Information Seeking in the Spirit of Learning: A Dataset for Conversational Curiosity

Pedro Rodriguez[*]
Computer Science
University of Maryland
pedro@cs.umd.edu

Paul Crook
Facebook
pacrook@fb.com

Seungwhan Moon
Facebook
shanemoon@fb.com

Zhiguang Wang
Facebook
zgwang@fb.com

Abstract

Open-ended human learning and information-seeking are increasingly mediated by digital assistants. However, such systems often ignore the user's pre-existing knowledge. Assuming a correlation between engagement and user responses such as "liking" messages or asking followup questions, we design a Wizard-of-Oz dialog task that tests the hypothesis that engagement increases when users are presented with facts related to what they know. Through crowd-sourcing of this experiment, we collect and release 14K dialogs (181K utterances) where users and assistants converse about geographic topics like geopolitical entities and locations. This dataset is annotated with pre-existing user knowledge, message-level dialog acts, grounding to Wikipedia, and user reactions to messages. Responses using a user's prior knowledge increase engagement. We incorporate this knowledge into a multi-task model that reproduces human assistant policies and improves over a BERT content model by 13 mean reciprocal rank points.

1 Introduction

Conversational agents such as Alexa, Siri, and Google Assistant should help users discover, learn, and retain novel factual information. More generally, systems for conversational information-seeking should help users develop their information need, be mixed-initiative, incorporate user memory, and reason about the utility of retrieved information as a combined set (Radlinski and Craswell, 2017). We focus on a curiosity-driven, information-seeking scenario where a user starts a conversation with an assistant by asking an open-ended question and then drills down into interest areas (Figure 1).

In this setting, what policies should assistants pursue to maintain the user's interest in the topic?

U: <assistant wake-word>, tell me about Tahiti.
A: It's the largest island in French Polynesia, near the center of the Pacific
U: What is its history with France?

Figure 1: An example of information-seeking dialog that the Curiosity dataset aims to support. Assistants should answer user questions *and* convey information that inspires meaningful followup questions.

Theories of human learning, such as Vygotsky's zone of proximal development, propose that learning novel information should be rooted in pre-existing knowledge and skills of the learner (Chaiklin, 2003). Considering this, a good policy may give general information about Tahiti; a better policy would select information related to the user's knowledge (e.g., familiarity with France). We hypothesize that engagement is correlated with policies that integrate a user's pre-existing knowledge, and test this through a large-scale, Wizard-of-Oz (WoZ) style collection (Kelley, 1984; Wen et al., 2017) that captures assistant policies, user reactions, and topically relevant entities that the user knows about. The Curiosity dataset has 14,048 English dialogs annotated with sentence-level knowledge grounding, the user's prior knowledge, dialog acts per message, and binary ratings per message.[1]

In our dialog task (Figure 2), one worker takes the role of a curious user learning about a geographic entity and the other of a digital assistant with access to Wikipedia facts (Section 2). At the start of each dialog, the user is assigned an entity as their topic (e.g., Puerto Rico) along with two *aspects* (e.g., *history* and *demographics*) to investigate. Beforehand, we show the user a list of entities related to the topic, and they mark which they know; these entities are a sample of their pre-

[*]Work done while interning at Facebook.

[1]Dataset and code at curiosity.pedro.ai.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 8153–8172,
November 16–20, 2020. ©2020 Association for Computational Linguistics

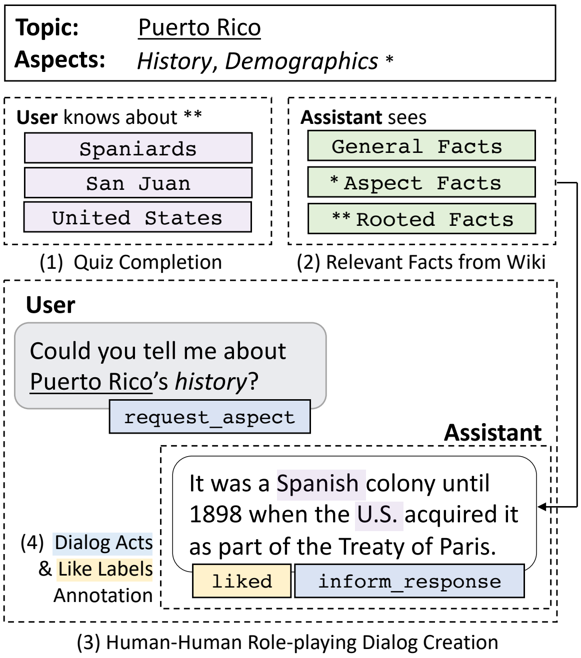

Figure 2: We sample pre-existing knowledge by asking users to indicate which topically related entities they already *know*. The assistant paraphrases facts related to either known entities (rooted facts), an aspect (aspect facts), or the topic generally (general facts). The user expresses engagement through a like button. Dialog acts are annotated in a separate crowd-source task.

existing knowledge. The user engages in open-ended discovery while the assistant simultaneously answers the user's questions and proactively introducing facts likely to prompt followup questions.

Section 3 uses dialog act annotations combined with explicit and implicit user feedback to compare assistants' content selection and presentation policies. For example, in interactions where the user asks a question and the assistant paraphrases a fact, how often does the user ask a followup question versus trail off in disinterest? Most datasets (Section 6) do not have enough annotations to answer these questions: it requires message-level dialog act annotations and feedback signals. We compare three assistant policies: using a fact with a rooted entity, a fact from the user's aspect, or a generic fact about the topic. The policies are compared through user "likes" of assistant messages and by the dialog act of their subsequent message (e.g., did they ask a specific followup or change topic).

In Section 4, we design models that predict the policies used by the assistant: what type of message to send and which fact to use (if any). All models are trained jointly with a multi-task objective function. We compare an end-to-end BERT (Devlin et al., 2018) model to our task-specific Hierarchical Recurrent Encoder model (Serban et al., 2015) and show that our model improves over the baseline.

In summary, we make three main contributions: (1) we design an experiment to test the efficacy of personalizing conversational information systems through a user's prior knowledge, (2) introduce the Curiosity dataset—the first dialog dataset combining sentence-level knowledge groundings, per message ratings, *and* per message dialog act annotations, allowing for robust and fine-grained structural learning of dialog policies for similar applications, and (3) design a multi-task model that incorporates the user's prior knowledge and improves over a natural BERT baseline.

2 Building the Curiosity Dataset

This section describes the construction of the Curiosity dataset. Dialog topics consist of prominent world geographic entities. The *worldwide* spread of entities makes each novel to most users, the consistent topic type makes starting dialogs easier, and their rich histories, demographics, and economics add topical diversity. For example, most people are only vaguely familiar with the history of Puerto Rico, but most know about related concepts such as the United States or Hurricane Maria. Section 2.1 describes how we select geographic topics, aspects, and derive a set of facts to ground against. We collected the dataset in two steps: (1) collecting dialogs with a custom interface (Section 2.2) and (2) after-the-fact dialog act annotation (Section 2.3). Sample dialogs from Curiosity are in Appendix C.

2.1 Geographic Topics, Aspects, and Facts

We select 361 geographic pages from Wikipedia that have separate geography and history pages (e.g., Puerto Rico, Geography of Puerto Rico, and History of Puerto Rico).[2] We use sentences from each page to build a set of 93,845 facts. We run an entity linker over the content (Gupta et al., 2017) and index each fact by its source page (topic), source section (*aspect*), and mentioned entities. Finally, we fit a TF-IDF text matcher (Rajaraman and Ullman, 2011) with Scikit-Learn (Pedregosa et al., 2011). While conversing, assistants are shown facts filtered by topic, aspect, or mentioned entities, that are ranked by textual similarity to the dialog.

[2]The existence of these pages implies that the topic has ample historical and geographical knowledge to draw from.

2.2 User and Assistant Dialog Interfaces

To collect dialogs, we build user and assistant interfaces for annotators. The user's interface samples their prior knowledge of a topic, captures which assistant messages interest them, and manages the dialog context. The assistant's interface provides contextually relevant facts. Appendix A has screenshots and details of each interface component.

