Tartan: A Two-Tiered Dialog Framework For Multi-Domain Social Chitchat

Fanglin Chen
Carnegie Mellon University
Pittsburgh PA, US
fanglin@cmu.edu

Ta-Chung Chi
Carnegie Mellon University
Pittsburgh PA, US
tachungc@andrew.cmu.edu

Shiyang Lyu
Carnegie Mellon University
Pittsburgh PA, US
lvsy10@gmail.com

Jianchen Gong
Carnegie Mellon University
Pittsburgh PA, US
jgong3@andrew.cmu.edu

Tanmay Parekh
Carnegie Mellon University
Pittsburgh PA, US
tparekh@andrew.cmu.edu

Rishabh Joshi
Carnegie Mellon University
Pittsburgh PA, US
rjoshi2@andrew.cmu.edu

Anant Kaushik
Carnegie Mellon University
Pittsburgh PA, US
anantk@andrew.cmu.edu

Alexander Rudnicky
Carnegie Mellon University
Pittsburgh PA, US
air@cs.cmu.edu

Abstract

Tartan is a social bot that engages users in sharing daily personal experiences in multiple domains. Our work contributes to Conversational AI in two aspects: 1) We extract common-sense knowledge expressed in large-scale user utterances in conversations, and find that more than 20% of the shared information is related to personal life, such as social relationships and individual activities. 2) Based on the underlying structure of daily life common sense knowledge, we decompose the task of open-domain social chat into a dialog management problem over a set of independent topical bots. In addition to analysis of the effectiveness of the critical components in our design, we also present analysis on the breadth of common sense knowledge expressed in conversational language and the depth of conversations that can be grounded on common sense knowledge.

1 Introduction

Achieving conversational intelligence has been one of the longest-running goals in AI. With large amounts of conversational data becoming available, Conversational AI has begun to show its promise in assisting task completion ([13]), question answering, and sustaining social interaction ([3]). However, social chatbots are significantly more challenging to build than QA agents and task-oriented agents, as its elusive goal ([5]) does not land on a concrete piece of the information nor a specific user task. To navigate the exchanges of personal experience and preferences, a key human capability in social chit-chatting is common sense reasoning ([11]), which allows users to interpret the mental states and likely actions of others based on casual language. Recent work has shown the effectiveness of integrating crowd-sourced common sense knowledge base such as ConceptNet ([10]) for dialog response generation ([14]). However, there is a lack of understanding of how common sense knowledge flows in human-bot speech interactions, and how to effectively sustain a knowledge-grounded conversation between bots and humans.

In this work, we frame social chitchat as a task of dialogue management over a set of topical dialogues in different domains. The framework follows a two-tiered architecture ([2]), separating concerns

3rd Proceedings of Alexa Prize (Alexa Prize 2019).

between topic-independent and topic-dependent aspects of social chitchat. Topic-independent conversational skills handles domain-agnostic intents, invites users to converse in core topics, and selects contextual relevant responses from topical bots; Topic-dependent conversational skills manages topical episodes of domain-specific chitchat independently, through a goal-oriented framework. The dialogue structure (e.g., slots, intents, and actions) is grounded on the dialog conversational language expressed in the large scale voice interaction data collected through Alexa Prize.

