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

Pedro Rodriguez[1]*, **Paul Crook**[2], **Seungwhan Moon**[2], **Zhiguang Wang**[2]
[1]University of Maryland, Computer Science
[2]Facebook Assistant
pedro@cs.umd.edu, {pacrook, shanemoon, zgwang}@fb.com

Abstract

Open-ended human learning and information-seeking are increasingly mediated by technologies like digital assistants. However, such systems often fail to account for the user's pre-existing knowledge, which is a powerful way to increase engagement and to improve retention. 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 that relate to their existing knowledge. Through crowd-sourcing of this experimental task we collected and now open-source 14K dialogs (181K utterances) where users and assistants converse about various aspects related to geographic entities. This dataset is annotated with pre-existing user knowledge, message-level dialog acts, message grounding to Wikipedia, user reactions to messages, and per-dialog ratings. Our analysis shows that responses which incorporate a user's prior knowledge do increase engagement. We incorporate this knowledge into a state-of-the-art multi-task model that reproduces human assistant policies, improving over content selection baselines by 13 points.

1 Introduction

Conversational agents such as Alexa, Siri, or Google Assistant[1] 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, fact-seeking scenario where a user initiates a conversation with a digital assistant by asking an open-ended question and then drilling down into areas that are of interest, e.g.,

"<assistant wake-word>, tell me about Tahiti."
"Tell me more about its demographics."
"How about the cuisine?"

In such a setting, what policies should digital assistants pursue to maintain the user's interest in the topic? Theories of human learning such as Vygotsky's zone of proximal development propose that learning novel skills or information should be based to pre-existing knowledge and skills of the learner (Chaiklin, 2003). Considering this, a good policy might give general information about Tahiti; a better policy would select information related to the user's prior knowledge. We hypothesize that user engagement is strongly correlated with policies that integrate a user's pre-existing knowledge, and test it through a large-scale, Wizard-of-Oz (WoZ) style collection (Kelley, 1984; Wen et al., 2016) with a carefully instrumented interface to collect the assistant's policies and user's reactions. The resulting Curiosity dataset consists of 14,048 dialogs annotated with sentence-level knowledge grounding, user's prior knowledge, dialog acts per utterance, and message-level preferences.[2]

In our dialog task (Figure 1), one crowd-worker takes the role of a curious user learning about a prominent geographic entity and the other that of a digital assistant with access to a broad set of Wikipedia facts. 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., *infrastructure* and *education*) to investigate. The topic is also associated with various entities. The user engages in open-ended discovery about the topic; the assistant's goal is to simultaneously answers

*Work done while interning at Facebook.
[1]Facebook does not own the preceding trademarks.
[2] Dataset and code at curiosity.pedro.ai.

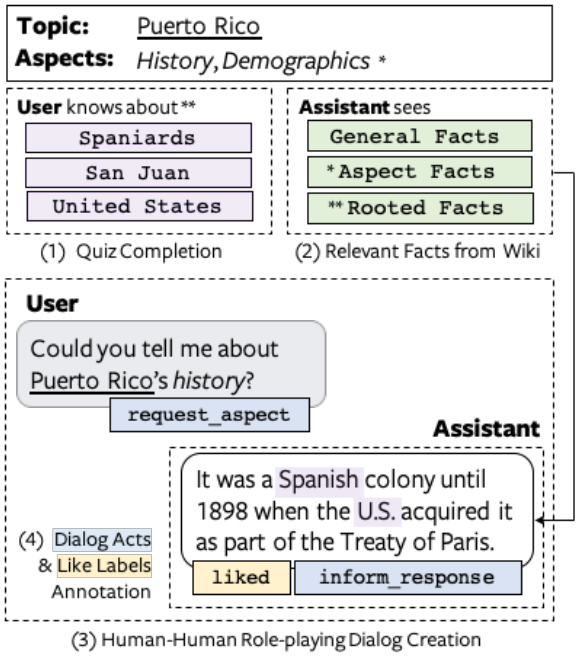

Figure 1: In the task, a user learns about a geographic entity (Puerto Rico) with emphasis on two aspects. The teacher incorporates Wikipedia facts into their messages; facts are rooted in the user's prior knowledge, aspect specific, or generic to the topic. The user explicitly expresses engagement with a like button and implicitly through their message's dialog act.

the user's questions while pro-actively introducing facts likely to prompt followup questions. For example, if the assistant knew of a user's familiarity with astronomy when providing information about Puerto Rico, then the user is more likely to engage with and remember facts about the Arecibo Observatory. Section 2 describes the dialog collection steps and the interface components that record assistant policies and user reactions.

Section 3 uses dialog act annotations combined with explicit and implicit user feedback to compare the assistant's content selection and presentation policies. For example, for interactions where the user asks a question and the assistant replies with content from a specific fact, how often does the user ask a followup question versus trail off in disinterest? Most datasets do not have sufficient annotation to answer these questions (see Section 6 for a detailed comparison to other knowledge-grounded datasets): 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). Following previous work, we use BERT to encode messages (Devlin et al., 2018), a Hierarchical Recurrent Encoder model (Serban et al., 2015) to encode dialog state, and jointly train the model with a multi-task objective function. Our experiments show that our model improves over baselines, and ablation studies show the importance of including the user's prior knowledge.