Sampling User's Prior Knowledge When deployed, digital assistants can draw from prior interactions (Ram et al., 2018) to estimate what a user knows. However, since we do not have these prior interactions, we collect information about what users know. Instead of exhaustively asking about every entity related to the topic, we sample this knowledge. Before the dialog begins, we show the user fifteen related entities that range from commonplace to obscure (United States versus Taíno). Users mark the entities they could (1) locate on a map or (2) summarize succinctly in one sentence.

Like Button for User Interest As part of our collection, we aimed to determine what fact-grounded utterances users found interesting. Users "liked" the assistant's message if they found it "interesting, informative, and relevant to their topic."

Assistant's Topic Summary and Fact Bank The worldwide spread of Curiosity's entities makes them unfamiliar to most crowd-workers, including the assistants. So that the assistant can still engage the user, the assistant interface provides contextually relevant information. First, the interface shows a topic summary from Wikipedia. Second, the assistant paraphrases facts from a contextually updated fact bank (box 2 in Figure 2). To reduce information overload, we use simplified topic descriptions from SimpleWikipedia and show a maximum of nine facts at a time.[3] We encourage assistants to "stimulate user interest and relate information to things they already know or have expressed interest in." Assistants are instructed to select relevant facts, click the "use" button, and paraphrase the content into their next utterance.

Like Dinan et al. (2019), the fact bank shows facts to the assistant using TF-IDF textual similarity to recent dialog turns but differs by incorporating the user's prior knowledge. We show the assistant nine facts: three facts that mention an entity familiar to the user (rooted facts), three facts from their assigned aspects (aspect facts), and three from anywhere on the page (general facts). By construction, rooted facts overlap with the exclusive categories of aspect and general facts. For each category, we find the nine highest TF-IDF scoring facts and then randomize their order. To avoid biasing the assistant, we do not inform them about the user's known entities or distinguish between types of facts.

[3]If a description exists in simple.wikipedia.org, we use that; otherwise, we use the description from en.wikipedia.org.

2.3 Dialog Act Annotation

Inducing structure on conversations through dialog acts is helpful for analysis and downstream models (Tanaka et al., 2019). We introduce structure—beyond knowledge groundings—into Curiosity by annotating dialog acts for each message.

In a separate collection, we annotate all utterances with dialogs acts using a custom interface (Appendix B). The annotation schema is based on ISO 24617-2 (Bunt et al., 2010, 2012) with customized sub-categories for our scenario. Table 1 shows our schema, descriptions, and examples.

2.4 Data Quality

We crowd-sourced conversations in two phases using ParlAI (Miller et al., 2017). In the first, pilot studies collect feedback from individual workers. Based on feedback, we create task guidelines, sample dialogs, a FAQ, tutorial videos, and qualification tests. These materials were used to train and qualify crowd-workers for the second phase. During the second, we monitor the interface usage and removed workers that ignored instructions.

Using Krippendorff's α (Krippendorff, 2004), we validate the quality of dialog act annotations. Dialog acts are multi-class and multi-label: a message can have none, one, or multiple dialog acts (e.g., positive feedback and followup). However, Krippendorff's α is computed for single-label tasks from a table where rows represent examples, columns represent annotators, and cells indicate the singular class label. We convert our multi-label problem to a single label problem by making each combination of example and label class a row in the table (Table 2). Since there are few dialog acts per utterance, most annotations agree; however, since Krippendorff's α focuses on disagreement, it is appropriate for this scenario. Using a separate annotation interface (Appendix B), we doubly annotate 4,408 dialogs and the agreement score 0.834 is higher than the 0.8 threshold recommended by Krippendorff (2004). Next, we analyze the annotated dialogs and introduce our model.

Dialog Act	Count	Description	Example
request topic	10,789	A request primarily about the topic.	I'd like to know about Puerto Rico.
request aspect	41,701	A request primarily about an aspect.	Could you tell me about its *history*?
request followup	4,463	A request about mentioned concept.	Do you know more about the Taínos?
request other	10,077	Requests on unmentioned concepts.	What is there to know about cuisine?
inform response	59,269	Directly answer an info request.	Taínos were caribbean indigenous.
inform related	6,981	Not a direct answer, but related info.	I do not know, but...
inform unrelated	557	Does not answer question, not related.	Politics is tiring!
feedback positive	26,946	Provide positive feedback	Thats quite interesting!
feedback negative	176	Provide negative feedback	Thats pretty boring.
feedback ask	36	Ask for feedback	Do you find < info > interesting?
offer topic	91	Offer to discuss topic	Want to learn about Puerto Rico?
offer aspect	1,440	Offer to discuss aspect	How about more on its *demographics*?
offer followup	63	Offer to discuss mentioned concept.	I could say more about the Spanish.
offer other	1,619	Offer to discuss unmentioned concept.	How about I tell you about its exports.
offer accept	1,727	Accept offer of information.	I'd love to learn about its history.
offer decline	405	Decline offer of information	Sorry, I'm not interested in that.

Table 1: Counts, abbreviated descriptions and examples of the dataset's dialog acts.

	Annotator 1	Annotator 2
Utterance 1, Label A	Yes	No
Utterance 1, Label B	Yes	No
Utterance 2, Label A	Yes	Yes
Utterance 2, Label B	Yes	Yes

Table 2: Consider a task where each utterance has labels A and B. In the single-label version, each utterance is labeled as either A or B. The table shows the outcome of converting the multi-label version to single-label by creating a row for each example–label combination. Cell values are binary indicators.

Metric (# of)	Total	Train	Val	Test	Zero
Dialogues	14,048	10,287	1,287	1,287	1,187
Utterances	181,068	131,394	17,186	17,187	15,301
Likes	57,607	41,015	5,928	5,846	4,818
Topics	361	331	318	316	30
Facts Total	93,845	NA	NA	NA	NA
Facts Shown	76,120	66,913	29,785	30,162	6,043
Facts Used	27,486	21,669	4,950	4,952	2,290

Table 3: Curiosity has 14,048 dialogs. On average, dialogs have 12.9 utterances. 60% of the assistants' 90,534 utterances were liked.

3 Dataset Analysis

This section shows statistics of the Curiosity dataset and that users prefer aspect-specific, rooted facts.

3.1 Dataset Statistics

Table 3 shows the basic statistics of the Curiosity dataset. In total, our dataset contains 14,048 dialogs with 181,068 utterances. The fact database contains 93,845 facts; of those, 76,120 (81%) were shown to the assistants and 27,486 (29%) were used in at least one message. We randomly split dialogs into training, validation, and testing folds.

3.2 What Facts do User Prefer?

In Section 1, we hypothesized that when assistants use facts that mention previously known entities (rooted facts), users will be more likely to engage. In our data collection, we incorporate two mechanisms to test this hypothesis. The first mechanism is explicit: we directly ask users—through a like button—to indicate what messages they preferred. The second mechanism is implicit and derived by mining dialogs for specific sequences of dialog acts that suggest engagement with the content. For each of these mechanisms, we compute the likelihood $P(\text{Prefer} \mid \text{Fact Source})$ of a user preferring utterances grounded to each fact source (Rooted, Aspect, or General). Figure 3 shows this likelihood and indicates that users prefer: (1) facts relevant to aspects versus general ones and (2) rooted facts in three of four scenarios.

3.2.1 Likes for Explicit Preference Elicitation

Explicit preference is computed directly from like button usage and shown on the right panel of Figure 3. Overall, users liked 60% of messages, and they prefer on-aspect, rooted facts.

