cosine similarity for InferSent sentence embedding falls below the model's selected threshold (0.9), so for that instance, the system would return a non-answer. In the second example, the test example's question is 'So where were you raised?' and both TF-IDF and BERT models pick up the right question in the KB: 'Where are you from?'. However, for all baselines, we are below the threshold, so that the system would return a non-answer. The system would be correct for the InferSent case as a similarity score of 0.741 is not

	Top Ranked Similar Question in the KB' (Cosine Similarity)		
Test Question	**TF-IDF**	**InferSent**	**BERT**
"So is this your natural hair color?"	"Is this your natural hair color?" (0.972)	"Is this your natural hair color?" (0.847)	"Is this your natural hair color?" (0.965)
"So where were you raised?"	"Where are you from?" (0.458)	"When did you graduate?" (0.761)	"Where are you from?" (0.825)
"Oh my God. That could have been. That's wild. And do you miss home when you're here in Abu Dhabi?"	"Do you miss home a lot?" (0.424)	"What can I do for fun in Abu Dhabi?" (0.815)	"Do you have siblings who study in New York or Abu Dhabi?" (0.803)
"When are you heading back home?"	"Are most people back home Orthodox?" (0.566)	"Where are you from?" (0.774)	"How far are you guys from the city?" (0.822)

Table 9: Top ranked question similarity for some meaningful examples in the test set for all three baselines.

high enough to be confident it corresponds to the right answer, but it is not correct to state that there is no answer in the KB. The examples point to the direction of separating the two problems ('Do we have an answer in the KB?' and 'Which one?').

The third example seems to show to what meanings different sentence embeddings give importance. Though all the answers do not make it above the thresholds, TF-IDF correctly identifies the best matching question in the KB. TF-IDF seems to spot the 'miss home' keywords, whereas InferSent and BERT seem to pick up on a location ('Abu Dhabi'). BERT might also give importance to a certain meaning of the word 'home' in the test question, matching something related to 'siblings' in the KB.

The fourth example shows that TF-IDF again functions by keywords matching ('back home'), whereas InferSent and BERT embeddings seem to capture different meanings of the words 'back home'. InferSent picks up provenience, and BERT looks like identifying geographic distance.

These examples show how different sentence embeddings weight keywords or try to capture meanings or other elements within a sentence. They also show that the vector space model they infer to sentences is not comparable and may mean completely different things. Cosine similarities in different vector spaces have different scales too: we could also notice it in Table 5 where TF-IDF's results tend to be more evenly spread across thresholds whereas InferSent and BERT have values more clumped towards high cosine similarities.

5. Outlook

Our work on building a TOIA corpus and working on baselines has exposed the following challenges.

Threshold selection We performed a confidence threshold selection with a simplistic heuristic. We are looking into separating the problems of establishing if an answer exists in the KB and what is the best answer. Regarding the confidence threshold to select for establishing if the system can answer a question, two directions for improvement include expanding the KB building upon the work of Traum et al. (2015a), and automating the scoring of chatbot responses as in the work by Yuwono et al. (2019).

Word tagging and entity recognition As pointed out in the error analysis, we could already improve some baseline results by exploring and implementing models for word tagging, entity recognition, or semantic parsing. In this way, a TOIA's dialogue manager should be able to differentiate between sentences that have the same meaning and words, apart for just one word (usually the predicate's object) that points to a completely different answer. An example of this would be "Do you have any siblings?" vs. "Do you have any pets?".

Accounting for annotated ordering The annotator indicated the order of answers from the most plausible in a given point of the conversation to the least plausible. We could develop a more sophisticated evaluation metric for taking into account this information. To better assess a language model, we can compare the answer-ranking provided by the model with the ranking provided by the annotator. An idea would be to modify the Recall@k metric such that we give weight to the ordering of the first k ranked answers rather than counting the mere presence or absence of correct answers within the first k ranked answers.

Human evaluation Different or additional evaluation methodologies could be drawn from the HCI community. For example, recent work by Amershi et al. (2019) propose eighteen generally applicable design guidelines for human-AI interaction for practitioners working on the design of applications and features that use AI, and to researchers interested in the development of guidelines for human-AI interaction design. Moreover, in the context of unstructured multi-turn dialogue modeling, the most used automatic evaluation metrics are biased and correlate poorly with human judgments of response quality (Lowe et al., 2017). Improving the human evaluation framework is indeed an important research direction for the data set proposed in this paper as well as dialogue systems in general, and the work by Li et al. (2019) proposes a novel, exciting perspective.

Expand corpus creation One point emphasized throughout this work is the importance of the data size. For a single avatar, more data shall be recorded and annotated, as suggested by Artstein et al. (2015). To understand how much data is enough for a time-offset conversation, one approach

could be to create synthetic avatars by borrowing dialogues from other data sets such as movie scripts. Another exciting approach would be to expand the corpus by creating more avatars and to collect a large number of user interactions.

Using richer features The baselines presented here do not use further information added by the category annotations, the two different modes (university and personal) of conversation, or the multi-turn nature of the dialogue data sets. Machine Learning approaches can leverage the additional information for both classifying a question as having an existing answer in the KB or not, and better ranking answers. Moreover, structuring the train and test data by combining more dialogue turns can be used to fine-tune deep learning models such as BERT in a broader context. For example, such a structure might address what the best next answer is for a context like a -question triplet. To improve the task of selecting the right answer, we could leverage the sequential dialogues and pairing of every question with plausible answers and sampling implausible answer from the KB, building on the approach introduced by Lowe et al. (2015). The hypothesis would be that a system that tracks a dialogue sequentially might be better suited for both selecting the right answer and engaging the user in a social conversation. We considered the implementation of deep learning models, especially RNNs and LSTMs, but setting up a model comparable to one that produces state-of-the-art results led us to abandon this approach because the number of parameters far exceeds the size of the data sets examined in this work.