In summary, we make three primary contributions: (1) we design an experiment to test the efficacy of personalizing conversational information systems through a user's prior knowledge and (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) present baseline multi-task conversational models incorporating both dialog contexts and user's prior knowledge.

2 Building the Curiosity Dataset

This section describes the construction of the Curiosity dataset. The dialog are focused on prominent geographic entities distributed throughout the world. The *worldwide* geographic spread of entities makes each topic novel to most users. The consistent topic type makes it easier to start a dialog, while the associated rich histories, demographics, economics, etc., allows for a diverse set of dialogs. For example, most people are only vaguely familiar with the history Puerto Rico, but most know about related concepts such as the United States, Astronomy, or Hurricane Maria. Users can start conversations with questions about entities or features common across geography such as *demographics*, *economy*, or *government* and use it as a starting point to sojourner through its history.

The dataset construction consisted of building an interface for users and assistants (screenshots in Appendix A), collecting the dialogs, and annotating dialog acts. Section 2.1 describes how we select geographic topics, aspects, and derive a set of facts to ground against. Next, we describe how we incorporate these into WOZ interfaces for users and assistants (Section 2.2). Section 2.4 de-

signs an ISO-24617-2-based dialog act annotation schema (Bunt et al., 2010, 2012) to measure engagement with facts.

2.1 Geographic Topics, Aspects, and Facts

We obtain 361 geographic entities from Wikipedia by finding the subset of Wikipedia pages that also have separate geography and history pages (e.g., Puerto Rico, Geography of Puerto Rico, and History of Puerto Rico).[3] The existence of these pages is a signal of topical breadth and depth.

We take the text of these pages and build fact bank of 93,845 sentences for assistants to use. Similarly to Linked WikiText-2 (Logan et al., 2019), we run an entity linker over the content (Gupta et al., 2017). Next, we 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) which we use as a component in providing the teacher contextually relevant facts.

2.2 User and Assistant Dialog Interfaces

To collect dialogs, we build use and assistant dialog interfaces. The user's interface samples their prior knowledge of a topic, measures which assistant messages they find interesting, and manages the dialog context. The assistant's interface is primarily aimed at providing contextually relevant facts about the topic. Appendix A contains screenshots and details of each interface.

Sampling User's Prior Knowledge While deployed digital assistants can draw from prior interactions, we cannot, so instead we must incorporate this as part of the data collection. Instead of exhaustively asking about every entity related to the topic we sample this knowledge. Before the dialog begins, we show the user a sample of fifteen related entities that range from commonplace to obscure (United States versus Taíno). Users were told to mark the entities they could (1) locate on a map or (2) explain succinctly in one sentence.

Like Button for User Interest As part of our collection, we wanted to discover what kinds of fact-grounded utterances users found interesting. One direct measure was to elicit preferences through a like button next to each assistant message. Users were asked to "like" the assistant's message if they found it "interesting, informative, and relevant to their topic."

Assistant's Topic Summary and Fact Bank Most crowd-workers are not deeply familiar with most geographic entities which would—ordinarily—make them poor teachers to other crowd-workers. We alleviate this issue through an interface that provides contextually relevant facts to assistants. First, we impart a general understanding of the topic. Throughout the dialog, the assistant can read a brief description of their topic taken from simple.wikipedia.org or en.wikipedia.org. Second, the assistant can incorporate facts from a contextually updated fact bank (green box in Figure 1). They are told to select relevant facts, click a "use" button, and paraphrase the content into their next utterance.[4] We encourage them to "stimulate user interest and relate information to things they already know or have expressed interest in."

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. Specifically, we show the assistant a total of nine facts: three facts that mention an entity the user is familiar with (rooted facts), three facts from their assigned aspects (aspect facts), and three from anywhere on the page (general facts).[5] By construction, rooted facts overlap with the exclusive categories of aspect and general facts. For each category, we show the highest scoring facts (TF-IDF) and then randomize the order of all nine facts.[6] To avoid biasing the assistant, we do not inform them about the user's known entities or distinguish between types of facts (e.g., rooted, aspect, or general facts).

2.3 Conversation Data Collection

We crowd-sourced conversations in two phases using a customized version of ParlAI (Miller et al., 2017). In the first phase, we ran pilot studies and collected feedback from individual workers. Based on feedback, we created task guidelines,[7] tutorial videos, qualification tests, and in-tool instructions; we used these to train and qualify crowd-workers for the second phase. During this second phase, we monitored the usage of interface elements and re-

[3] We use the 07/23/19 dump and remove non-geo pages.

[4] We disable paste to discourage verbatim copying.

[5] Feedback from pilot collections showed six facts was too sparse and twelve overwhelmed workers.

[6] We also drop repeatedly unused facts.

[7] Includes extended instructions, examples of good and bad dialogs, and frequently asked questions.

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.

moved workers that were blatantly disregarding instructions. As with any large data collection, these steps do not solve all quality issues, but they greatly improved the quality of the Curiosity dataset.