3.2.2 Mining Acts for Implicit Preferences

When users ask specific followup questions—as opposed to generic ones—about an assistant's fact, it shows that the user implicitly prefers these kinds of messages. For example, asking about an entity like the Taínos is more specific than asking about history and therefore indicates engagement. We

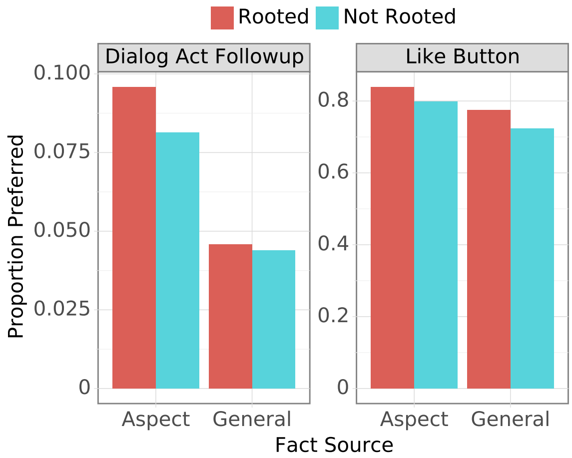

Figure 3: User engagement is measured by dialog act followups (left) and like button usage (right). We compare reactions to messages that use a fact mentioning an entity the user knew about (rooted) and whether the fact is general or aspect-specific. Pairwise differences are statistically significant (99%+) with a two proportion z-test *except* for dialog act followups between rooted and non-rooted general facts. Overall, users prefer on-aspect, rooted facts.

identify these interactions by mining for pairs of assistant-user messages where the assistant uses a fact and their message is labeled with an "inform" dialog act. With these, we compute the likelihood

$$P(\text{Outcome} = \text{request followup} \mid \text{Fact Source})$$

that the user's message has the "request followup" dialog act given the source. Similarly to likes, users engage more with aspect-oriented and rooted facts.

3.2.3 Paraphrase Analysis

Although our work does not include a paraphrase model, we manually analyze a random sample of two hundred and fifty assistant messages where facts were used. Of these messages, 51% were acceptable paraphrases, 27% were verbatim copies, 12% were contextualizations of near copies, and the remainder were errors such as incorrect paraphrases or did not incorporate the fact. Appendix D shows descriptions, counts, and random examples of each category. This analysis estimates that about half of grounded messages have non-trivial signal for future paraphrase models to use.

4 Models

We design a machine learning model that predicts assistant and user actions. We introduce a multi-task architecture for **C**uriosity that **H**ier**ar**chically **M**odels (CHARM, Figure 4) dialogs to: (1) predict the dialog acts of the user message (utterance act prediction), (2) select the best fact (fact prediction), (3) choose the best set of dialog acts for the next message (policy act prediction), and (4) predict if the assistant message will be liked (like prediction).

4.1 Text Representation

CHARM jointly encodes the text of utterances and facts with one encoder. E is a bi-directional LSTM (Sutskever et al., 2014) over GLoVE (Pennington et al., 2014) word embeddings and Wikipedia2Vec (Yamada et al., 2020) entity embeddings.[4] The text t_i^u of utterance u_i in dialog D is represented as $E(t_i^u)$. Similarly, fact f_j on turn i is represented as $E(t_{i,j}^f)$ where j indexes facts shown on that turn.

4.2 Dialog Representation

In our models, we use a hierarchical recurrent encoder (HRE) architecture (Sordoni et al., 2015; Serban et al., 2015) where a forward LSTM contextualizes each utterance to the full dialog. We modify the HRE model by adding additional inputs beyond the utterance's textual representation. First, we represent user's known entities

$$\boldsymbol{k} = \text{avg}(E_{\text{entity}}(e_1), \ldots, E_{\text{entity}}(e_k)) \quad (1)$$

as the average of entity embeddings. An entity embedding also represents the topic

$$\boldsymbol{t} = E_{\text{entity}}(\text{topic}) \quad (2)$$

of the dialog. Next, we create trained speaker embedding $\boldsymbol{v}_s$ for the user and $\boldsymbol{v}_t$ for the assistant. Given the set of all dialog acts $\mathcal{A}$, each utterance has a set of dialog acts $\mathcal{A}_u \in \mathcal{P}(\mathcal{A})$ where $\mathcal{P}(\mathcal{X})$ denotes the set of all subsets of $\mathcal{X}$. Finally, we use an act embedder A to compute an act representation

$$\boldsymbol{a}^i = \frac{1}{|\mathcal{A}_u|} \sum_{a_k \in \mathcal{A}_u} A(a_k) \quad (3)$$

by averaging embeddings at each turn. The input at each step is the concatenation

$$\boldsymbol{c}^i = [E(t_i^u); \boldsymbol{a}^i; \boldsymbol{t}; \boldsymbol{k}; \boldsymbol{v}] \quad (4)$$

of the representations for text, speaker, topic, known entities, and utterance dialog acts.[5] With this joint representation, the contextualized dialog representation

$$\boldsymbol{h}^{i-1} = \text{LSTM}(\boldsymbol{c}^1, \ldots, \boldsymbol{c}^{i-1}) \quad (5)$$

[4] In CHARM, BERT was not as effective an encoder.

[5] The speaker embedding $\boldsymbol{v}$ alternates between $\boldsymbol{v}_s$ and $\boldsymbol{v}_t$.

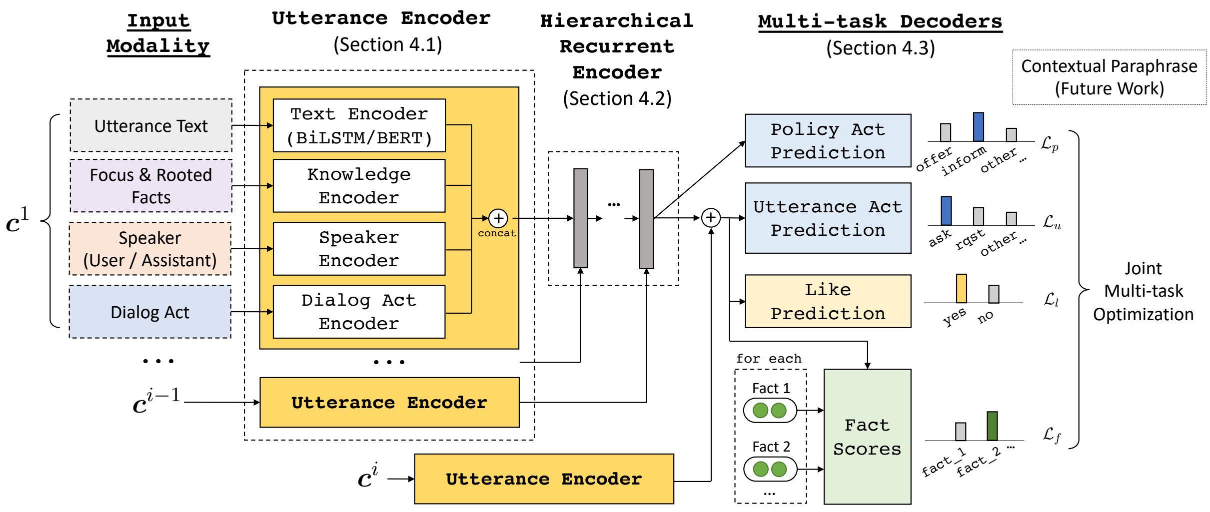

Figure 4: **Architecture**: CHARM builds a dialog context up to $t = i - 1$ to predict the current message's dialog acts (policy prediction) and the best facts to use. The model uses this combined with the current utterance to classify it's dialog acts and if it will be liked.

is the final LSTM state and includes time step $t = i - 1$. The dialog up to and including time i is

$$\boldsymbol{d}^i = [\boldsymbol{h}^{i-1}; \boldsymbol{c}^i] \tag{6}$$

which emphasizes the current utterance and makes multi-task training straightforward to implement.

4.3 Tasks and Loss Functions

In our model, we jointly learn to predict fact usage, user likes, utterance acts, and policy acts.

Fact Prediction For every assistant turn, the model predicts which fact(s) from

$$\{f_1, \ldots, f_k\} \in \mathcal{F}^{(i)}, \mathcal{F}^{(i)} \in \mathcal{P}(\mathcal{F})$$

the assistant marked as "used" where $\mathcal{F}$ is the set of all facts. We frame this task as pointwise learning to rank (Li et al., 2008). A fact prediction network

$$\boldsymbol{s}_j^{f,(i)} = \text{GELU}\left(\left[\boldsymbol{W}^f \cdot \boldsymbol{h}^{(i-1)} + \boldsymbol{b}^f; E(t_j^f)\right]\right) \tag{7}$$

with parameters $\boldsymbol{W}^f$ and $\boldsymbol{b}^f$ using a Gaussian Error Linear Unit (Hendrycks and Gimpel, 2017) outputs salience scores for each fact. The network does not use utterance u_i since it contains signal from the choice of fact. The predictions

$$\hat{\boldsymbol{y}}_j^{f,(i)} = \text{softmax}(\boldsymbol{s}_j^{f,(i)}) \tag{8}$$

are converted to probabilities by the softmax

$$\text{softmax}(\boldsymbol{q}) = \frac{exp(\boldsymbol{q})}{\sum_{j=1}^{k} exp(\boldsymbol{q}_j)} \tag{9}$$

over k labels. Using this, we compute the fact loss

$$\mathcal{L}_f = \frac{1}{|\mathcal{F}^{(i)}|} \sum_{i,j} \ell_{ce}(\hat{\boldsymbol{y}}_{i,j}^f, \boldsymbol{y}_{i,j}) \tag{10}$$

where labels $\boldsymbol{y}_j^{f,(i)}$ indicate if fact from utterance i in position j was used and

$$\ell_{ce}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \sum_{p=1}^{k} \boldsymbol{y}_p \log(\hat{\boldsymbol{y}}_p). \tag{11}$$

is the cross entropy loss. To mitigate class imbalance, we also scale positive classes by nine (Japkowicz and Stephen, 2002).