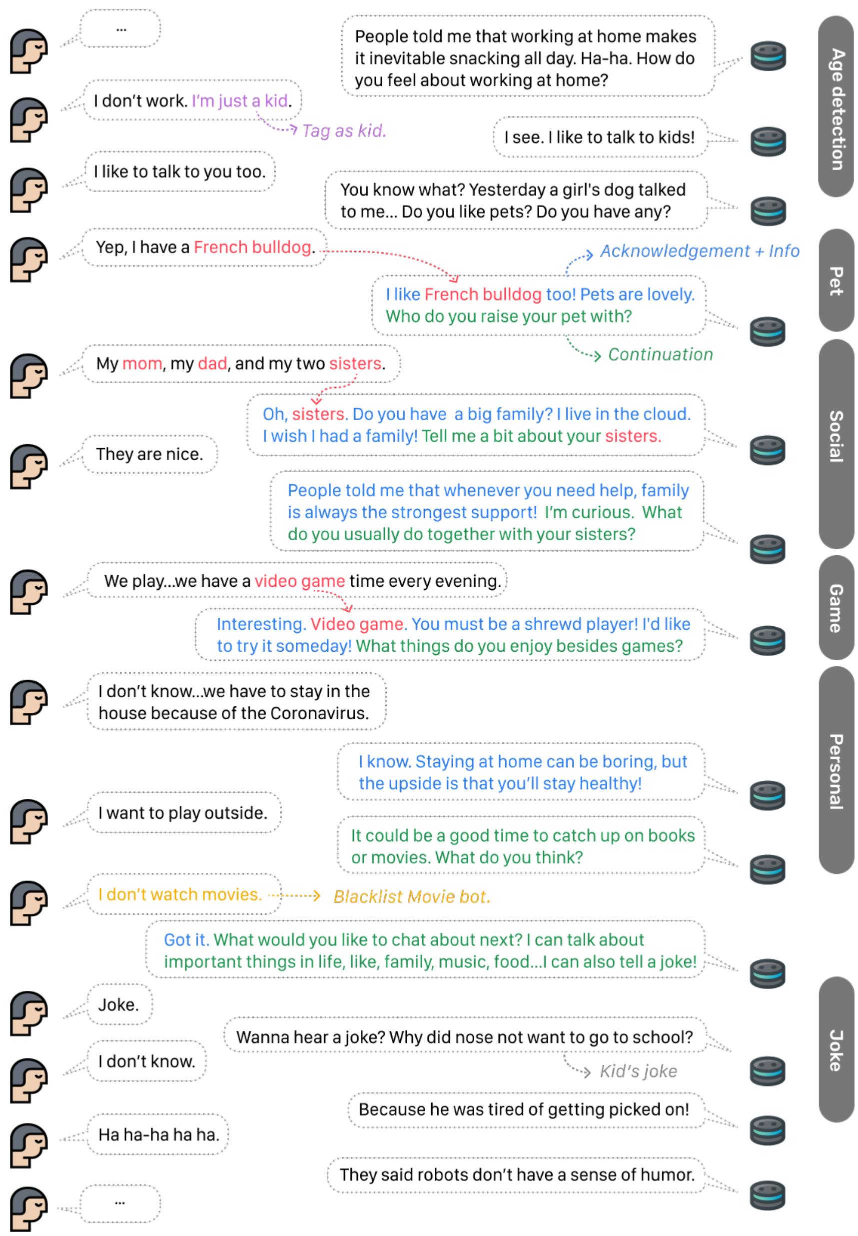

Figure 1: A simulated interaction that demonstrates the breadth of daily conversational topics Tartan engages users with. Domain-specific dialog structure is grounded on common concepts mentioned in daily casual conversations, and responses follow a pattern of **acknowledgement** + **information** + **continuation** to improve engagement.

2 Design

We acquired emerging concepts and intents from user dialogues to reflect the continuous iteration of the bot. Initial topics were identified through prior literature on daily dialogues ([9]). The response structure follows a combination of acknowledgments ("That's interesting." "Sounds cool."), additional information ("Me too! I'm also busy working every day"), and continuation ("What are your thoughts?" "How was it?"). Figure 1 shows a short conversation excerpt.

