

A View From The Crowd: Evaluation Challenges for Time-Offset Interaction Applications

Alberto M. Chierici and Nizar Habash

Computational Approaches to Modeling Language (CAMEL) Lab

New York University Abu Dhabi, UAE

{alberto.chierici, nizar.habash}@nyu.edu

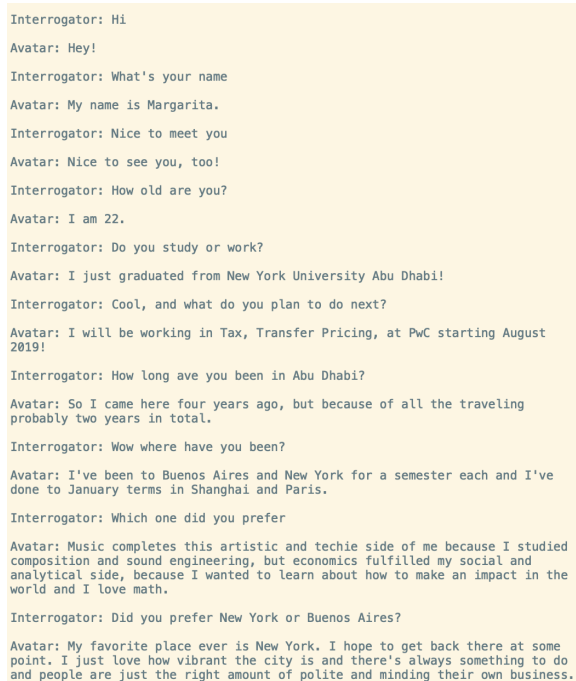
Abstract

Dialogue systems like chatbots, and tasks like question-answering (QA) have gained traction in recent years; yet evaluating such systems remains difficult. Reasons include the great variety in contexts and use cases for these systems as well as the high cost of human evaluation. In this paper, we focus on a specific type of dialogue systems: Time-Offset Interaction Applications (TOIAs) are intelligent, conversational software that simulates face-to-face conversations between humans and pre-recorded human avatars. Under the constraint that a TOIA is a single output system interacting with users with different expectations, we identify two challenges: first, how do we define a ‘good’ answer? and second, what’s an appropriate metric to use? We explore both challenges through the creation of a novel dataset that identifies multiple good answers to specific TOIA questions through the help of Amazon Mechanical Turk workers. This ‘view from the crowd’ allows us to study the variations of how TOIA interrogators perceive its answers. Our contributions include the annotated dataset that we make publicly available and the proposal of Success Rate @k as an evaluation metric that is more appropriate than the traditional QA’s and information retrieval’s metrics.

1 Introduction

Time-Offset Interaction Applications (TOIAs) (Artstein et al., 2015) are a sort of chatbot applications that lie between Question Answering (QA) and Information Retrieval (IR). They differ from QA in that a TOIA’s task is **not** about demonstrating comprehension of a text span (Rajpurkar et al., 2016; Reddy et al., 2019) but selecting a single (one-shot) appropriate answer from a restricted set of answers, a problem also known as Answer Retrieval (AR) or retrieval-based dialogue (Boussaha et al., 2019).

Ideal TOIA interactions are expected to mirror a dialogue with a real person, including all the pos-



```
Interrogator: Hi
Avatar: Hey!
Interrogator: What's your name
Avatar: My name is Margarita.
Interrogator: Nice to meet you
Avatar: Nice to see you, too!
Interrogator: How old are you?
Avatar: I am 22.
Interrogator: Do you study or work?
Avatar: I just graduated from New York University Abu Dhabi!
Interrogator: Cool, and what do you plan to do next?
Avatar: I will be working in Tax, Transfer Pricing, at PwC starting August 2019!
Interrogator: How long ave you been in Abu Dhabi?
Avatar: So I came here four years ago, but because of all the traveling probably two years in total.
Interrogator: Wow where have you been?
Avatar: I've been to Buenos Aires and New York for a semester each and I've done to January terms in Shanghai and Paris.
Interrogator: Which one did you prefer
Avatar: Music completes this artistic and techie side of me because I studied composition and sound engineering, but economics fulfilled my social and analytical side, because I wanted to learn about how to make an impact in the world and I love math.
Interrogator: Did you prefer New York or Buenos Aires?
Avatar: My favorite place ever is New York. I hope to get back there at some point. I just love how vibrant the city is and there's always something to do and people are just the right amount of polite and minding their own business.
```

Figure 1: Initial ten turns of a casual dialogue between an interrogator and a TOIA avatar using our best performing retrieval model.

sible directions it may take — which naturally has great ramification on evaluation: when two people meet and engage in casual conversation, questions may range over different topics and depending on the answer to a specific question, different conversational question-answer threads may unravel. Also, not unexpected, different answers to a specific question can be acceptable and not cause a change in the overall conversational flow. So, how can we answer the question *what is a ‘good’ (i.e., ‘right’, ‘correct’ or ‘relevant’) answer?*