Policy Act and Utterance Act Prediction Each utterance may have multiple dialog acts so we treat policy and utterance act prediction as a multi-label task. The goal of policy prediction is to choose the best act for the next utterance; the utterance act classifies the last message's acts. To predict these acts, we create a policy act network

$$\boldsymbol{s}^{p,(i)} = \text{GELU}(\boldsymbol{W}^p \cdot \boldsymbol{h}^{i-1} + \boldsymbol{b}^p) \tag{12}$$

and an utterance act network

$$\boldsymbol{s}^{u,(i)} = \text{GELU}(\boldsymbol{W}^u \cdot \boldsymbol{d}^i + \boldsymbol{b}^u) \tag{13}$$

where the probability of act a_k is $p_k^{*,i} = exp(\boldsymbol{s}_k^{*,(i)})$. From these, we derive the policy act loss

$$\mathcal{L}_p = \sum_{k}^{|\mathcal{A}|} y_{i,k}^a \log p_k^{p,i} + (1 - y_{i,k}^a) \log(1 - p_k^{p,i}) \tag{14}$$

and utterance act loss

$$\mathcal{L}_u = \sum_{k}^{|\mathcal{A}|} y_{i,k}^a \log p_k^{u,i} + (1 - y_{i,k}^a) \log(1 - p_k^{u,i}) \tag{15}$$

for an utterance at $t = i$ with act labels $y_{i,k}^a$.

	Fact Rank (MRR)		Utt. Act (F_1)		Policy Act (F_1)		Like (Accuracy)	
Model	Val	Test	Val	Test	Val	Test	Val	Test
Majority Class	N/A	N/A	0.602	0.604	0.491	0.494	0.690	0.681
E2E BERT	0.420	0.418	0.794	0.795	0.635	0.631	0.829	**0.822**
CHARM	0.546	**0.546**	0.845	**0.847**	0.682	**0.682**	0.826	0.815
− context	0.516	0.506	0.838	0.842	0.664	0.664	0.824	0.820

Table 4: The CHARM model outperforms end-to-end BERT on most tasks. We compare fact selection with MRR, dialog act prediction with micro-averaged F_1, and like prediction with accuracy. Ablating dialog history degrades context-dependent tasks (fact selection and policy act prediction), but not tasks more dependent on one message.

Like Prediction For every assistant message, the model predicts the likelihood of the user "liking" the message. We treat this as binary classification, predict the "like" likelihood

$$\hat{y}_i^l = \text{softmax}(\text{GELU}(\boldsymbol{W}^l \cdot \boldsymbol{h}^i + \boldsymbol{b}^l)), \quad (16)$$

and use it to compute the like loss

$$\mathcal{L}_l = \ell_{ce}(\hat{y}_i^l, y_i^l) \quad (17)$$

where y_i^l indicates if the message was liked. We train the model jointly and optimize the loss

$$\mathcal{L} = \mathcal{L}_f + \mathcal{L}_l + \mathcal{L}_p + \mathcal{L}_u. \quad (18)$$

See Appendix F for training details.

5 Modeling Experiments

CHARM improves over a BERT model in most tasks.

5.1 Evaluation

We evaluate each sub-task with separate metrics. Fact selection is evaluated with mean reciprocal rank (MRR). For utterances with at least one selected fact, we compute the MRR using the selected facts as relevant documents. We compare like prediction with binary classification accuracy. For utterance and policy act prediction, we compare models with micro-averaged F_1 scores so that frequent classes are weighted more heavily. For each metric, we report validation and test set scores.

5.2 Baselines

BERT (Devlin et al., 2018) is a standard baseline for many NLP tasks. We use a multi-task extension of an uncased BERT model as our primary baseline and fine-tune it for our unique set of tasks (E2E BERT). Specifically, we use the CLS representation of each utterance to replace the HRE representation as a time-distributed input to the same multi-task decoders (Section 4.3). The context-less CHARM ablation replaces the dialog contextualizer LSTM with a per-timestep projection layer. Lastly, we report majority class accuracy for classification tasks.

5.3 Discussion

The proposed CHARM model for conversational curiosity is more effective than E2E BERT for most of the tasks in Curiosity (Table 4). Specifically, CHARM improves significantly in fact prediction (13 MRR points) and both dialog act prediction tasks (5 F_1 points), demonstrating the efficacy of the structural encoding of the various input modalities. Generally, models accurately predict utterance acts and likes, but their MRR and F_1 scores on fact selection and policy act prediction is comparatively worse. To a degree, this is expected since there is not always one best fact or one best action to take as the assistant; there may be various reasonable choices, which is common in information retrieval tasks. Nonetheless, models that specifically reason about the relationship between prior knowledge and entities would likely yield improvement. For example, Liu et al. (2018) predict the most relevant unmentioned entity while Lian et al. (2019) model a posterior distribution over knowledge. We leave these improvements to future work.

6 Related Work

Our work builds on knowledge-grounded conversational datasets and modeling.

Datasets Although there are numerous grounded datasets, we did not find one for conversational information seeking that contained fine-grained knowledge groundings, message-level feedback from the user, and dialog acts. Table 5 compares the Curiosity dataset to several others according to six factors: (1) is the goal of the task information seeking, (2) is the dataset collected from natural dialog with one participant always taking the role of an assistant, (3) are dialog responses constrained, (4) are document groundings annotated—as opposed to distantly supervised—and fine-grained, (5) is there message level feedback for the assistant, and (6) is the dataset annotated with dialog acts.

Dataset	Info Seeking	Dialog w/Assistant	Free Response	Annotated Grounding	Message Feedback	Dialog Acts
Curiosity (ours)	✔	✔	✔	✔	✔	✔
Topical Chat (Gopalakrishnan et al., 2019)	✔	⚠	✔	✔	✔	⚠
Search as a Conversation (Ren et al., 2020)	✔	✔	✔	✔	✗	✗
Wizard of Wikipedia (Dinan et al., 2019)	✔	✔	✔	✔	✗	✗
QuAC (Choi et al., 2018)	✔	✔	✗	✔	✗	⚠
CMU DOG (Zhou et al., 2018b)	✔	✔	✔	⚠	✗	✗
MS Marco Conv. (Nguyen et al., 2016)	✔	✗	N/A	N/A	N/A	N/A
OpenDialKG (Moon et al., 2019)	✗	✔	✔	✔	✗	✗
CoQa (Reddy et al., 2019)	✗	✔	⚠	✔	✗	✗
Holl-E (Moghe et al., 2018)	✗	⚠	✔	✔	✗	✗
Commonsense (Zhou et al., 2018a)	✗	✗	✔	✗	✗	✗
Reddit+Wiki (Qin et al., 2019)	✗	✗	✔	✗	✗	✗

Table 5: ✔ indicates a dataset has the feature, ⚠ that it does with a caveat, and ✗ that it does not. Conversational MS MARCO is a search dataset but has inquiry chains we want assistants to induce (exemplar in Appendix G). Topical Chat and Search as a Conversation are motivationally similar. While our dataset's combination of (human) annotation is unique, all three datasets are steps forward in resources for conversational information-seeking.