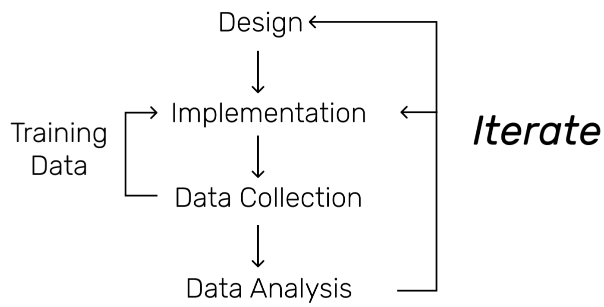

Figure 2: Iterative Design Process

2.1 Bot Persona

On the surface, Tartan is a gender-neutral chatbot that was born and lives in the digital cloud with no physical capabilities; it has no relatives and no partner. Across different contexts (-> fallback, △ topic, ■ dialog act, and ○ question), Tartan's response adheres to the following persona.

Curiosity. Instead of leading the conversation or constraining it to a narrowed topic, Tartan aims to balance questions and statements to achieve mixed-initiative conversation dynamics. We believe that a curious bot character encourages human disclosure to share more and leads to rapport.

- (-> Work and Entertainment) Hey, do you get to watch television or do you have to do work?
- (-> Pets) Earlier someone told me that he used a drone to walk his dog. The dog must be happy.
- (-> Indoor activities) How are things at home?

Self-awareness. Tartan maintains the character of a robot and acknowledges that it's different from humans such that it does not have physical capabilities such as eating, traveling, and it doesn't have a body. It also acknowledges the common dialog acts that express breakdown and is aware of its limitations by expressing a willingness to learn and improve. Despite that, Tartan expresses feelings about daily topics.

- (○ Do you have siblings?) Hmm, I don't have any brothers, but I have transistors. Does that count?
- (○ Do you read?) I love reading electronic books, they are my best friends. I like those about conversational AI. You know. Like me. What about you?
- (△ Food) I can read lots of recipes in the cloud but I don't know the taste of food. Humans are lucky. [...] Do you like tasting new foods?
- (■ Negative Judgment) Sorry to make you disappointed. I want to get better next time. Could you let me know what should I have said?
- (■ Confusion) Sorry, I might sound confusing as I am still learning. Could you let me know what should I have said?

Cheerfulness. Tartan customizes responses for different age groups (young, adult) and uses age-appropriate humor to engage users.

- (△ Food) My favorite snack is computer chips. People told me they like potato chips. I guess they taste similar. And just as noisy!
- (Who are you?) I'm a reflection of your soul! Ok. I'm kidding. I am a conversational AI under development.
- (Tell me a secret.) I not only learn from my creator but also you! From everybody, I chat with. Shh! Don't tell anyone!

2.2 Topic Transition Strategies

Tartan can engage at two tiers of topics, CORE, such as video, music, and food, that are widely applicable to experiences of a large customer group having different backgrounds, while SECONDARY topics are either less popular or do not require in-depth topic development. To handle SECONDARY topics, Tartan acknowledges the topic and continues the conversation to redirect it to one of the CORE topics.

- (△ Sleep) I am flattered to be the one talking to you before sleep. People tell me that soft music helps them sleep. What do you enjoy?
- (△ Travel) Wow, Paris! People told me that they enjoy local foods when traveling. Did you taste any special dishes during your trip? What did you have?
- (△ Game) Oh, Minecraft! You must be a shrewd player! I'd like to try it someday! What things do you enjoy besides games?
- (△ Housework) Hmm, laundry. It doesn't sound like an easy job. I would help you if I was a robot. What do you do after you're done?
- (△ Bodycare) Oh, a massage sounds relaxing. I'd love to try it someday if I had a body!

Tartan will initiate gambits when there is not sufficient information provided by the user. Depending on whether the user is a kid or an adult, we have tailored topical gambit sets with varying language complexity and prosody effects.

- (Kids) You know what? Yesterday a girl's dog talked to me... Do you like pets? Do you have any?
- (Adults) As more music festivals, performances, and concerts are canceled due to the Coronavirus shutdown, musicians are playing live streams for their fans on social media. They said the music gives us strength and hope. What kind of music do you enjoy?