We explore this question using a publicly available dataset that was manually annotated by its avatar maker – the Margarita Dialogue Corpus (MDC) (Chierici et al., 2020). The best performing

IR model we could produce had a low Recall@1 – 24% on the development (dev) set and just below 10% on the test set. When trying to interact with this avatar, one would expect to get a wrong answer about 1/4th of the times or less. However, when chatting with the avatar using the best performing model (Figure 1), we could see that the system wasn't so bad in entertaining and holding a conversation. Hence we asked a 'crowd' of human annotators to give their opinion, and we learned that the task to define the correct answers isn't straightforward, primarily due to misaligned expectations about answer relevance. We explore a number of metrics and single out Success Rate @ k (SR@ k) as the most pertinent metric for optimizing TOIAs. Our contributions include the annotated dataset that we make publicly available, and the introduction of SR@ k as the best metric for evaluating TOIAs.

We present previous work on TOIAs and related datasets in Section 2. Sections 3, 4, and 5 introduce the corpus, retrieval models, and annotation process we used, respectively. We present our results and analysis in Section 6, and discuss further in Section 7. In Section 8, we share how to apply this work's conclusions to develop better avatars.

2 Related Work

We present a number of recent TOIAs, and data sets relevant for their study and development. While most of the related work focuses on large corpora, working with small datasets and addressing evaluation issues of TOIAs are interesting, practical problems both for the IR (what is 'relevant?') and the NLP communities (transfer learning and low resources corpora).

2.1 Recent TOIAs

TOIAs have applications in a number of practical scenarios. For example, they are used for keeping historical memories (Traum et al., 2015b), job interview practice for young adults with developmental disabilities,¹ and building digital humans across different industries.^{2,3} The most recent TOIAs involve significant production costs, they are mainly used as museum attractions or training prototypes for the army, and they require recording about 2,000 answers for building an avatar (Nishiyama et al., 2016; Jones, 2005). While these works focus more

¹<https://ict.usc.edu/prototypes/vita/>

²www.soulmachines.com

³www.storyfile.com

on the overall system architecture, components and the avatar creation methodologies, their evaluation has seldom been addressed. Furthermore, research into time-offset interactions needs to generalize and streamline the avatar development process. A first attempt made by Abu Ali et al. (2018) goes towards this direction and includes the possibility to chat with the avatars in different languages. We develop our TOIA using their open-sourced architecture. Building and democratizing access to this technology is an interesting problem, and defining the right evaluation setup is a critical step forward.

2.2 The Evaluation Problem

Traum et al. (2015a) report that their TOIA gives relevant direct answers to 60-66% of user utterances, and that seems to be good enough from "informal impressions from current testing at a museum." However, we don't have to date a rigorous study about how multiple users of TOIAs evaluate such interactions. Regarding the evaluation task in adjacent fields such as IR and QA, this is often criticized and remains an open problem (Liu et al., 2016). IR systems focus on the relevance of a set of documents retrieved and 'relevance' itself is a notion not exempt from criticism (Manning et al., 2008). The evaluation metrics mostly reported are the Mean Average Precision (MAP) and the Mean Reciprocal Rank (MRR). Applying these metrics to question-answer (q-a) retrieval makes it difficult to compare systems. In the context of TOIAs, we only care about the single retrieved answer as the most relevant. Moreover, MAP and MRR are influenced by how many relevant q-a pairs exist or are retrieved by the system (more on this in Section 7) so it's difficult to compare results across different datasets or annotation methodologies of the same dataset. Other metrics like Recall@ k also depend on the number of relevant q-a pairs. For example Lowe et al. (2015) report Recall@ k by picking the right answer and 10 randomly sampled distractors, rather than computing a relevance score between a question and all the possible answers available in the knowledge base. This way Recall@10 would always give 100%, making it difficult to judge how good the system would be from a user perspective in a practical implementation.

QA system evaluation is not necessarily relevant for TOIAs as the QA task is more about reading comprehension than the ability to retrieve an answer from a knowledge base and engage in a free-

form dialogue format. Moreover, such systems often use text generation models which we didn't use in our TOIA. Text generation methodologies are usually evaluated with n -gram based metrics (Merdivan et al., 2020) such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005), which are often criticized for their poor alignment with human judgement (Chen et al., 2019). Across all these works as well as the datasets presented for study free-form conversations, there is a gap in addressing the question of what is a 'good' answer. This is an important question to address not only for evaluating the relevant NLP tasks, but also for defining an annotation methodology.

2.3 Relevant Dialogue Data

Conversational questions have challenging phenomena not present in existing reading comprehension datasets. Recent datasets that focus on free-form human dialogues and include human annotations are CoQA (Reddy et al., 2019) and HUMOD (Merdivan et al., 2020). CoQA is a large scale reading comprehension dataset that improves a dataset like SQuAD (Rajpurkar et al., 2016) by including questions that depend on conversation history and by ensuring the naturalness of answers in a conversation. HUMOD instead takes inspiration from the Cornell's movie dialogue corpus (Danescu-Niculescu-Mizil and Lee, 2011) by adding human annotations to it. The Douban Conversation Corpus (Wu et al., 2016) contains dialogues between people sampled from Douban, a popular social network in China. The dataset is public and open domain — people chat about movies, books, music, etc. These datasets are both large scale and address different tasks, whereas TOIAs usually involve much smaller datasets. A system like Traum et al. (2015b)'s has a Knowledge Base (KB) of about 2,000 answers. We used the Margarita Dialogue Corpus (MDC) made available by Chierici et al. (2020), which has a KB of 431 answers, as well as a set of complete annotated dialogues.