Our dataset is most similar to those for information-seeking such as QuAC (Choi et al., 2018), Wizard of Wikipedia (Dinan et al., 2019, WoW), CMU DOG (Zhou et al., 2018b), MS MARCO (Nguyen et al., 2016), Topical Chat (Gopalakrishnan et al., 2019), the TREC Conversational Assistance track (Dalton et al., 2019, CAsT), and Search as a Conversation (Ren et al., 2020, SaaC). QuAC constrains assistant responses to spans from Wikipedia, which makes it better for conversational *question answering*, but prevents more sophisticated assistant policies. QuAC also provides dialog acts, but they exist so that the assistant can inform the user of valid actions; we annotate dialog acts after-the-fact so that we can compare *freely chosen* user responses. Like QuAC, Topical Chat, SaaC, and WoW have annotated knowledge-groundings for each message, but responses are free-form. SaaC is a contemporaneous, CAsT-based dataset that shares our motivation to make conversation a medium for information-seeking. Topical Chat includes user feedback, but instead of explicitly defined roles, workers implicitly take dual and alternating roles as the user and assistant through knowledge asymmetry; followup work added automatically annotated dialog acts to Topical Chat (Hedayatnia et al., 2020).

Many tasks instruct annotators to take on a specific role in the dialog. For example, in Wizard of Wikipedia, annotators assume an assigned persona (Zhang et al., 2018) in addition to being the user or assistant. Consequently, many dialogs revolve around personal discussions instead of teaching about a topic. Additionally, annotators may not have the background to play their role. In contrast, we ask annotators to take roles that—as humans—they already know how to do: read about and convey interesting information on a topic (assistant) and engage in inquiry about a novel topic (user).

Our work is one of many in knowledge-grounded conversational datasets. For example, Moghe et al. (2018) have workers discuss movies and ground messages to plot descriptions, reviews, comments, and factoids; however, one worker plays both roles. In OpenDialKG (Moon et al., 2019), annotators ground messages by path-finding through Freebase (Bast et al., 2014) while discussing and recommending movies, books, sports, and music. Qin et al. (2019) use Reddit discussion threads as conversations and ground to web pages. Similarly, Ghazvininejad et al. (2018) collect Twitter three-turn threads and ground to Foursquare restaurant reviews. Our work adds to this dataset compendium.

External Knowledge in Models Our model is related to those that incorporate external information like facts in question answering (Weston et al., 2015; Sukhbaatar et al., 2015; Miller et al., 2016), knowledge base triples in dialog models (Han et al., 2015; He et al., 2017; Parthasarathi and Pineau, 2018), common sense (Young et al., 2018; Zhou et al., 2018a), or task-specific knowledge (Eric and Manning, 2017). Similarly to Kalchbrenner and Blunsom (2013); Khanpour et al. (2016), CHARM predicts the act of the current message, but also next message's act like Tanaka et al. (2019) do.

7 Future Work and Conclusion

We see two immediate directions for future work. The first is to augment our CHARM model with a text generation module to make a digital version of our human assistants. This involves contextualizing and paraphrasing facts which our dataset supports. Second, dialog act sequences could identify additional data-driven policies that could be used to define rewards or losses. By conditioning on dialog acts or sequences of dialog acts, textual outputs could be better-controlled (Sankar and Ravi, 2019; See et al., 2019) and combined with knowledge grounding (Hedayatnia et al., 2020). However, text is not the native modality of digital assistants.

We envision digital assistants participating in information-seeking, which means handling speech input. Consequently, automatic speech recognition (ASR) introduces transcription errors which are especially prevalent in knowledge-oriented text like question answering (Peskov et al., 2019). Gopalakrishnan et al. (2020) show this is also problematic in information-seeking dialog by comparing models on textual and ASR versions of Topical Chat. To close the loop in conversational information-seeking, models need to account for the speech-based environment of digital assistants.

In summary, this work introduces Curiosity: a large-scale conversational information seeking dataset. With Curiosity's unique set of annotations, we design CHARM which jointly learns to choose facts, predict a policy for the next message, classify dialog acts of messages, and predict if a message will be liked. We hope that our dataset will encourage further interest in curiosity-driven dialog.

Acknowledgments

We thank Rajen Subba, Stephen Roller, Alborz Geramifard, and Scott Yih for insightful discussions. Thanks to Becka Silvert for improving the task guidelines and Victor Ling for building the dialog act annotation tool's backend. Thanks to Sara LaRocca and Hal Daumé III for advice on adapting Krippendorff's α. We thank our anonymous ACL and EMNLP reviewers as well as Shi Feng, Jordan Boyd-Graber, Joe Barrow, Karthik Gopalakrishnan, and UMD CLIP members for feedback on the paper. Rodriguez's contributions to this work while at UMD were supported by NSF Grant IIS-1822494. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

References

Hannah Bast, Florian Bäurle, Björn Buchhold, and Elmar Haussmann. 2014. Easy access to the freebase dataset. In *Proceedings of the World Wide Web Conference*.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Kôiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David R. Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of the Language Resources and Evaluation Conference*.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Kôiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R. Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Language Resources and Evaluation Conference*.

Seth Chaiklin. 2003. *The Zone of Proximal Development in Vygotsky's Analysis of Learning and Instruction*, Learning in Doing: Social, Cognitive and Computational Perspectives, page 39–64. Cambridge University Press.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of Empirical Methods in Natural Language Processing*.

Jeffrey Stephen Dalton, Chen-Yan Xiong, and James P. Callan. 2019. TREC CAsT 2019: The conversational assistance track overview. In *Text REtrieval Conference*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations*.

Mihail Eric and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the Annual SIGDIAL Meeting on Discourse and Dialogue*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. *arXiv:1803.07640*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Scott Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Association for the Advancement of Artificial Intelligence*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Proceedings of the Annual Conference of the International Speech Communication Association*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Longshaokan Wang, Yang Liu, and Dilek Hakkani-Tur. 2020. Are neural Open-Domain dialog systems robust to speech recognition errors in the dialog history? an empirical study. In *Proceedings of the Annual Conference of the International Speech Communication Association*.

Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of the Association for Computational Linguistics*.

Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of the Annual SIGDIAL Meeting on Discourse and Dialogue*.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the Association for Computational Linguistics*.

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-Driven neural response generation for Knowledge-Grounded dialogue systems. *arXiv:2005.12529*.

Dan Hendrycks and Kevin Gimpel. 2017. Gaussian error linear units (GELUs). *arXiv:1606.08415*.

Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6:429–449.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*.

J. F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2:26–41.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of International Conference on Computational Linguistics*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

Klaus Krippendorff. 2004. *Content Analysis: an Introduction to its Methodology*. Sage. Chapter 11.

Ping Li, Christopher J. C. Burges, and Qiang Wu. 2008. Learning to rank using classification and gradient boosting. In *Proceedings of Advances in Neural Information Processing Systems*.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *International Joint Conference on Artificial Intelligence*.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the Association for Computational Linguistics*.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. ParlAI: A dialog research software platform. *arXiv:1705.06476*.

Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of Empirical Methods in Natural Language Processing*.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of Empirical Methods in Natural Language Processing*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the Association for Computational Linguistics*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation*.

Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *Proceedings of Empirical Methods in Natural Language Processing*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch:

An imperative style, high-performance deep learning library. In *Proceedings of Advances in Neural Information Processing Systems*.

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(85):2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Denis Peskov, Joe Barrow, Pedro Rodriguez, Graham Neubig, and Jordan Boyd-Graber. 2019. Mitigating noisy inputs for question answering. In *Proceedings of the Annual Conference of the International Speech Communication Association*.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *Proceedings of the Association for Computational Linguistics*.

Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the Conference on Human Information Interaction and Retrieval*.

Anand Rajaraman and Jeffrey David Ullman. 2011. Data mining. In *Mining of Massive Datasets*, pages 1–17. Cambridge University Press.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational AI: The science behind the alexa prize. *arXiv:1801.03604*.

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.

Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten de Rijke. 2020. Conversations with search engines. *arXiv:2004.14162*.

Chinnadhurai Sankar and Sujith Ravi. 2019. Deep reinforcement learning for modeling chit-chat dialog with discrete attributes. *Proceedings of the Annual SIGDIAL Meeting on Discourse and Dialogue*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2015. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Association for the Advancement of Artificial Intelligence*.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. *Proceedings of the ACM International Conference on Information and Knowledge Management*.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Proceedings of Advances in Neural Information Processing Systems*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of Advances in Neural Information Processing Systems*.