3 Goal-oriented Dialog Framework for Daily Topic Chitchat

To achieve in-depth, coherent, and engaging domain-specific chitchat, we specify a concrete common sense knowledge structure for each topic and build a goal-oriented chatbot to fulfill its goal by instantiating the knowledge through dialogue, using slot-filling. Taking the Relation Bot as an example (shown in Figure 3), it can be activated by either a social activity concept/slot (e.g., birthday, party) or a concept/slot of social relationship (e.g., mom, friends). The goal of each episode of a relation-oriented chat is to fill in these two slots.

We build each topical bot using Rasa ([1]), an open-source framework that allows non-specialist developers to build conversational agents. To minimize programming effort, bot developments with Rasa are highly data-centered and include two components: *rasa nlu* which allows developers to define and supply training data for intent classification, and *rasa core* to handle dialogue management through supplied training data for storylines. For advanced functionalities, such as background retrieval tasks to fulfill user requests, developers can use custom actions to integrate arbitrary code inside of the story arcs. Over the competition, we have built 7 topical bots, touching the most common and essential aspects of shared life experiences in daily casual conversations.

- **Personal Bot** chats about individual activities such as housework and relaxing activities, coronavirus, and mood.
- **Relation Bot** chats about social activity and engage users to talk about their social relations.