2.4 Dialog Act Annotation

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

After dialog collection, we annotated all utterances with dialogs acts using a custom interface (screenshots in Appendix B). Following prior dialog work, we base our annotation schema on the ISO 24617-2 standard (Bunt et al., 2010, 2012) and customize sub-categories for our scenario. Functionally, we introduce finer grain distinctions for requests; superficially, we rename categories to avoid confusing annotators. Table 1 shows our annotation schema, descriptions, and brief examples.

Before annotating the full dataset, we first annotated the first 4,408 dialogs to decide whether to collect multiple annotations per dialog. In this first set, we annotated each dialog twice to measure inter-annotator agreement. Dialog act annotation is multi-class and multi-label: an utterance can have none, one, or multiple dialog acts (e.g., positive feedback and followup request). We adapt Krippendorff's α to this case as detailed in Appendix B.1. The computed agreement 0.834 is higher than the 0.8 significance threshold recommended by Krippendorff (2004) so we annotate the remaining dialogs only once.

The combination of dialog acts, knowledge groundings, and likes makes Curiosity unique. We analyze and model these signals next. Sample dialogs from Curiosity are included in Appendix C.

3 Dataset Analysis

Now we show basic statistics of the Curiosity dataset and use it to show that users consistently prefer topically relevant, rooted facts.

3.1 Dataset Statistics

Table 2 shows the basic statistics of the Curiosity dataset. In total, our dataset contains 14,048 dialogs with 181,068 utterances. Our fact database contains a total of 93,845 facts; of those, 76,120 were shown to the assistants and 27,486 were used in at least one message. For experiments, we first split-off thirty random topics and their dialogs as a zero-shot set and then split the remaining dialogs into training, validation, and testing folds.

3.2 What Facts do User Prefer?

In Section 1, we hypothesized that when assistants incorporate facts rooted in the user's prior knowledge that they will more likely remain engaged in the topic. In our data collection, we incorporated two mechanisms for testing this hypothesis. The first mechanism is explicit: we directly asked users—through the like button—to indicate what messages they preferred. The second mechanism is implicit and derived by mining dialogs for a specific sequence of dialog acts that suggest engagement with the content. For

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 2: The Curiosity dataset consists of 14,048 dialogs with an average of 12.9 utterances per dialog. Of 93,845 unique facts, 81% were shown at least once and 29% were used by an assistant at least once. About 60% of the assistants' 90,534 utterances were liked.

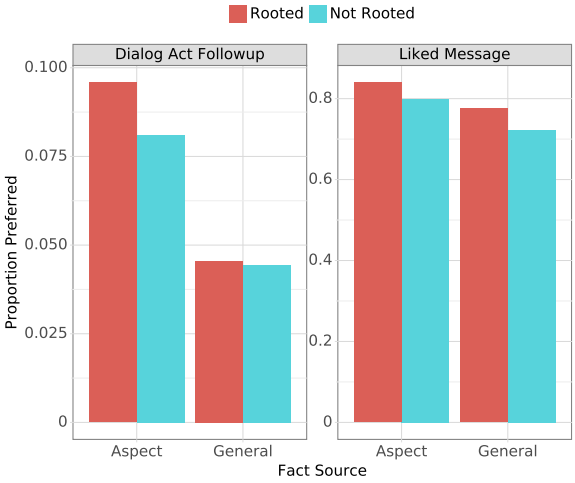

Figure 2: We measure user engagement through dialog act followups (left) and like button usage (right). Differences are statistically significant (99%+) in all comparisons *except* for dialog act followups between rooted and non-rooted general facts. Statistics were computed with a two proportion z-test. Users prefer on-aspect, rooted facts.

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). The likelihood in the data—shown in Figure 2—demonstrates that users prefer: (1) facts relevant to aspects versus general ones and (2) they prefer rooted facts in every case except for general facts measured by followups.

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 2. Users liberally use the like button (60% of messages are liked); nonetheless, the trend to prefer on-aspect, rooted facts is reproduced in likes.

3.2.2 Mining Acts for Implicit Preferences

We measure implicit preference by finding interactions where the assistant informs the user of some knowledge and the user engages directly by asking a targeted followup question. A specific followup question is a direct measure of engagement and implicit measure of preference. For example, asking about an entity like the Taínos is more specific than asking about history. To mine these patterns, we search for sequences of assistant-user messages where the assistant message is labeled with an "inform" dialog act, the assistant message uses a fact, and record the fact source. With the user exposed to a fact-grounded message from a specific source, we compute the preference likelihood

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

that the user's message is labeled as "request_followup." While this pattern is comparatively rare, the trend mirrors that of the like button usage: user give priority to aspect-oriented facts and then to rooted facts.

4 Models

We construct machine learning (ML) models that predict assistant and user actions. Concretely, our model (1) predicts the dialog acts of the user message (utterance act classification), (2) selects the best fact (fact prediction), (3) chooses the best set of dialog acts for the next message (policy act prediction), and (4) if the assistant message will be liked (like prediction).

4.1 Text Representation

Our model requires text representations of utterances and facts. We represent the textual content t_i^u of utterance u_i in dialog D as $E(t_i^u)$; E is an arbitrary text encoder that ouputs a fixed-size representation. Similarly, the representation of fact f_j on turn i is $E(t_{i,j}^f)$ where j indexes facts shown on that turn.[8] Our experiments compare two encoders. In the first, E is a bi-directional LSTM (Sutskever et al., 2014) over word embeddings initialized with GLOVE (Pennington et al., 2014) and entity embeddings initialized with Wikipedia2Vec (Yamada et al., 2020). The second encoder uses the CLS representation from uncased BERT (Devlin et al., 2018) without entity embeddings. In both cases, the output of the encoder is the primary input to a hierarchical dialog encoder.