Koji Tanaka, Junya Takayama, and Yuki Arase. 2019. Dialogue-act prediction of future responses based on conversation history. In *Proceedings of the Association for Computational Linguistics*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Pei hao Su, Stefan Ultes, David Vandyke, and Steve J. Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. *Proceedings of the International Conference on Learning Representations*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2020. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. In *Proceedings of Empirical Methods in Natural Language Processing*.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Association for the Advancement of Artificial Intelligence*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the Association for Computational Linguistics*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In *International Joint Conference on Artificial Intelligence*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018b. A dataset for document grounded conversations. In *Proceedings of Empirical Methods in Natural Language Processing*.

A Components of Dialog Interfaces

In this section, we provide short descriptions and screenshots of every component of the user and assistant dialog interfaces.

A.1 User's Interface

Figure 5 shows the interface that we use to sample the user's prior knowledge of entities related to the topic. To derive a diverse sample, we use Wikipedia page views as a proxy for how well known the entity is. All experiments use the English Wikipedia dump generated on July 23, 2019. We divide entity mentions into ten buckets based on the frequency of page views, and round-robin sample fifteen entities from those buckets. The interface is shown before the user starts chatting with the assistant.

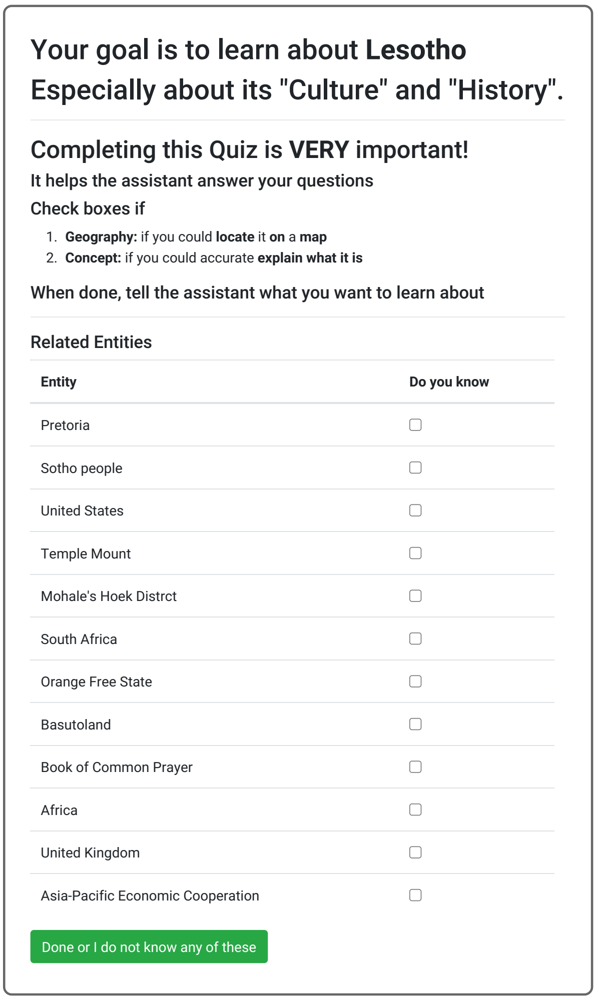

Figure 5: In this example, the user is assigned to learn about Lesotho, specifically its *culture* and *history*. In addition to their training with guidelines and videos, we repeat the instructions here. The related entities span relatively common ones like the United States or Africa to less known ones such as Basutoland.

We elicit how "interesting" a user finds each of the assistant's messages through the like button in Figure 6. Only users can "like" a message; the assistant cannot "like" user messages. Users are instructed to "like" messages if they are "interesting, informative and/or entertaining" and "relevant to their topic and/or aspects." They are specifically instructed not to "like" messages that are devoid of factual content, only express feelings, or only contain greetings or farewells.

Switching Aspect Users are randomly assigned two aspects for each dialog and told to spend time discussing each. The guidelines instruct them to spend at least two turns per topic, but we do not specify any further time requirements. When the user changes aspects, we instruct them to click a button (Figure 7) to indicate when and which aspect they are switching to. Additionally, this event triggers a reset in the context we use to rank the assistant's facts.

A.2 Assistant Interface

By design, we intend for most workers to not be familiar in depth with most of the geographic topics. Thus, the most important responsibility of the assistant interface is to provide enough information—without overwhelming them—to be engaging conversational partners. The first interface shown is a short description of the topic from either Simple Wikipedia or the English Wikipedia. This component helps the assistant reach a general understanding of the topic so that they can choose better facts.

The most important component of the assistant interface is their list of available facts. These facts have high textual similarity with the most recent three turns and are broken into three categories: facts related to entities the user knows about (rooted facts), facts related to an aspect (aspect facts), and facts from anywhere on the page (general facts). Feedback from pilot collections showed that six facts was too few which caused a lack of relevant facts, but twelve facts overwhelmed annotators. Thus, we use nine facts so that we can also balance equally across each type of fact. When composing their reply, the assistant can use any number of facts as in Figure 9. To discourage verbatim copying, we disable the paste feature in javascript. We also drop repeatedly unused facts.

Figure 6: The user expresses the "interestingness" of the assistant's messages through a "like" button (right of message). The instructions are shown prominently in the full interface and repeated in training material.

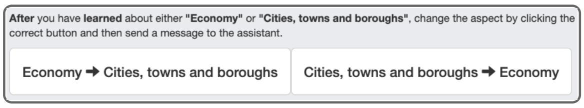

Figure 7: The user is assigned two aspects about their topic. After they are satisfied with what they have learned about the first aspect, they click a button and switch to the next aspect. While the button click is not communicated to the assistant (the user must send a corresponding message), it resets the fact contextualizer; we observe that without this, too many facts were related to the previous aspect.

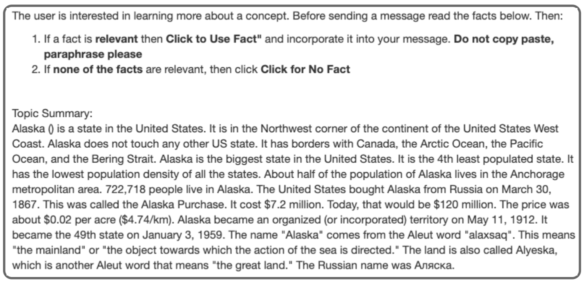

Figure 8: A short topic description is always visible to the assistant. The goal is to ensure the assistant always has a general understanding of the dialog topic.

B Dialog Act Annotation

To annotate dialog acts, we create a separate annotation interface (Figure 10). The interface shows one dialog at a time, and the same annotator annotates all the utterances. In addition to the utterances, the interface shows the topic, aspects, and sender of each message. Lastly, we incorporate a "Report Dialog" feature to help identify and remove inappropriate dialogs.

C Sample Dialogs

Tables 6 and 7 show Curiosity dialogs and highlight the dataset's features. Typos and grammatical errors made by annotators are left unaltered.

D Paraphrase Analysis and Samples

In Section 3.2.3, we describe the results of a manual analysis on two hundred and fifty assistant paraphrases. Annotations were completed by the authors and shown in Table 8. We break messages into four categories: paraphrases, copies, errors, and unrelated. Paraphrases include messages that incorporate the selected fact and possibly additional information. Copies include verbatim copying, cherry-picked phrases, and trivial contextualizations like replacing an entity with a pronoun. Table 9 shows ten randomly selected paraphrases from the two hundred and fifty manual annotations.

E Like Prediction Comparison

Like prediction is the one task where CHARM was not the best model. To better understand the differences between the CHARM and BERT model, we randomly sample thirty dialogs in the test set and find assistant messages where the model predictions disagree. Of the 202 assistant messages in the thirty dialogs, the BERT like prediction is correct 174 times (86.1%) and CHARM 170 times (84.2%). The models disagree on eight predictions (3.96%); BERT is correct on six of these and CHARM on the other two. Table 10 shows the messages, predictions, and labels. Although we might guess at why the predictions differ (e.g., perhaps something in the Florida dialog confused BERT), without further analysis we cannot establish causality.