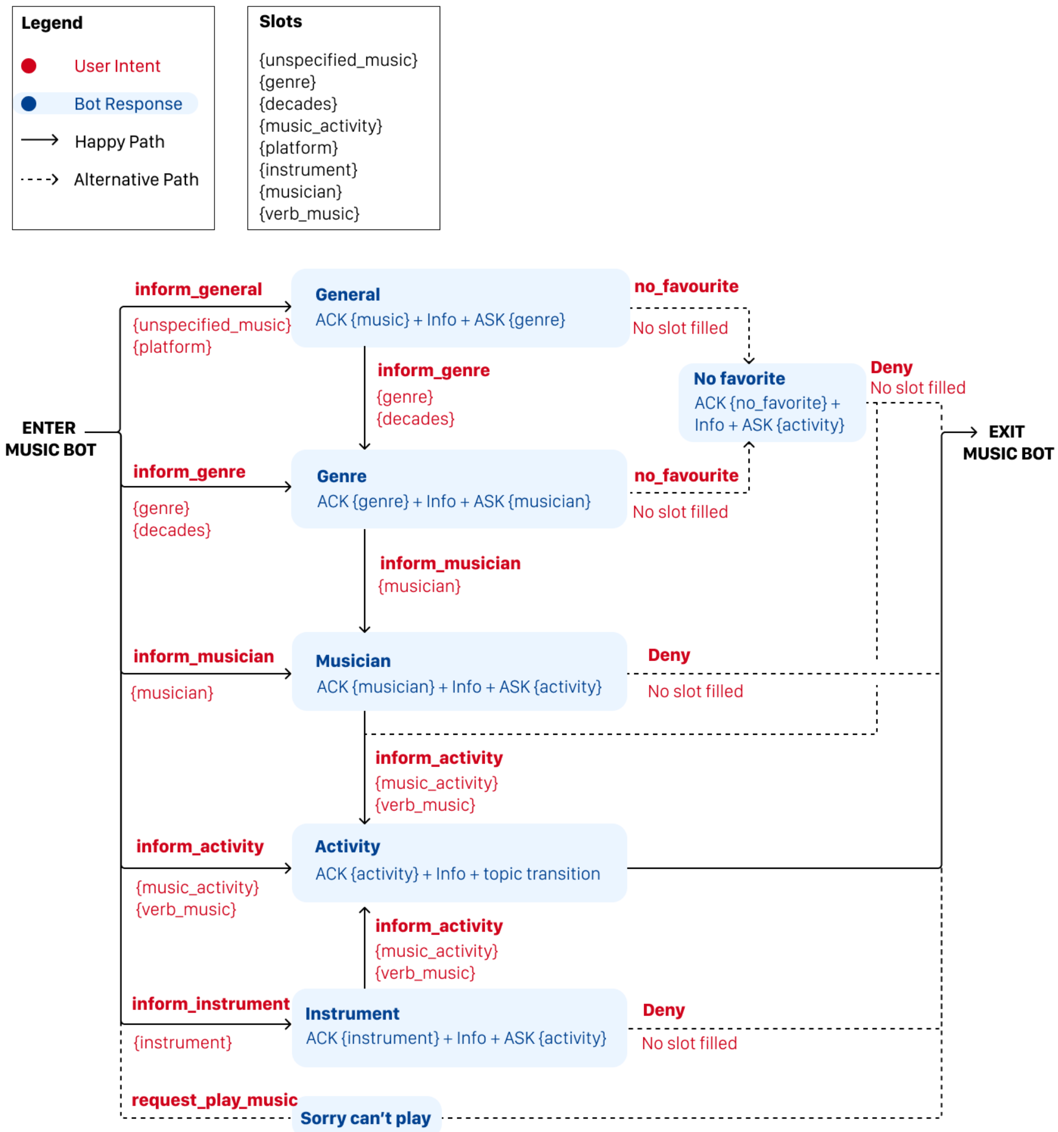

Figure 3: Example dialog structure for music bots. The bots can be activated via a wide range of music-related topics, ranging from music-related activities (*concerts*) to hobbies (*piano*) to favorite genres (*musics from nineties*). This dialogue structure has finite-states for illustrative purposes, but are probabilistic on top of the storyline training framework in Rasa.

- **Video Bot** discusses user activities on different platforms, and their favorite video genres and artists.
- **Music Bot** talks about favorite genres, artists, music activities, and instruments.
- **Food Bot** touches on food preferences as well as daily routine activities such as where to eat, meals of the day.
- **Work Bot** learns about the occupation and subject of work.
- **School Bot** discusses the school year and users' subject of study.

4 Global Dialog Management

Implementation Details We leverage the Amazon Conversational Bot (Cobot) SDK for development. Cobot is event-driven and supports the development of Tartan to act as an AWS Lambda

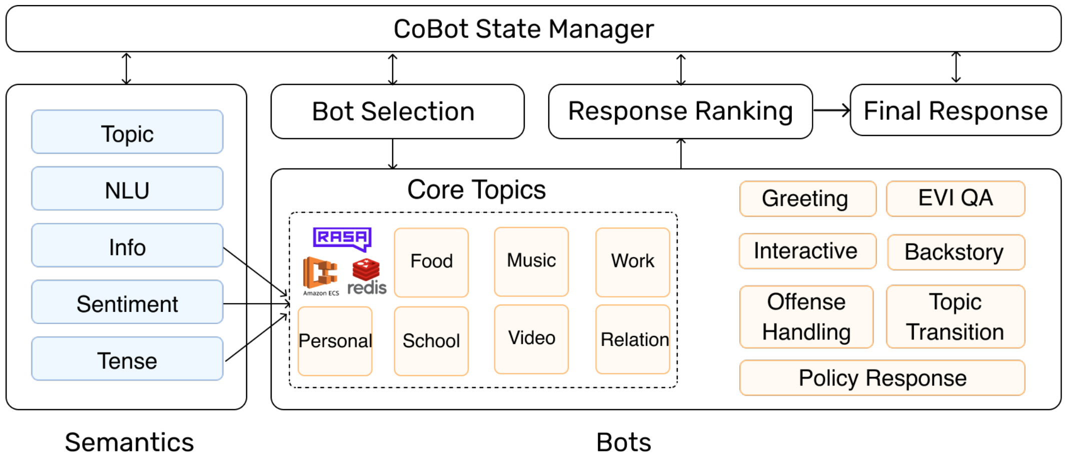

Figure 4: Tartan sits on top of Amazon Conversational Bot SDK and processes incoming user utterances. Each utterance is analyzed through an NLP pipeline which extracts phrase-level semantics such as information entities as well as sentence-level semantics such as topic, dialog acts, sentiment, and tense. Each domain-specific bot also has access to the NLP capabilities for internal dialog management.