2.4 Deep Retrieval-Based Dialogue Systems

State-of-the-art results have been achieved very recently on Answer-Retrieval tasks using deep learning architectures (Wu et al., 2016; Humeau et al., 2019; Roller et al., 2020). We used more straightforward techniques for this work as we want to focus on human evaluation rather than AR techniques. Moreover, the data size for the TOIA we

use — and for TOIAs in general — is too small for deep learning. We manage to overcome this limitation for a sentence similarity model (more on this in the next section) and plan to leverage transfer learning in future work.

3 The Margarita Dialogue Corpus

Chierici et al. (2020) recorded twenty dialogues with twenty different interrogators who were each instructed to engage in a 15-minute conversation with a TOIA's avatar maker. They then used ten randomly picked dialogues to define the training set (in the original data, these dialogues are labeled as 'train' but here we call them 'development' or 'dev' set as we use them as such). They used these dialogues as the inspiration for defining the KB of q-a pairs the avatar maker recorded in the TOIA. The MDC comprises conversations 'on-topic' and 'wild': half of the conversations are about the university attended by the avatar maker and half did not have a set topic — the interrogator was instructed to get to know the avatar maker as one would do when meeting a person for the first time. For the original dialogues and KB statistics, we point to the original MDC paper tables. Here we limit ourselves to mention a few highlights. The KB is not in dialogue format. There are 431 unique answers and 758 unique questions. The answers in the KB correspond to the videos the avatar maker recorded for powering the TOIA. Some questions have more than one possible answer, and some answers have more than one possible question. In total, the MDC KB comprises 892 self-contained q-a pairs. In addition to the KB, the MDC includes dev and test dialogues comprising 340 and 319 q-a pairs, respectively. Each dialogue has 33 turns on average.

Figure 2 shows the distribution of frequent trigram prefixes for the MDC's KB questions and answers, and for the dialogues dev set. Because of the free-form nature of questions, we have a richer variety of questions in the dialogues than the KB. While nearly half of the KB questions are dominated by 'what' questions, the dev questions are distributed across multiple question types. Several sectors indicated by prefixes *I*, *that*, *so*, and *it* are frequent in the dev set but are completely absent in the KB. This indicates that dialogues are highly conversational whereas the KB is not, and while a large portion of questions in the dev set are *do*, *I*, and *what* type of questions, an equally large number are made of different types of questions.

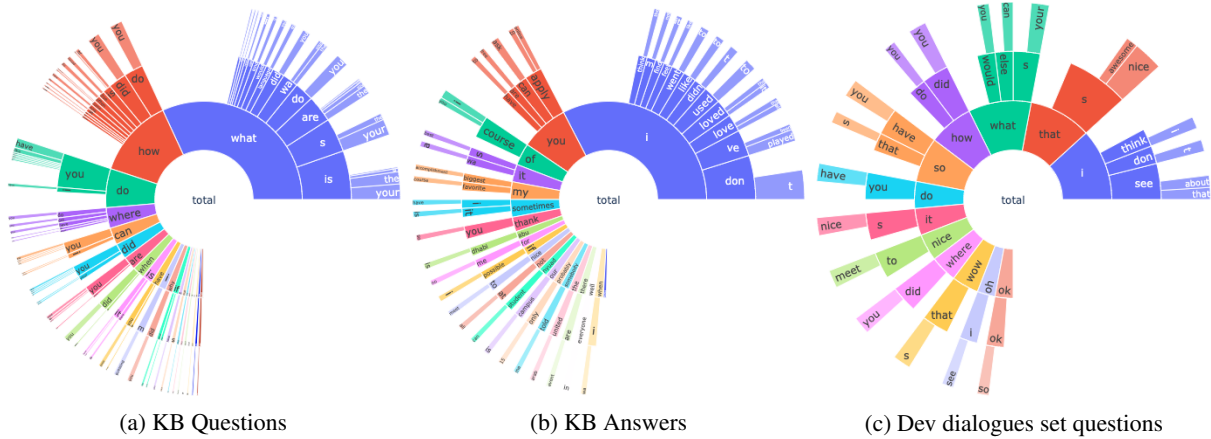


Figure 2: Distribution of trigram prefixes of questions and answers in the Margarita Dialogue Corpus’ KB (a, b), and questions in the MDC’s dialogue sets (c).

4 Retrieval Models

We used five models for retrieving answers for the questions in the MDC dialogue dataset, and for shortlisting the top candidate responses for the ‘crowd’ annotation task.