[8] In our model, the text encoders share parameters.

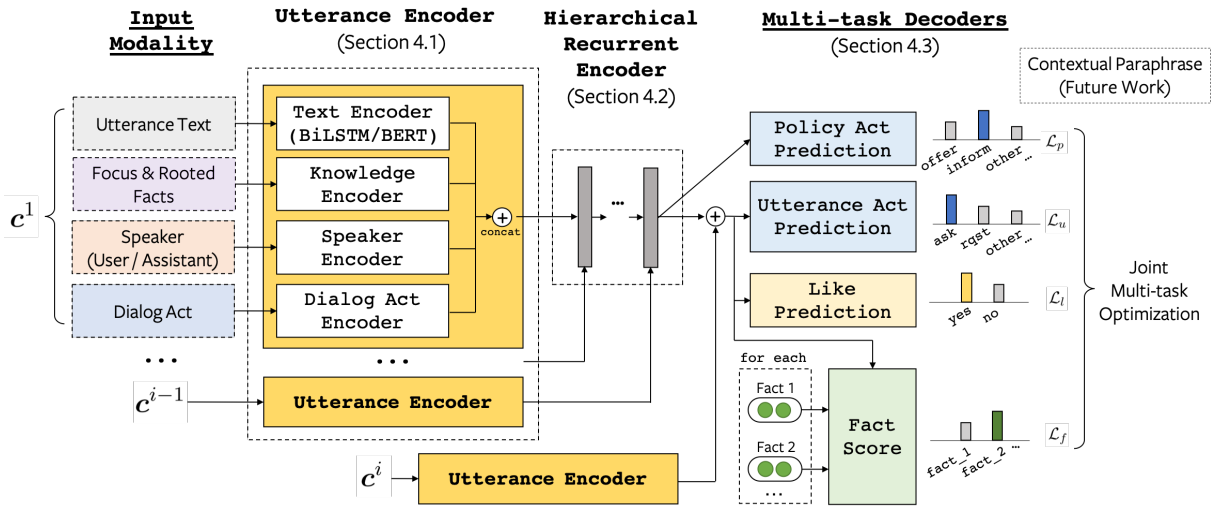

Figure 3: **Architecture**: Our model builds a dialog context up to $t = i - 1$ to predict the current message's dialog acts (policy prediction) and the best facts to use. The model uses this combined with the current utterance to classify it's dialog acts and if it will be liked. We leave building a paraphraser to mimic the assistant to future work.

4.2 Dialog Representation

In our models, we follow a similar 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. We use the same embeddings to represent the topic

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

of the dialog. Next, we create trainable 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 to 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.[9] With this joint representation, the contextualized dialog up to and including $t = i - 1$ becomes

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

by taking the final state of the LSTM. The dialog up to and including time i is

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

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

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

with parameters $\boldsymbol{W}^f$ and $\boldsymbol{b}^f$ and 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

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

are converted to probabilities by the softmax

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

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

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

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

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

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

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

Policy Act and Utterance Act Prediction Since each utterance may have multiple dialog acts we treat policy and utterance act prediction as a multi-class, multi-label task. The objective of the policy prediction is to choose the acts that the next utterance should have while the utterance act classifies the acts of a specific message. To predict these acts, we create a policy act network

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

and an utterance act network

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

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

$$\mathcal{L}_p = \sum_{k}^{|\mathcal{A}|} y^a_{i,k} \log p^{p,i}_k + (1 - y^a_{i,k}) \log(1 - p^{p,i}_k) \quad (14)$$

and utterance act loss

$$\mathcal{L}_u = \sum_{k}^{|\mathcal{A}|} y^a_{i,k} \log p^{u,i}_k + (1 - y^a_{i,k}) \log(1 - p^{u,i}_k) \quad (15)$$

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

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

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

and use it to compute the like loss

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

where y^l_i 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 D for training details.

5 Modeling Experiments

Our experiments show that our model significantly improves over baselines, and leave-one-out ablation studies show that taking advantage of the user's prior knowledge is important.

5.1 Evaluation

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

5.2 Baselines

We create baselines for like classification and fact selection. For like classification, we compare against the majority class (liked); for fact selection we use a TF-IDF-based ranker similar to Chen et al. (2017). Similarly, we use a majority class classifier as the dialog act baseline. Our matcher implementation uses word-level unigrams and bigrams, and inverse document frequencies are computed from the sentences in our Fact set. The facts are ranked by cosine similarity with the dialog text.

5.3 Discussion

Most HRE models for conversational curiosity improve significantly over both baselines on Curiosity (Table 3). Note also that the HRE+BiLSTM model outperforms the BERT-based counterpart, which could be due to the effective use of the wiki2vec entity embeddings (Section 4.1). 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 and this is not captured by these metrics. 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.