Self-narrative long recordings An alternative, and perhaps more challenging, route for the development of time-offset interaction with avatars is for the avatar maker to record a long, self-narrative video. The system would then operate with reading comprehension algorithms to play only the video clip snippet that corresponds to the best answer to an interrogator's question.

Context transferability One other area of investigation is using the same avatar for a different context. For instance, the Margarita Dialogue Corpus avatar maker created two corpora for two different contexts, namely providing information about New York University Abu Dhabi and speaking about herself when introduced to a stranger. Other contexts of interest could be self-narrative for a curriculum. The exciting aspect will be to study if there are parts of the dialogue that we can transfer between contexts, e.g., the avatar's talking style, jargon, or vocabulary.

6. Conclusion and Further Work

This work proposed an original approach to collect and annotate data for training and evaluating a Time-Offset Interaction Application (TOIA). We work with two types of data sets: an intuition-based, single-turn knowledge base, and in-context, multi-turn annotated dialogues. The data annotation involves the avatar maker running a 'post-hoc' wizard of Oz. We make the Margarita Dialogue Corpus, including the recorded video clips of the avatar maker's answers, available to the research community.[4] We implemented three baselines for laying down the basis to improve the answer selection for a TOIA that allows anyone to create self-made avatars in a relatively short time frame and low cost though we realize further work is needed for streamlining the process. While the baselines report on single-turn metrics, the Margarita Dialogue Corpus is also suited for research into unstructured, multi-turn dialogues with low-resources data and transfer learning tasks. Some of the results observed point to interesting research paths such as improving the best answer confidence threshold, defining the correct setup for human evaluation, or understanding how much data is enough for a TOIA.

In the future we plan to continue improving the models for answer selection, streamlining the evaluation methodology, and for engineering the best user experience. We plan to expand the models to work in multilingual settings, building on the bilingual avatars introduced by Abu Ali et al. (2018). We also plan to create additional avatars for other avatar makers.

[4] http://resources.camel-lab.com/

Acknowledgments

We would to thank David Traum for helpful conversations. We also would like to thank the NYUAD TOIA team (Dana Abu Ali, Muaz Ahmad, Hayat Al Hassan, Paula Dozsa, Ming Hu, and Jose Varias) for making the TOIA software they created available to us.

Bibliographical References

Abu Ali, D., Ahmad, M., Al Hassan, H., Dozsa, P., Hu, M., Varias, J., and Habash, N. (2018). A bilingual interactive human avatar dialogue system. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 241–244.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., et al. (2019). Guidelines for human-ai interaction.

Artstein, R., Leuski, A., Maio, H., Mor-Barak, T., Gordon, C., and Traum, D. (2015). How many utterances are needed to support time-offset interaction? In *The Twenty-Eighth International Flairs Conference*.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fast, E., Chen, B., Mendelsohn, J., Bassen, J., and Bernstein, M. S. (2018). Iris: A conversational agent for complex tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 473. ACM.

Gao, J., Galley, M., Li, L., et al. (2019). Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298.

Guo, P. J. and Seltzer, M. I. (2012). Burrito: Wrapping your lab notebook in computational infrastructure.

Jones, A., Unger, J., Nagano, K., Busch, J., Yu, X., Peng, H.-Y., Alexander, O., Bolas, M., and Debevec, P. (2015).

An automultiscopic projector array for interactive digital humans. In *ACM SIGGRAPH 2015 Emerging Technologies*, page 6. ACM.

Khatri, C., Hedayatnia, B., Venkatesh, A., Nunn, J., Pan, Y., Liu, Q., Song, H., Gottardi, A., Kwatra, S., Pancholi, S., et al. (2018). Advancing the state of the art in open domain dialog systems through the alexa prize. *arXiv preprint arXiv:1812.10757*.

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., and Socher, R. (2016). Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning*, pages 1378–1387.

Li, M., Weston, J., and Roller, S. (2019). Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y., and Pineau, J. (2017). Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*.

Nishiyama, M., Miyauchi, T., Yoshimura, H., and Iwai, Y. (2016). Synthesizing realistic image-based avatars by body sway analysis. In *Proceedings of the Fourth International Conference on Human Agent Interaction*, pages 155–162. ACM.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142. Piscataway, NJ.

Ritter, A., Cherry, C., and Dolan, B. (2010). Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180. Association for Computational Linguistics.

Schatzmann, J., Georgila, K., and Young, S. (2005). Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *6th SIGdial Workshop on DISCOURSE and DIALOGUE*.

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Shum, H.-Y., He, X.-d., and Li, D. (2018). From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26.

Traum, D., Georgila, K., Artstein, R., and Leuski, A. (2015a). Evaluating spoken dialogue processing for time-offset interaction. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 199–208.

Traum, D., Jones, A., Hays, K., Maio, H., Alexander, O., Artstein, R., Debevec, P., Gainer, A., Georgila, K., Haase, K., et al. (2015b). New dimensions in testimony: Digitally preserving a holocaust survivor's interactive storytelling. In *International Conference on Interactive Digital Storytelling*, pages 269–281. Springer.

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A., and Mikolov, T. (2015). Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Wilcock, G. (2012). Wikitalk: A spoken wikipedia-based open-domain knowledge access system. In *Proceedings of the workshop on question answering for complex domains*, pages 57–70.

Williams, J., Raux, A., Ramachandran, D., and Black, A. (2013). The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413.

Yuwono, S. K., Wu, B., and D'Haro, L. F. (2019). Automated scoring of chatbot responses in conversational dialogue. In *9th International Workshop on Spoken Dialogue System Technology*, pages 357–369. Springer.