(1) TF-IDF q-Q: Let q be a query from a user (in our case, a question in the MDC dialogue dataset), and Q a question annotated in the MDC KB. We vectorized q and Q using a TF-IDF vectorizer trained on the KB, and computed the shortest distance between q and Q with cosine similarity. We used the sci-kit learn Python library for the TF-IDF vectorizer (Pedregosa et al., 2011).

(2) Okapi BM25 q-Q: Okapi BM25 (Trotman et al., 2014) is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document. We used the Rank-BM25 implementation in Python.⁴ Since BM25 was the worst performing approach, we do not report on it further due to limited space.

(3) BERT q-Q: BERT is a large deep learning model architecture, and one of 2018’s breakthroughs in NLP (Devlin et al., 2018). We computed the sentence embedding for each q and Q by taking the mean of BERT pre-trained layers. The cosine similarity between embeddings gives us the ranking function for computing how close a query in the dialogues is to a question in the KB.

(4 and 5) Fine-tuned BERT q-A: We fine-tuned BERT on answer selection as a classification task. Let A be an answer in the KB. For every Q - A pairs

in the KB, we labeled them as 1’s to indicate a relevant match. We then sampled a number of irrelevant (or ‘wrong’) matches for every question, and labeled them as 0’s. We tried different sampling ratios, namely drawing one wrong match for every correct one (1:1), ten wrong ones (1:10), a hundred (1:100) and using all the available utterances (1:All). To increase the data size further and better generalize for questions phrased differently, we augmented the train data by sampling synthetic questions using the methodology proposed by Wei and Zou (2019) and their Python implementation.⁵ We fine-tuned BERT for 3 more epochs (we chose a few epochs as advised by Dodge et al. (2020)) using Wolf et al. (2019)’s Transformers library. We only report on **BERT q-A 1:100** and **BERT q-A 1:All** as they were the best performing.

5 Crowd Annotations

We developed a web interface (Figure 3) for collecting the annotations from the ‘crowd’ using the crowdsourcing platform Amazon Mechanical Turk (AMT). Full anonymity of the users were maintained and the ERB review of the host institution didn’t raise ethical concerns.

For each question in the MDC dialogue dataset, we took the union of the top-10 answers retrieved by the five different retrieval techniques described above. On average, each question has about 24 selected answers. Using a sliding window on all the dialogue questions, we selected three conversation turns, and appended the prediction as a fourth turn (interrogator-avatar-interrogator-predicted avatar response) without specifying who was whom, and

⁴<https://pypi.org/project/rank-bm25/>

⁵https://github.com/jasonwei20/eda_nlp

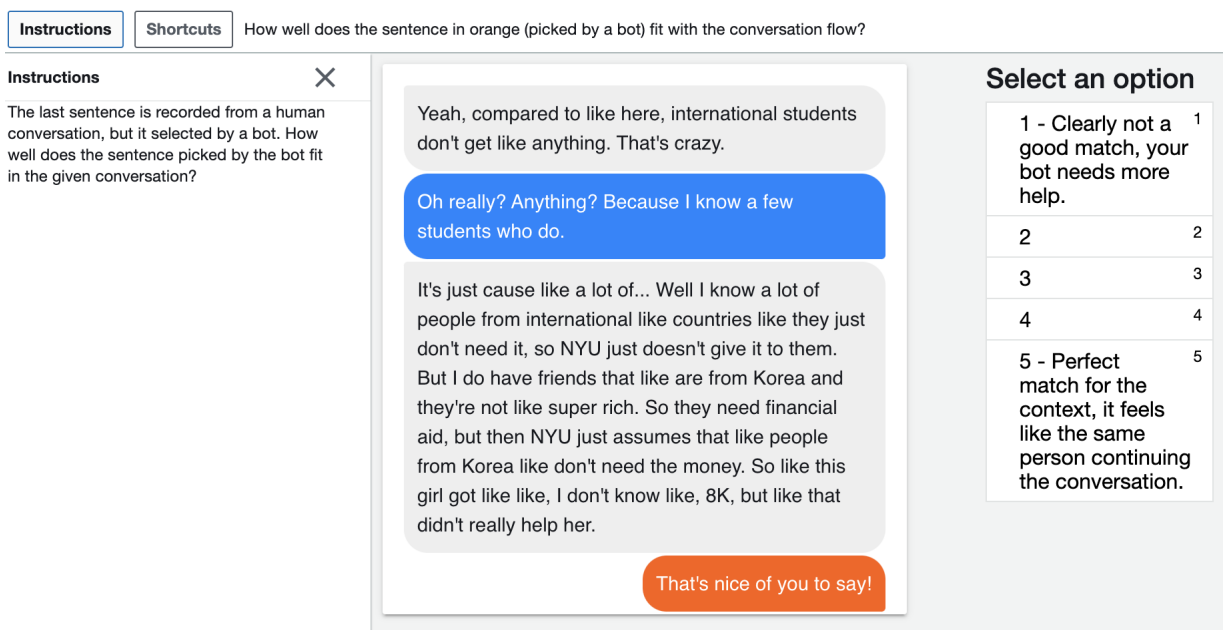


Figure 3: Annotation interface displayed to the human annotators.