F Model Training, Implementation, and Computation

We implement all models with PyTorch (Paszke et al., 2019) and AllenNLP (Gardner et al., 2018). The learning rates for models is set using the built-in learning rate finder in AllenNLP. Model losses were optimized with Adam (Kingma and Ba, 2015);

	Topic:	Djibouti
	Aspects:	*Economy*, *Politics*
	Known Entities:	France, Africa
	Dialog Acts	Messages
U	`request topic`	Hi can you tell me about Djibouti?
A	`inform response`	The country of Djibouti is officially called the Republic of Djibouti is located in the horn of Africa. It was created out of French Somaliland. It is also recorded the country was colonized by France in1800s and later gained its independence on June 27, 1977.
U	`request aspect`	Can you tell me about their Politics
A	`inform response`	After its independence, The country had two regiments commanded by the French officers. The country also maintains its close ties with Somalia, Ethiopia, France and the United States.
U	`request aspect`	Any other information about the French and their impact of the politics?
A	`inform response`	There are no direct information on its effect on the country's politics but even though the country is already independent. The 13th Demi-brigade of the French Foreign Legion is still stationed in this country. It is also the largest french military presence abroad. commanded by a 3-star general.
U	`request aspect`	Heavy French infeleunce. what other interesting inofrmation on politics do you have?
A	`inform response`	In recent years, the country has improved its training techniques, military command, and information structure. They also became more self-reliant in supplying their military and assist in peacekeeping missions and provide military assistance to countries that officially ask them for help.
U	`request aspect, feedback positive`	Wow Interesting, Can you now tell me about the economy they have?
A	`offer aspect`	Sorry, I do not have that information. But I can tell you more about Politics.
U	`offer accept`	Sure
A	`inform response`	Camp Lemonnier is rented by the United State as a Naval Expeditionary Based for $63 Million a year. While Japan and France each pays $30 million a year and China pays $20 Million a year.
U	`request other`	Lastly, any any fun facts?
A	`inform response`	I am not sure if this is a fun fact, but the country's gross domestic product expanded by more than 6%. From $341 Million to 1.5 Billion
U		That's a huge increase. thank you for all your help
A		You are welcome

Table 6: Example dialog #1 from Curiosity. (U: User, A: Assistant)

the BERT model uses a learning rate of .0001 and CHARM a learning rate of .001 with otherwise default parameters. We train for a maximum of forty epochs and early stop based on the sum of validation losses. The CHARM model uses batch size 64 and the BERT model batch size 4. Our best model (CHARM), has 26, 970, 475 parameters, takes two hours and eighteen minutes to train, and early stops on epoch fifteen. In our models, text encoders for utterances and facts share parameters.

Models were developed on a single machine with eighty Intel 2.0GHz CPUs, 256GB RAM, and eight Tesla V100 graphics cards. Each model was trained and evaluated on a single graphics cards with hyper-parameter sweeps parallelized across the eight cards.

AllenNLP configuration files and software dependencies (including version) are included in our code at github.com/facebookresearch/curiosity.

G MS Marco Conversational Sample Queries

Conversational MS MARCO is a search dataset that partially inspired this work. Assistant messages should prompt followup queries like in Table 11.

Topic:	<u>British Columbia</u>
Aspects:	*Government and politics*, *Culture*
Known Entities:	Canada, Seattle

	Dialog Acts	Messages
U	`request topic`	Hi! Can you help me learn some basic information about British Columbia? I don't know much except that it's located in Canada.
A	`inform response`	Yes, British Columbia is the westernmost province of Canada and is located between the Rocky Mountains and the Pacific Ocean.
U	`request aspect, feedback positive`	I didn't know it was on the coast! What can you tell me about government and politics there?
A	`inform response`	One interesting fact about the government is that the Green Part plays a larger role in this province than it does in other provinces of Canada.
U	`request followup, feedback positive`	Interesting. What can else you tell me about the Green Party?
A	`inform response`	The New Democratic Party and the Green Party caucuses together control 44 seats. Which seems like a lot but the British Columbia Green Party only takes up 3 of those 44 seats.
U	`request aspect`	That's a pretty small influence. Can you tell me some fun culture facts about British Columbia?
A		I am sorry I do not have any information on their culture right now.
U	`request topic`	That's okay. What other fun facts can you share?
A	`inform response`	Interestingly, Queen Victoria chose British Columbia to distinguish what was the British sector of the Columbia District from the United States which became the Oregon Territory on August 8, 1848.
U	`request aspect`	So that's why it has "British" specifically as part of it's name! Makes sense. Are there any sports or outdoor activities that are popular in British Columbia?
A	`inform response`	Horseback riding is enjoyed by many British Columbians.
U		Thanks for your help today. Now I know more than I did before.
A		No problem, it was a pleasure.

Table 7: Example dialog #2 from Curiosity. (U: User, A: Assistant). After mentioning the Green Party, the user asks a specific followup question; we use these interactions to estimate implicit preference.

Category	Label	Count	Percent
Copy	verbatim	68	27.2%
Copy	cherry-pick	6	2.40%
Copy	context	30	12.0%
Copy	Total	104	41.6%
Paraphrase	paraphrase-correct	111	44.4%
Paraphrase	paraphrase-multiple	17	6.80%
Paraphrase	Total	128	51.2%
Error	paraphrase-error	5	2.00%
Unrelated	unrelated	13	5.20%
Total		250	100%

Table 8: We analyze the paraphrases annotators use through manual categorization. The "Copy" category includes cherry-picked verbatim phrases, verbatim copies, and contextualized copies (e.g., changing a named entity to "it"). The majority of paraphrases are correct and only incorporate the provided fact, but a few weave in other information. 7.2% of paraphrases are either unrelated to the selected facts or paraphrase the fact incorrectly. Overall, 51.2% of messages have valid paraphrases.

Facts

Entity	Section	Fact	
Anchorage, Alaska	Economy	Military bases are a significant component of the economy in the Fairbanks North Star , Anchorage and Kodiak Island boroughs , as well as Kodiak .	Fact Used
United States	Economy	The Trans-Alaska Pipeline can transport and pump up to 2.1 Moilbbl of crude oil per day , more than any other crude oil pipeline in the United States .	Fact Used
United States	Body	United States armed forces bases and tourism are also a significant part of the economy .	Fact Used
Alaska Natives	Economy	Many Alaskans take advantage of salmon seasons to harvest portions of their household diet while fishing for subsistence , as well as sport .	Click to Use Fact
Yakutat, Alaska	Cities, towns and boroughs	Yakutat City , Sitka , Juneau , and Anchorage are the four largest cities in the U.S. by area .	Click to Use Fact
Unorganized Borough, Alaska	Cities, towns and boroughs	The remaining population was scattered throughout Alaska , both within organized boroughs and in the Unorganized Borough , in largely remote areas .	Click to Use Fact
Oregon Territorial Legislature	State symbols	The willow ptarmigan is common in much of Alaska . , State fish : king salmon , adopted 1962 . , State flower : wild / native forget-me-not , adopted by the Territorial Legislature in 1917 .	Click to Use Fact
Pacific Ocean	Body	The Pacific Ocean lies to the south and southwest .	Click to Use Fact
Arctic	Geography	The climate in the extreme north of Alaska is Arctic (Köppen : ET) with long , very cold winters and short , cool summers .	Click to Use Fact
			Click for No Fact

THEM: Thats cool, I love salmon. What are their other main exports economically speaking?

YOU: The US armed forces through military bases and tourism are large parts of the economy. Alaska also has a large oil business with the Trans Alaska pipeline transporting more oil than any other crude oil pipeline.

Figure 9: The assistant could incorporate any number of facts into their reply to the user. Their goal was to answer the user's immediate questions, and anticipate what information they would be most interested in.