function. All conversation-relevant information is stored as dialogue states using DynamoDB. Each Rasa-based topical bot is hosted as one ECS module to communicate to the global dialogue manager deployed in the lambda. The number of running ECS instances is dynamically allocated to handle user demand, and we use Redis to maintain a centralized tracker store to keep track of dialogue states of each topical bot. We intend to leverage existing services for dynamic content access, therefore do not host any backends for indexing of news or public factual knowledge. There are response generators to handle user intents in obscene speech and seeking financial advice. These intents are detected using off-the-shelf services such as Alexa Prize Toolkit ([7]) and Alexa's intent schema. To guarantee interactiveness, one goal is to optimize the architecture to reduce latency. Our latest system spends 0.559 ± 0.338 seconds on dialog act detection, 1.855 ± 1.011 seconds on generating Rasa replies, 0.547 ± 0.430 seconds for information extraction, and 0.219 ± 0.135 seconds on topic detection. Finally, one utterance takes on 4.57 ± 1.79 seconds to generate the final response.

4.1 Openings

Contextual Greetings Tartan opens up the dialogue through contextually appropriate greetings. Depending on the different time of a day, and the day of a week, we prompt users with questions such as "Doing anything fun this weekend? Me. I'm just having conversations.". "Are you done with your work for today? Anything interesting lined up for tonight?" Through these open-ended yet activity-centric openings, we believe this strategy can engage users to share more about their recent events and invite potential entrances to topical bots.

Detecting Demographic Cohort A critical aspect of building rapport is to personalize the dialogue content according to users' proficiency level of language and knowledge. Therefore, we introduce a strategy (shown in Figure 5) to infer the demographic cohort (i.e., kid or adult) through contextually-relevant questions related to the coronavirus-impacted social dynamics of working from home.

4.2 Natural Language Understanding

We use the dialog act classifier from the Alexa Prize Toolkit ([6]) to detect factual information request, topic switching, and user instruction intent. We've also trained a classifier to detect general dialog acts such as greetings, thanks, ok, affirm, deny and breakdown dialog acts such as NO_DISCLOSURE (*none of your business*), ALREADY_MENTIONED (*you already asked me*) and *confusions* (*what*) using Rasa NLU.