always starting the dialogue snippet from an interrogator’s question and ending with the avatar’s answer. We chose four turns because it seems to give an optimal context size by looking at the annotations performed on the HUMOD dataset (Merdivan et al., 2020). So we have 339 dialogue snippets for the dev set and 341 for the test set. Each human annotator could rate as many snippets as they wanted in one task. On average, they rated 23 sampled dialogue conversations. They were asked to rate the last reply of the dialogue snippet on a 1–5 scale according to the dialogue context (where 1: Clearly not a good match; 5: Perfect match for the context). Each dialogue–reply pair is rated by three different annotators. For each dialogue context, there are on average 72 annotated answers (24 times 3), resulting in a total dataset size of 24,291 annotations for the dev dialogues and 24,555 for the test dialogues. In order to maintain high quality responses in the data, we defined a blacklist of annotators who gave poor quality annotations as follows. We forced each annotator to give a rating for the ‘gold answer’ given by the avatar maker in the dialogues data. If the annotator gave a rating lower than 4 (i.e., 1, 2 or 3) to the gold answer, we removed them from the annotations. While this blacklisting methodology is quite restrictive (we lose about 36% annotations), we have a large enough number of annotations left for our purposes.

Rater 1 vs Rater 2	κ (dev)	κ (test)
Closest two ratings	0.51	0.50
Lowest two ratings	0.23	0.20
Highest two ratings	0.07	0.11
Random two ratings	0.10	0.04

Table 1: Inter-annotator agreement computed using Cohen’s kappa score (κ) for the dev set and the test set.

6 Results and Analysis

We analyze our annotations in terms of interannotator agreement and the relationship between the crowd’s opinion and the best retrieved answers by the models. We then report the IR metrics on the models we decided to study.

6.1 Inter-annotator Agreement

We computed the weighted Cohen’s kappa score (Cohen, 1968) between human ratings to compute inter-annotator agreement excluding the blacklisted annotations. Following the approach of Merdivan et al. (2020), we calculated the weighted kappa score for different configurations of three ratings for each different context-predicted answer pair. We calculated weighted kappa score for the closest two (as a majority voting) ratings, the highest two ratings, the lowest two ratings, and on a random selection of two ratings from the three ratings of each predicted answer. For example, if a dialogue snippet is rated 1, 2 and 5, we keep the closest two

	Gold	TF-IDF	BERT	BERT	BERT
		q-Q	q-Q	q-A	q-A
				1:100	1:All
		<i>dev</i>			
Average	4.53	4.03	3.99	4.17	4.01
%Gold		89.0	88.2	92.1	88.5
Rank		2	4	1	3
		<i>test</i>			
Average	4.59	3.01	2.98	3.47	3.25
%Gold		65.6	64.9	75.7	70.9
Rank		3	4	1	2

Table 2: Average ratings assigned to the gold and top retrieval model choices in dev and test sets. %Gold specifies the ratio of model average rating to gold average rating. Rank specifies the performance rank of the retrieval model.

(1 and 2) and randomly assigned them to **Rater 1** and **Rater 2**. Table 1 shows the results of each combination for the MDC’s dialogue dev set and test set. The inter-annotator agreement seems consistent between the test and dev set apart for the highest two ratings and the random two ratings scenarios. However, both scenarios register quite a low agreement between annotators. Though representing moderate agreement, the highest is between the closest two ratings, showing that the ‘crowd’ seems more in agreement on a majority vote than on the lowest two ratings, highest two, or random two ratings. Moreover, agreeing on the worst answers (lowest two ratings) seems easier than agreeing on the best answers (highest two ratings). As our analysis points out, the poor level of agreement between annotators isn’t about the quality of the annotations. It rather shows the difficulty of defining what is a good answer for a TOIA’s interrogator.

6.2 Crowd Ratings of Retrieval Top Choices

Next we consider the average rating given by the AMT workers to the gold answer, and to the top retrieved reply by our four models. We include the ratings to all snippets excluding blacklisted annotations for both dev and test. We drop 35% of the annotations for the top retrieved answers due to blacklisting, consistently with the overall drop reported above. See Table 2 for the averages, percentage of the gold answer (i.e. how close to the gold answer is a model), and model ranking. The standard deviation of the average ratings for the gold answer is 0.35 in dev and 0.39 in test (because of blacklisting, we only keep ratings 4 and 5 for

Retrieval	ρ (dev)	ρ (test)
TF-IDF q-Q	0.25	0.10
BERT q-Q	0.16	0.08
BERT q-A 1:100	0.30	0.13
BERT q-A 1:All	0.29	0.15

Table 3: Spearman’s Rank Correlation Coefficient (ρ) between each retrieval model and the human ratings for the dev set and the test set.

the gold answer).⁶ The standard deviation of the retrieval models ranges from 1.17 to 1.33 in dev and 1.20 to 1.38 in test.