	Fact Rank		Utterance Act		Policy Act		Like	
	MRR		Micro-F_1		Micro-F_1		Accuracy	
Model	Val	Test	Val	Test	Val	Test	Val	Test
Majority Class			0.602	0.604	0.491	0.494	0.690	0.681
TF-IDF	0.415	0.408						
HRE+BERT	0.374	0.377	0.767	0.768	0.654	0.653	0.830	0.822
HRE+BiLSTM	0.546	**0.546**	0.845	0.847	0.682	0.682	0.826	0.815
- acts	0.549	0.545					0.830	0.822
- facts			0.845	0.846	0.681	0.682	0.819	0.811
- known	0.379	0.380	0.759	0.762	0.666	0.666	0.831	**0.825**
- likes	0.543	0.545	0.847	**0.850**	0.685	**0.688**		

Table 3: We compare MRR for fact selection, micro-averaged F_1 for dialog acts, and accuracy for likes. Ablating prior knowledge leads to absolute drops of 16.6% in MRR, 8.5% in utterance act F_1, and 1.6% in policy act F_1.

Ablation Study We analyze the influence of each input and label category (facts, dialog acts, and likes) by running a leave-one-out ablation study. For each category, we ablate the inputs, labels, and thus losses.[10] The exclusion of the users' prior knowledge has the largest adverse effect on the model with an absolute drop in fact MRR of 16.6%. However, ablating other input and label categories does not show any consistent trends. Overall, prior knowledge is the most important input—aside from utterances—for the models.

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. For example, a new TREC track on Conversational Assistance (Dalton et al., 2019) was created to promote interest in creating resources and evaluations for conversational information-seeking.

Table 4 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 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.[11] Of these datasets, ours is most similar to those aimed at information-seeking such as Quac (Choi et al., 2018), Wizard of Wikipedia (WoW) (Dinan et al., 2019), CMU DOG (Zhou et al., 2018b), MS MARCO (Nguyen et al., 2016), and Topical Chat (Gopalakrishnan et al., 2019).

Unlike most datasets, Quac constrains the response of the assistant to a span from Wikipedia. This makes it better for conversational *question answering*, but worse for training assistant policies for knowledge discovery (e.g., what fact would prompt followups from users). Quac also provides dialog acts, but these exist so that the assistant can inform the user of valid actions to take; we annotate dialog acts after-the-fact so that we can compare *freely chosen* user responses.

Like Quac, Topical Chat and WoW have annotated knowledge-groundings for each message, but user and assistant responses are both free form. Topical Chat includes user feedback for each message, but does not have dialog act annotations and participants take symmetric roles (i.e., there is no defined user or assistant). Symmetric roles is helpful for building grounded chit-chat systems, but not as helpful for building assistant systems to guide users in knowledge discovery.

In crowdsourcing it is common to 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 their role as the user or assistant. The outcome is that many dialogs revolve around personal discussions rather than teaching about a

[10] We implement this by clamping inputs and losses to zero.

[11] User/student and assistant/teacher are interchangeable.

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

Table 4: A comparison of knowledge-grounded datasets. ✔ indicates a dataset has the feature, ⚠ that it does but with a caveat, and ✗that it does not. The conversational MS MARCO is a search dataset, but contains the types of inquiry chains we want assistants to induce (exemplar in Appendix E).

topic. Additionally, annotators may not have the background to play particular roles. One example from CMU DOG is discussing movies they not have seen; the provided passages are helpful, but no replacement for having watched the movie. 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 restaurant reviews from Foursquare. Our work adds to this compendium of grounded datasets.

External Knowledge in Models Our modeling is most similar to those that incorporate external factual information. This includes memory networks in question answering (Weston et al., 2015; Sukhbaatar et al., 2015; Miller et al., 2016), and in dialog models using knowledge bases (Han et al., 2015; He et al., 2017; Parthasarathi and Pineau, 2018), common sense (Young et al., 2017; Zhou et al., 2018a), or task-specific knowledge (Eric and Manning, 2017). Similarly to Kalchbrenner and Blunsom (2013); Khanpour et al. (2016), our model predicts the dialog act of the current utterance, but also predicts the dialog act of the next utterance as Tanaka et al. (2019) does.

7 Future Work

We see two possible directions for future work. The first is to augment our multi-task policy model with a text generation module to make a digital version of our human assistants. A thorough evaluation would compare human user ratings of digital and human assistants. Second, we show that dialog act annotations—when based on an appropriate schema—can be used to identify desirable policies based on the user's actual reaction. Another direction for future work includes annotating dialog acts in existing datasets with the goal of mining policies. Conditioning models on dialog acts should lead to better control over model outputs just as discrete control variables do (Sankar and Ravi, 2019; See et al., 2019). Generally, we are also excited by the possibilities in improving information-seeking systems.

8 Conclusion

We introduce Curiosity: a large-scale dataset for conversational information seeking. The dialog task centers around a curious user asking about aspects of diverse and prominent geographic entities throughout the world. We describe its collection and show that users prefer messages with facts related to previously known entities (rooted). With Curiosity's unique set of annotations, we build a

HRE model that jointly learns to choose facts, determine a policy for the next message, classify dialog acts of messages, and predict if a message will be liked. We show that our model improves over baselines in each of these tasks. Finally, we outline two concrete directions for future work in grounded dialog generation and dialog policy.