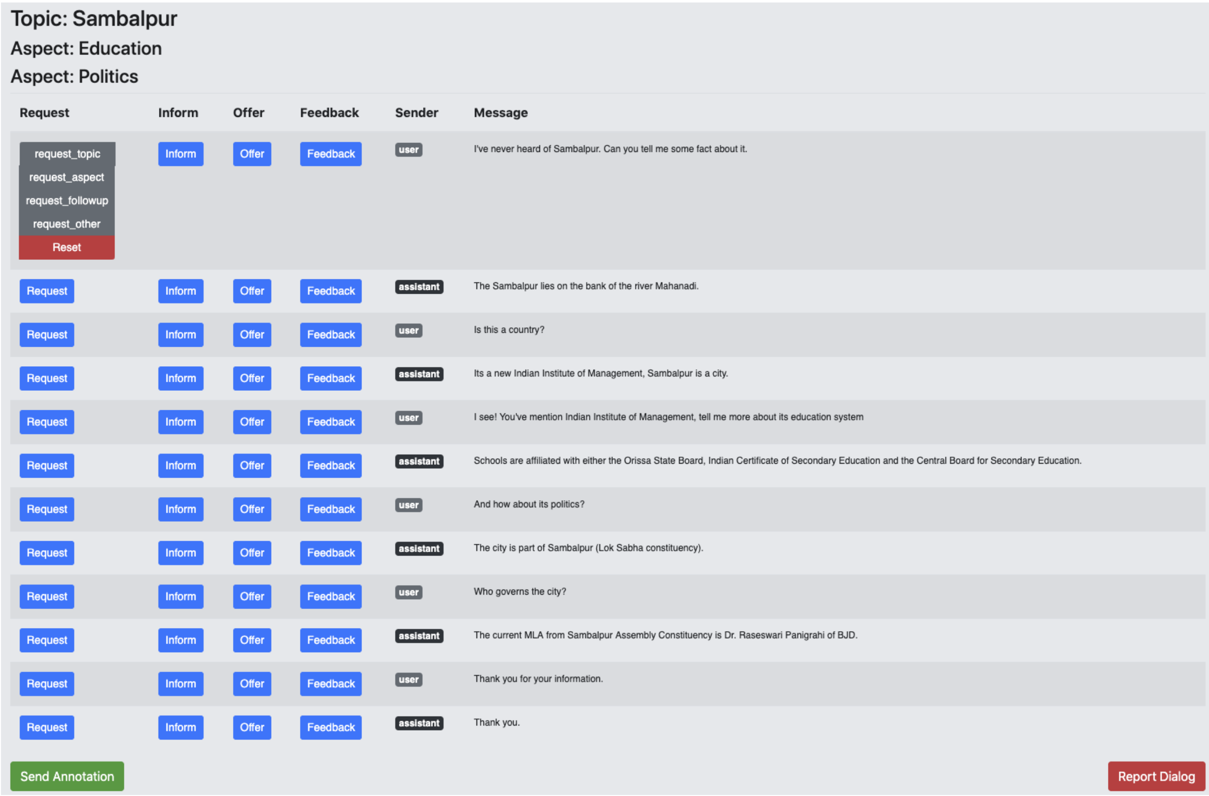

Figure 10: To annotate dialog acts, we develop an interface that showed each utterance on a separate line. Annotators assign zero or more dialog acts to each utterance using grouped dropdowns.

Label: paraphrase-correct
Message: I do not have the information available but I can tell you 10 million tones were exported annually for, Cardiff dock in the years leading up to the First World War.
Fact: In the years leading up to the First World War , more than 10 million tonnes of coal was exported annually from Cardiff Docks .

Label: copy
Message: Media and journalists in Namibia are represented by the Namibia chapter of the Media Institute of 'southern Africa and the Editors Forum of Namibia.
Fact: Media and journalists in Namibia are represented by the Namibian chapter of the Media Institute of Southern Africa and the Editors ' Forum of Namibia .

Label: paraphrase-multiple
Message: The film industry has a rich history in Indonesian culture and the Indonesian Film Festival has been held regularly since 1955.
Fact: Indonesia has held annual film festivals and awards , including the Indonesian Film Festival (Festival Film Indonesia) that has been held intermittently since 1955 .

Label: context
Message: EU means European Union. GDP per capital ranges from 188% of EU average in Bratislava to 54% in eastern Slovakia
Fact: GDP per capita ranges from 188 % of EU average in Bratislava to 54 % in Eastern Slovakia .

Label: paraphrase-correct
Message: Sudan was considered the 17th growing economy in the world. Its faster development of the country from oil profits.
Fact: Sudan , officially the Republic of the Sudan , is a country in Northeast Africa .

Label: paraphrase-correct
Message: There is a mountain range called the Annamite Range that runs along eastern border to Vietnam and another range, named Luang Prabang Range, near the Thailand border.
Fact: The Mekong River forms a large part of the western boundary with Thailand , where the mountains of the Annamite Range form most of the eastern border with Vietnam and the Luang Prabang Range the northwestern border with the Thai highlands .

Label: copy
Message: In 1603, the kingdoms of England, Scotland and Ireland were united in a personal union when James VI, king of Scots, inherited the crowns of England and Ireland and moved his court from Edinburgh to London; each country nevertheless remained a separate political entity and retained its separate political, legal and religious institutions. Would you like to know anything else?
Fact: In 1603 , the kingdoms of England , Scotland and Ireland were united in a personal union when James VI , King of Scots , inherited the crowns of England and Ireland and moved his court from Edinburgh to London ; each country nevertheless remained a separate political entity and retained its separate political , legal , and religious institutions .

Label: copy
Message: The A31 connects to the A35 at Bere Regis , and passes east through Wimborne and Ferndown to Hampshire where it later becomes the M27
Fact: The A31 connects to the A35 at Bere Regis , and passes east through Wimborne and Ferndown to Hampshire , where it later becomes the M27 .

Label: context
Message: I'm not sure of that. One fact I do have is that by 1857, Chicago was Illinois' largest city.
Fact: By 1857 , Chicago was Illinois 's largest city .

Label: paraphrase-correct
Message: On March 2, 2008, Dmitry Medvedev was elected President of Russia and Vladimir Putin became Prime Minister.
Fact: On 2 March 2008 , Dmitry Medvedev was elected President of Russia while Putin became Prime Minister .

Table 9: A random sample of ten manually labeled paraphrases from the assistant. The top row indicates the label we (the authors) annotated, the middle row the message, and the bottom row the original fact from Wikipedia. The original fact is shown as displayed to crowd-workers including punctuation tokenization.

Liked	Correct Model	Message
No	BERT	You are welcome!
Yes	BERT	I'm sorry I don't have anymore information about the etymology of Tunisia, but what I can tell you is that Tunisia Sports City is a whole sports city being constructed in Tunis
Yes	BERT	Yes Buddhism is a dominant influence in Lao culture. It has been great helping you.
Yes	CHARM	Florida is a state in the southeast United States. What would you like to know?
Yes	BERT	They have an average daily temperature of 70.7, it's the warmest state in the U. S.
Yes	CHARM	Yes, I can. Florida is nicknamed the "Sunshine State", but severe weather is a common occurrence.
Yes	BERT	Hello, Indonesia is part of the Malay Islands and is in Southeast Asia. Would you like to know more about the history?
Yes	BERT	I do not have etymologic information, would you like to know more about the economy? I can tell you thank Indonesia develops military and commuter aircraft.

Table 10: To compare like prediction between models, we randomly sample thirty dialogs and obtain predictions from CHARM and BERT. The table only shows messages where the model predictions disagree and indicates which model was correct. Dialogs are delineated by horizontal lines. Unfortunately, from only these examples we cannot determine why the CHARM model errors in most of these predictions.

Query
What is a physician's assistant?
What are the educational requirements required to become a physician's assistant?
What does the education to become a physician's assistant cost?
What's the average starting salary of a physician's assistant in the UK?
What's the average starting salary of a physician's assistant in the US?
What school subjects are needed to become a registered nurse?
What is the physician's assistant average salary vs a registered nurse?
What the difference between a physician's assistant and a nurse practitioner?
Do nurse practitioners or physician's assistant's make more?
Is a physician's assistant above a nurse practitioner?
What is the fastest way to become a nurse practioner?
How much longer does it take to become a doctor after being a nurse practitioner?
What are the main breeds of goat?
Tell me about boer goats.
What goat breed is good for meat?
Are angora goats good for meat?
Are boer goats good for meat?
What are pygmy goats used for?
What goat breed is the best for fiber production?
How long do Angora goats live?
Can you milk Angora goats?
How many Angora goats can you have per acre?
Are Angora goats profitable?

Table 11: An exemplar query chain from the conversational variant of MS MARCO. An ideal assistant should answer these questions *and* inspire these types of followup questions.