The results indicate that, although the crowd disagrees, they generally give high ratings to the best retrieved answers. So, annotators may disagree on many instances, but when the models retrieve sensible answers, these are recognized by the annotators. For this reason we decide not to resolve the annotator’s disagreements, and in the analysis that follows we use the average rating between the three (or less because of blacklisting) scores given by the crowd for each dialogue context-predicted reply pair. According to the crowd, the model with the best top choice is BERT q-A 1:100, and the model with the worst top choice is BERT q-Q.

6.3 Correlations Between Models and Annotations

We also computed the Spearman’s Rank Correlation Coefficients between the rankings produced by four of the models used for answer retrieval and the annotators ratings (always excluding the blacklisted annotations). The results are displayed in Table 3. While the correlations are weak (yet statistically significant as all the p-values approached 0), we can notice a mixed behavior. The models performing better (See Table 4) do not necessarily correlate more with human ratings. This is a ranking correlation. So the crowd may rank differently than the models’ answers but agree on the top ranked replies as we have seen earlier. Furthermore, on the 24 answer presented for each dialogue snippet on average, few ones are the top ranked by the models and the majority are ‘negative’ examples, where it’s easier to disagree or rank differently.

⁶For reference, the average of all ratings of the gold answers (i.e. without blacklisting) is 3.96 for dev and 3.76 for test, with corresponding standard deviation of 0.71 and 0.74 respectively.

Metric	Avatar Maker’s Annotations						Crowd’s Annotations					
	Rand	TF-IDF	BERT	BERT	BERT	Crowd	Rand	TF-IDF	BERT	BERT	BERT	Crowd
		q-Q	q-Q	q-A	q-A			q-Q	q-Q	q-A	q-A	
				1:100	1:All					1:100	1:All	
R@1	0.0	23.4	22.9	24.0	21.0	9.1	0.1	6.5	3.9	7.5	6.6	20.9
R@2	0.2	29.5	28.0	33.3	29.5	15.5	0.2	9.6	6.3	10.8	9.6	40.4
R@10	1.1	38.2	42.0	55.2	43.1	34.0	1.3	21.5	24.7	24.0	18.7	98.4
MRR	2.6	69.1	60.5	70.1	81.8	48.0	7.0	49.4	49.1	53.6	63.4	97.3
MAP	2.4	66.5	69.1	90.3	59.6	18.8	4.5	33.6	35.7	28.8	17.6	97.3
SR@1	0.0	32.4	31.8	33.2	29.1	12.6	0.3	36.6	36.0	42.3	36.9	97.3
SR@2	0.3	40.6	38.2	45.0	40.9	20.9	0.9	50.6	48.8	55.4	52.4	97.3
SR@10	1.5	49.1	51.5	66.8	57.1	43.5	5.6	75.3	80.7	82.4	79.5	97.3

Table 4: Information Retrieval metrics on the **dev** dialogues set for all the models, including a random selection model and using the crowd ratings as a retrieval model. On the left the models are assessed against the original annotations made by the avatar maker. On the right the models are assessed against the annotations from the crowd.

6.4 Versatile Questions and Answers

Excluding random noise or poor quality annotations, one hypothesis is that the more volatile (or the higher the disagreement in) the ratings for a given q-a pair, the more difficult it is to assign a ‘ground truth’ value to an annotation. To validate this hypothesis, we computed a more practical proxy of disagreement. The Coefficient of Variation (CoV) is defined as the standard deviation of the three ratings given on the same q-a pairs divided by their average. The CoV quantifies the variability of the ratings with respect to the average rating for a given q-a pair.

Let A be the set of questions with a CoV higher than the 75th percentile (0.50) and B the set of questions with a CoV lower than the median (0.25). A has 167 utterances, B has 239 and their intersection has 133. Set A less the intersection defines the ‘versatile’ questions, i.e. utterances that go well with many answers and generate high disagreement. Set B less the intersection represents ‘one-sided’ questions, i.e. questions that don’t go well with many answers, hence generate low disagreement. To confirm this expected behavior, we re-computed the Weighted Cohen’s kappa on the two versatile and one-sided questions. The uplift in agreement or disagreement confirmed our interpretation. E.g., for the one-sided questions, the inter-annotator agreement doubles on the highest two ratings, it improves by a few points for the closest two ratings and the lowest two ratings, while κ becomes negative for the one-sided questions. It seems that there are less versatile questions (34) than one-sided ones (106) in the dev dialogues.

Metric	Avatar Maker		Crowd			
	BERT	BERT	BERT	BERT		
	q-Q	q-A	1:100	q-Q	q-A	1:100
SR@1	7.8	11.9	13.5	25.2		
SR@2	12.2	19.7	23.3	35.5		
SR@10	22.6	36.1	75.8	68.2		

Table 5: SR@ k metrics on the **test** set only for the best performing models on SR@1 and SR@10.

Repeating the same analysis on the answers, we have 152 versatile, 237 one-sided, and 118 in the intersection.