Acknowledgements

We thank Rajen Subba, Stephen Roller, Alborz Geramifard, and Scott Yih for insightful discussions. Thanks to Becka Silvert for greatly improving task guidelines and Hal Daume for helpful tips regarding Krippendorff computations. Thanks to Shi Feng, Jordan Boyd-Graber, Joe Barrow, and CLIP members for useful feedback on paper drafts.

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.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the Association for Computational Linguistics*.

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*, abs/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*.

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*.

Dan Hendrycks and Kevin Gimpel. 2017. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *ArXiv*, abs/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: Thousand Oaks, CA. 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*.

Robert L Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. 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 preprint 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. *arXiv preprint arXiv:1611.09268*.

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, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

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.

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

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*, page 1–17. Cambridge University Press.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

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. 2016. 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 optimized tool for learning embeddings of words and entities from wikipedia. *arXiv preprint 1812.06280v3*.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2017. 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 appendix, we provide short descriptions and screenshots of every component of the user and assistant dialog interfaces.

A.1 User's Interface

Figure 4 shows the interface that we used to sample the user's prior knowledge of entities related to the topic. To derive a diverse sample, we used Wikipedia page views as a proxy for how well known it is. We divided entity mentions into ten buckets based on frequency of page views, and round robin sampled fifteen entities from those buckets. This interface was shown at the start of every dialog before the user sent the first message to the assistant.

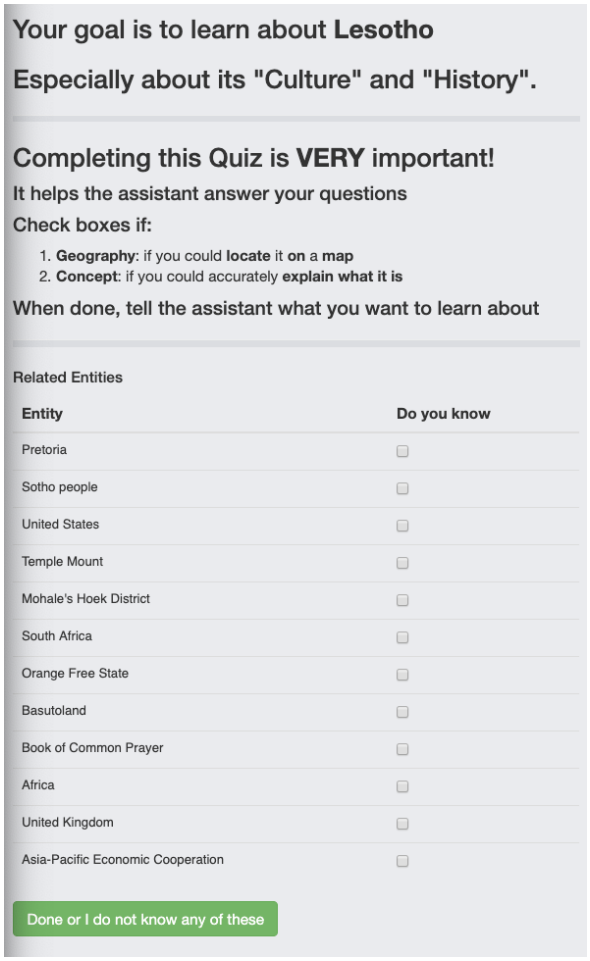

Figure 4: In this example, the user has been assigned to learn about Lesotho, specifically its *culture* and *history*. In addition to their training with guidelines and videos, we repeat the instructions here. The related entities span relatively common ones like the United States or Africa to less known ones such as Basutoland.

We elicit how "interesting" a user finds each of the assistant's messages through the like button in Figure 5. 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 were randomly assigned two aspects for each dialog and told to spend time discussing each. The guidelines instructed them to spend at least two turns per topic, but we do not specify any further time requirements. When the user changed aspects, we instructed them to click a button (Figure 6) to indicate when and which aspect they switched to. Additionally, this event triggered a reset in the context used to rank the assistant's facts.

A.2 Assistant Interface

By design, we intended 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 transform them into a just-in-time expert. The first interface shown was a short description of the topic from either Simple Wikipedia or the English Wikipedia. This component was designed to help the assistant reach a general understanding of the topic so they could choose better facts.

The most important component of the assistant interface was their list of available facts. These facts have high textual similarity with 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). When composing their reply, the assistant could use any number of facts as in Figure 8.

B Dialog Act Annotation

Figure 9 shows a screenshot of the customized dialog act annotation interface we created.

B.1 Krippendorff Score Calculation

Krippendorff agreement scores are typically computed in multi-class classification tasks where disagreements are more important than agreements. However, dialog act annotation is a multi-label and multi-class task. To our knowledge, there is no

Figure 5: The user expresses their opinion of 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 written and video training material.

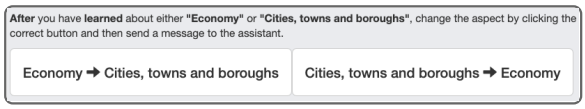

Figure 6: The user was assigned two aspects about their topic to learn about. After they are satisfied with what they have learned about the first aspect, we instructed them to click the button corresponding to their switch in aspect. While the button click is not communicated to the assistant (the user must send a corresponding message), it resets the fact contextualizer; we observed without this that too many facts were related to the previous aspect.