6.5 IR Metrics

Given that a TOIA’s ability to engage in a conversation depends critically on retrieving a correct answer from the knowledge base, we focus on models (and hence metrics) used for Information Retrieval. The IR metrics we use to evaluate the answer retrieval ability of our TOIA are Recall@ k (Lowe et al., 2015), Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and we include the Success Rate (SR@ k) inspired by the work of Sakata et al. (2019). Success Rate is the fraction of questions for which at least one correct answer is ranked among the top k .

Table 4 summarizes the retrieval metrics for four of the models we worked with. We added two benchmarks: a model that randomly picks any answer from the KB, and we used the average crowd ratings’ as a ‘retrieval’ model. When we compare the avatar maker’s annotations vs. the crowd’s annotations, we label an answer as ‘correct’ for the

Question (<i>redacted</i>) I think it's somewhere in California. Basically people are no longer afraid of being caught with meth. So like the whole city is like a capital of meth. So what do you feel about that? Does that change what you previously said?	BERT Kind of.
Avatar Maker No.	
Question So would you feel the same way about like meth, heroin, coke because technically those...?	BERT Sure.
Avatar Maker (Can you ask me something else?, I'd rather talk about something else, sorry.)	
Question (<i>redacted</i>) How about like your perception of like religion, like are you religious or?	BERT I started questioning my beliefs when I was a teenager after watching lots of documentaries and talking to agnostics alike.
Avatar Maker I am sure something bigger than us exists, but I don't follow any specific book or rituals.	
Question (<i>redacted</i>) How is your romantic life.	BERT Relationships are harder to maintain because of all the study-aways. Also, the campus is a bit isolated from the city so it's not as lively.
Avatar Maker I like to keep things like this private.	
Question Yeah, that's fascinating. So no toilet in your granny house?	BERT Sure.
Avatar Maker No.	
Question Bye!	BERT See you later!
Avatar Maker Goodbye!	

Figure 4: Examples in the dev set where the BERT q-A 1:100 system choice differs from the Avatar Maker's.

crowd, if it received an average rating ≥ 3.5 . The metrics consistently show that the best retrieval model looks like the fine-tuned BERT on the q-A relevance task with a training set ratio of 1:100 between correct and wrong Q-A pairs. We can now notice why $\text{Recall}@k$, MRR and MAP may not be adequate for the context of TOIAs (and for the wider task of q-a retrieval from a KB).

One would expect that AMT raters would be more generous in classifying answers as 'relevant' for a given question. In fact, there are often cases where a sensible answer gets retrieved by a model (Figure 4), but the avatar maker did not deem it as appropriate. Other utterances like yes/no, sure, and OK answers are relevant for many questions, but, as expected, the avatar maker would be more selective to choose which one between a yes or a no is an appropriate answer. However, the $\text{Recall}@k$, MAP and MRR look lower in the case where the models are assessed against the crowd annotated data. This is partly due to the models trained on the data annotated by the avatar maker, but mostly because the crowd is indeed more generous and the examples of relevant q-a pairs increased vs. the avatar maker's annotations. Moreover, MRR is highly influenced by the number of documents retrieved by a model. In fact, the trivial model retrieving all possible answers in the KB would give a 100% MRR. MRR is the only metrics for which it seems that the BERT q-A model with the 1:All sampling ratio performs better than the 1:100 ratio but, in reality, this is due to the model with the 1:All ratio retrieving more documents.

Including the $\text{SR}@k$ metrics makes things easier to assess. $\text{SR}@k$'s for models evaluated on the

crowd's annotations are consistently higher than the respective models assessed on the avatar maker's annotations with the only exception of the best model. BERT q-A (1:100) gives $\text{SR}@k$'s that are higher than $\text{R}@k$'s when evaluated on the data annotated by the avatar maker. The difference is even steeper on the data annotated by the crowd.

We evaluated the retrieval models versus the data with combined annotations, i.e. both by the AMT workers and the avatar maker. The results have negligible differences with respect to the assessment against the data annotated by the crowd, suggesting that the avatar-maker annotations are mostly included in the crowd annotations.

7 Discussion

7.1 Accurate vs Plausible Answers

Models that produce state-of-the-art results in other domains seem to not perform as well in the context of a TOIA. While a model like BERT q-A 1:100 retrieves *plausible* answers, the avatar maker's *accurate* answers differ (Figure 4, Table 2). This is also shown by the weak correlation between the IR models and the human annotators, and by the poor results the 'crowd model' generates on the data annotated by the avatar maker (Figure 4). For some answers, it seems that the avatar maker missed them when annotating the dialogues set due to human error. In fact, she had to manually go through 431 answers for 659 questions for a total of 284,029 look-ups. On the other hand, many questions require affirmative or negative answers, which both makes sense when evaluating a dialogue snippet but only one type of answer is correct for the avatar.