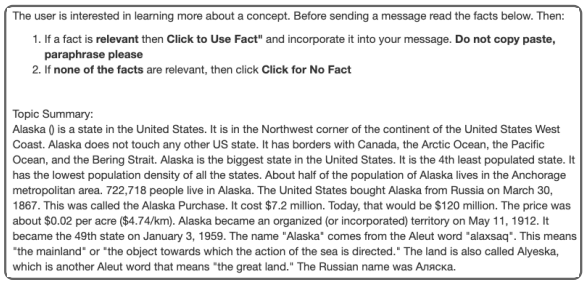

Figure 7: A short description of the topic is persistently shown to the assistant. The goal is to ensure the assistant always has a general understanding of the dialog topic.

standard method for calculating agreement scores in this scenario.

To compute our agreement score we convert the multi-label problem into a single-label problem. Typically, Krippendorff agreement scores are calculated from a matrix where rows represent examples, columns represent annotators, and entries indicate the class label. For example, Table 5 shows a typical way to represent two annotators, three examples, and binary labels. If the task were instead multi-label with binary labels, Table 5 shows how we convert the annotation of one utterance into multiple examples. Since there are few dialog acts per

	Annotator 1	Annotator 2
Utterance 1	A	B
Utterance 2	A	A

Table 5: A standard representation of annotations used to calculate Krippendorff agreement scores.

	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 6: Our representation of multi-label, multi-class annotations we use to calculate agreement scores. Each combination of label and example become one row in the matrix.

utterance, most annotations will agree; however, since Krippendorff focuses on disagreement it is appropriate for this scenario.

C Sample Dialogs from Curiosity

Table 7 and 8 show sample dialogs from Curiosity and the features of the dataset they demonstrate. Typos and grammatical errors are preserved (copy/paste were disabled in the collection tool to encourage contextual paraphrasing of grounded facts from Wikipedia.

D Model Training and Implementation Details

We optimize the model loss with Adam (Kingma and Ba, 2015) using .001 as the learning rate. We train for a maximum of twenty epochs, and early stop if the sum of the validation losses does not improve for three epochs. The BiLSTM model uses batch size 64 and the BERT model batch size 4. All models were implemented in PyTorch (Paszke

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

et al., 2017) and AllenNLP (Gardner et al., 2018)

E MS Marco Conversational Sample Queries

While MS MARCO is not a dialog dataset—it is a conversational *search* dataset—it is an inspiration for the direction of this work. Namely, responses from an agent should prompt query chains as in Table 9.

	Topic:	British Columbia
	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 8: Example dialog #2 from Curiosity. (U: User, A: Assistant). After mentioning the Green Party, the user asked a specific followup question. These are the interactions we mined to calculate implicit engagement with dialog acts.

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 9: Exemplar query chain from the conversational variant of MS MARCO.

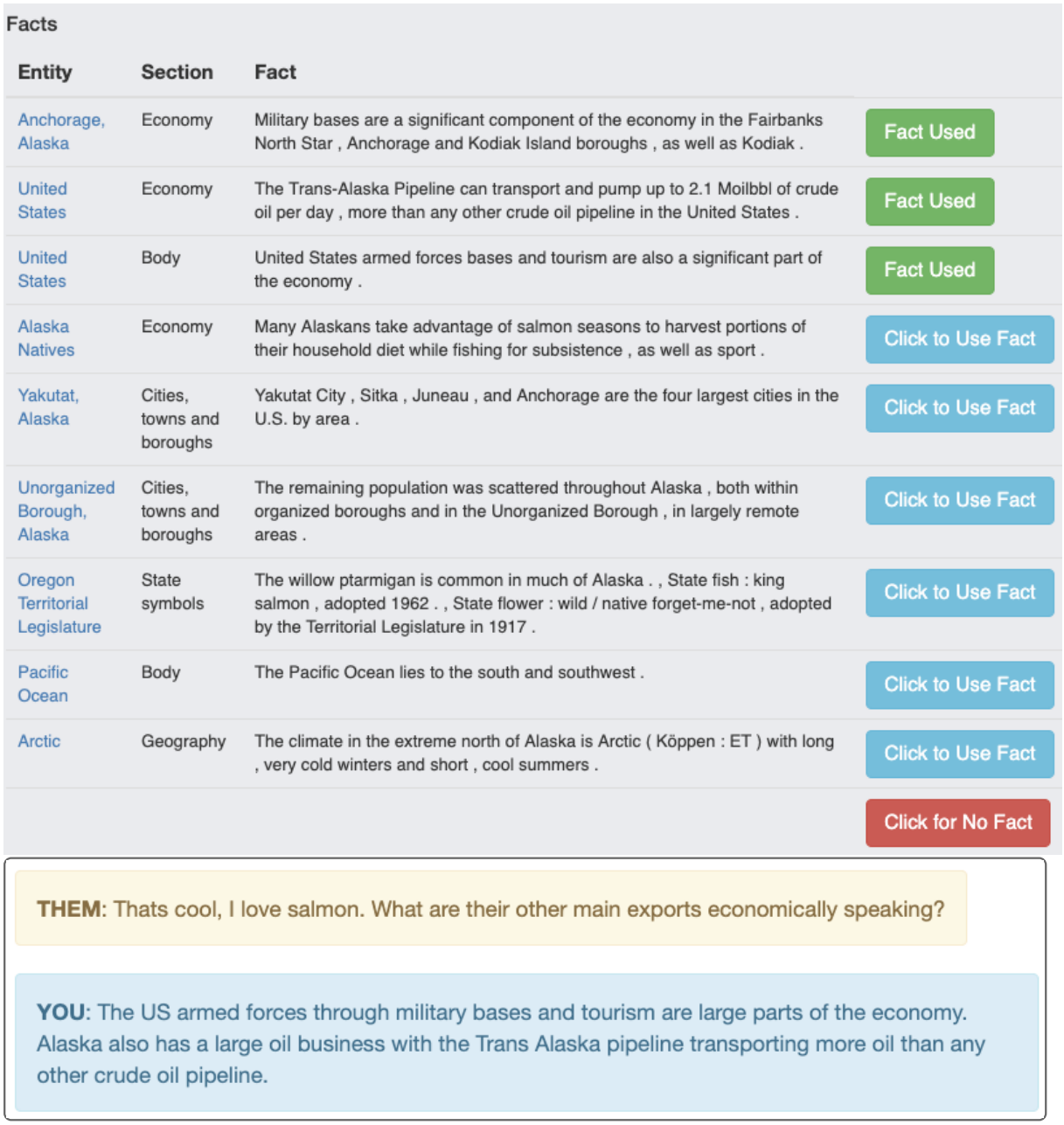

Figure 8: The assistant could incorporate any number of facts into their reply to the user. Their goal was to answer the user's immediate questions, and anticipate what information they would be most interested in.

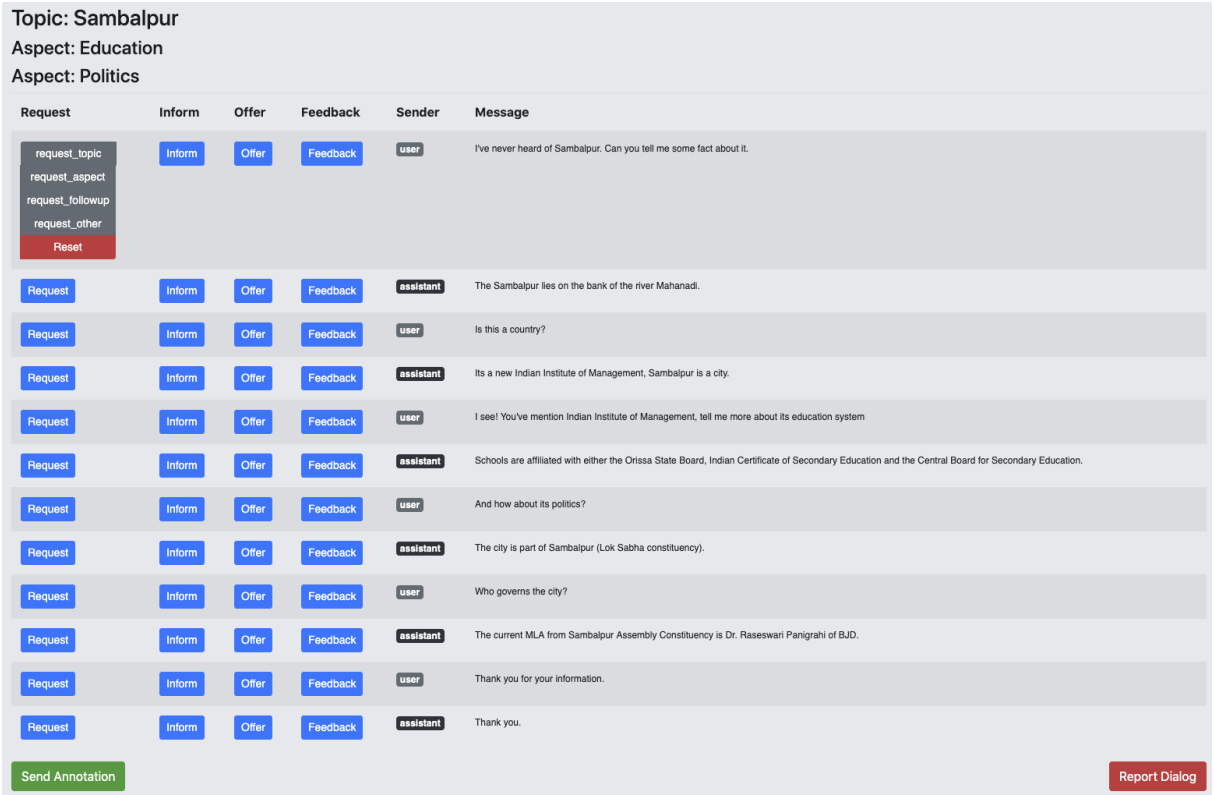

Figure 9: To annotate dialog acts, we developed an interface that showed each utterance on a separate line. Annotators could assign multiple dialog acts to each utterance. To reduce cognitive load, we grouped dialog acts into categories and showed a dropdown when a button is clicked.