7.2 Viable Metrics for TOIA Evaluation

A TOIA is a single-output system, where the best answer should be retrieved as the top ranked document, and there may be more than one answer that suits perfectly within a conversation turn. This makes traditional IR metrics unsuitable for optimizing time-offset systems, so we identified SR@1 as the metric that gives the best indication for the ability of the system to retrieve a ‘good’ answer. For $k > 1$, SR@ k gives us more insight into how to improve a model. For instance, the best performing model can retrieve a good answer in the top 10 ranked retrieved utterances in more than 80% of the cases. This information can be used to improve the system, e.g. by retrieving the top 10 answers using BERT q-A 1:100 and fine-tuning a re-ranking methodology that pushes on top the best answers. Table 5 shows the SR@ k metrics for the test set, and it’s interesting to notice that BERT q-Q yields a better SR@10 on the crowd’s annotations than the BERT q-A 1:100 model.

7.3 Limitations

We limited the study to a retrieval problem and we did not leverage the conversational format of the dialogues set. There are some turns where we can observe co-reference (a few examples can be seen in Figures 1 and 4). We manually annotated co-references in a sample of 100 dialogue turns and these happen in about 5% of the dialogues. So while the IR techniques produce some errors due to the conversational structure, this is not as material as to invalidate this study. Regarding the annotation methodology, a ‘fairer’ way to annotate the answers might have been to ask the AMT workers to give a rating for every question in the KB paired with every dialogue snippets in the dialogue sets. So when we use the crowd ratings as a model (Table 4), we are limited to the the answers that were rated by the human annotators. Rating all answers for every single question would be unpractical and picking the union of the top 10 retrieved answers from our models makes sure that the human raters could see an answer annotated by the avatar-maker for at least 66.8% of the questions (SR@10 of the best performing model, Table 4).

8 Conclusion and Future Work

We explored the challenge of defining what a ‘good’ answer is in the context of a TOIA by annotating a dataset used for creating an avatar, and evaluat-

ing human-avatar dialogues. We learned that the perceived ‘right’ answer for avatar interrogators differs from the avatar maker expectations partly because some questions and answers are too versatile, i.e., they go well with many answers and questions, respectively. Additionally, yes and no answers are equally perceived as relevant by users interrogating an avatar but would be right or wrong for a given avatar maker. We make all the human annotations we collected available to the research community. We challenged classical retrieval metrics and proposed that TOIA’s dialogue managers should optimize Success Rate @1. Success Rate @ k for different levels of k can help identify how to improve retrieval techniques.

Our future work includes recognizing versatile questions and answers, designing methods to elicit more precise answer recordings at the avatar creation stage, and forcing yes/no answers with acceptable degrees of confidence. We plan to use transfer-learning and one-shot learning for leveraging state-of-the-art results of deep neural models in the context of a TOIA. Addressing misaligned expectations between different user needs and picking the right metric are essential to improving the design, usability, and answer retrieval methodology of time-offset interaction applications.

Acknowledgments

The first author was supported by the New York University Abu Dhabi Global PhD Student Fellowship program. The support and resources from the High Performance Computing Center at New York University Abu Dhabi are gratefully acknowledged. We thank NYUAD’s CAMEL Lab member Bashar Alhafni for helpful discussions.

References

- Dana Abu Ali, Muaz Ahmad, Hayat Al Hassan, Paula Dozsa, Ming Hu, Jose Varias, and Nizar Habash. 2018. A bilingual interactive human avatar dialogue system. In *Proceedings of the 19th Annual SIG-dial Meeting on Discourse and Dialogue*, pages 241–244.
- Ron Artstein, Anton Leuski, Heather Maio, Tomer Mor-Barak, Carla Gordon, and David Traum. 2015. How many utterances are needed to support time-offset interaction? In *The Twenty-Eighth International Flairs Conference*.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved

- correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Basma El Amel Boussaha, Nicolas Hernandez, Christine Jacquin, and Emmanuel Morin. 2019. Deep retrieval-based dialogue systems: a short review. *arXiv preprint arXiv:1907.12878*.
- Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124.
- Alberto Chierici, Nizar Habash, and Margarita Bicec. 2020. The margarita dialogue corpus: A data set for time-offset interactions and unstructured dialogue systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 476–484.
- Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.
- Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.
- Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.
- Karen Spärck Jones. 2005. **ACL lifetime achievement award: Some points in a time**. *Computational Linguistics*, 31(1):1–14.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. **How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation**. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.
- Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.
- Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to information retrieval.
- Erinc Merdivan, Deepika Singh, Sten Hanke, Johannes Kropf, Andreas Holzinger, and Matthieu Geist. 2020. Human annotated dialogues dataset for natural conversational agents. *Applied Sciences*, 10(3):762.
- Masashi Nishiyama, Tsubasa Miyauchi, Hiroki Yoshimura, and Yoshio Iwai. 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.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.
- Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.
- Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. 2019. Faq retrieval using query-question similarity and bert-based query-answer relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1113–1116.
- David Traum, Kallirroi Georgila, Ron Artstein, and Anton Leuski. 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.

- David Traum, Andrew Jones, Kia Hays, Heather Maio, Oleg Alexander, Ron Artstein, Paul Debevec, Alesia Gainer, Kallirroi Georgila, Kathleen Haase, 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.
- Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, pages 58–65.
- Jason Wei and Kai Zou. 2019. [EDA: Easy data augmentation techniques for boosting performance on text classification tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6383–6389, Hong Kong, China. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2016. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*.