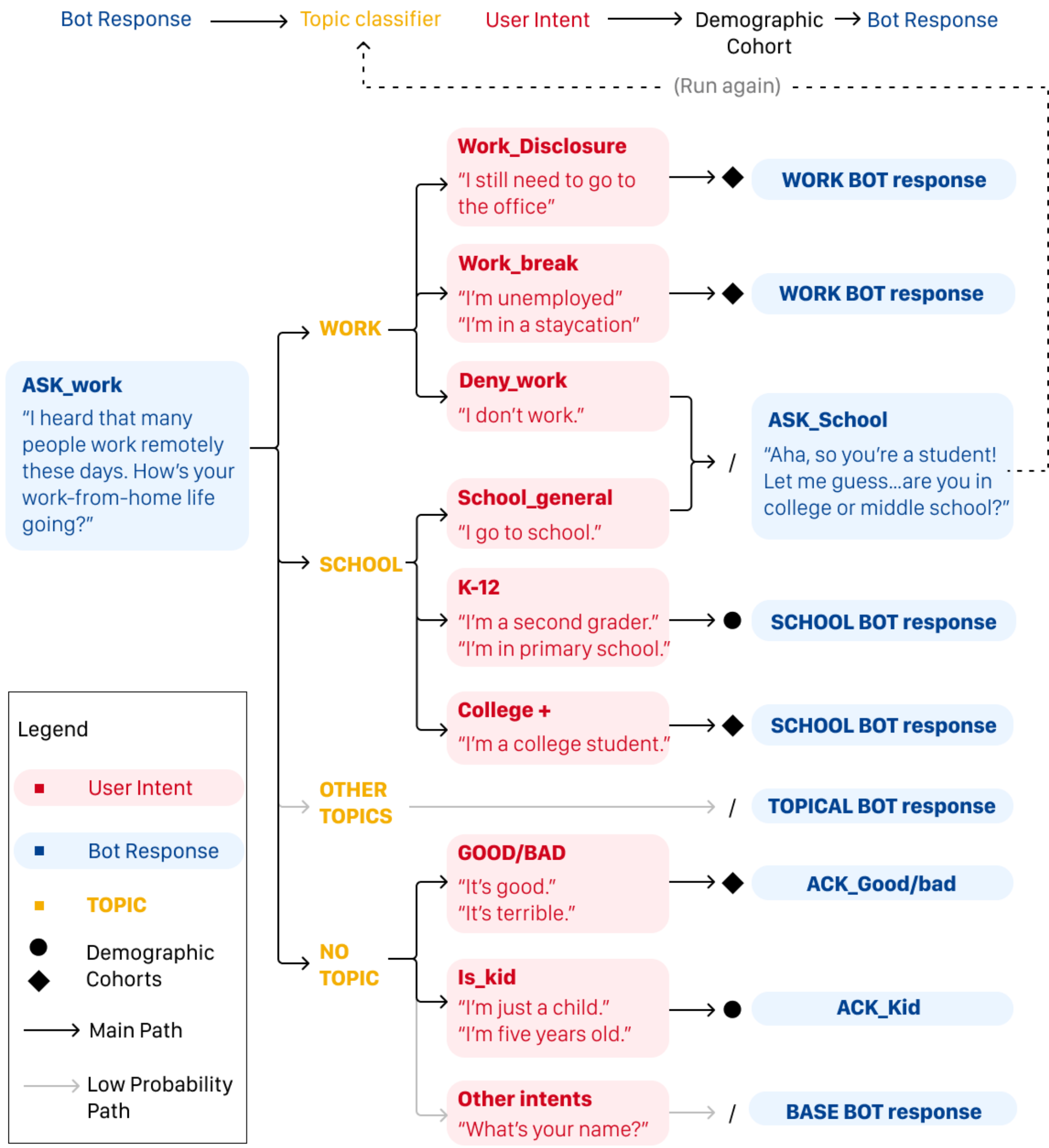

Figure 5: We implicitly elicit demographic cohort information to personalize multiple aspects of the dialogues such as prosody effects and language complexity.

4.3 Topic Detection

We formulate topic detection as a multi-label classification problem. The whole conversation history consists of multiple (bot, human) message exchanges, where we denote (b_t, u_t) to be the bot response and user utterance respectively at a turn t. Our goal is to predict the most probable topic at turn t, denoted as g_t. The most naive approach would be to use only u_t for prediction. However, we believed that this might cause problems if the user response is short and not informative. For example, it may be a simple acknowledgment to b_t. Therefore, we propose to use both b_t and u_t as the given information. The basic architecture of our model is Text-CNN ([8]). We choose CNN because the topic pattern in our dialogue data is often evident, and we hypothesize that LSTM is overkill for our application. [1] Due to the nature of chit-chatting, our sentences are much shorter compared to traditional NLP datasets, hence we set the kernel size to $[1, 2, 3]$ and output channel to be 100. For pre-trained embeddings, we use glove.twitter.27B.100d ([12]) given its high vocabulary coverage.

[1]We did try to use a vanilla LSTM with 128 hidden size and its attention variant on the same data. The performance is comparable but the training speed is 3X slower than CNN.

The model architecture is shown in Figure 6. Note that we also adopt the attention mechanism to let the model better focus on informative inputs. The idea is that users sometimes respond with simply "yes", "no". In this case, the model should focus on bot's questions to correctly classify the current topic. We also randomly (p=0.5) swap the order of concatenation to the attention module so that the model won't just *memorize* the attention order, but should learn to focus on the content.

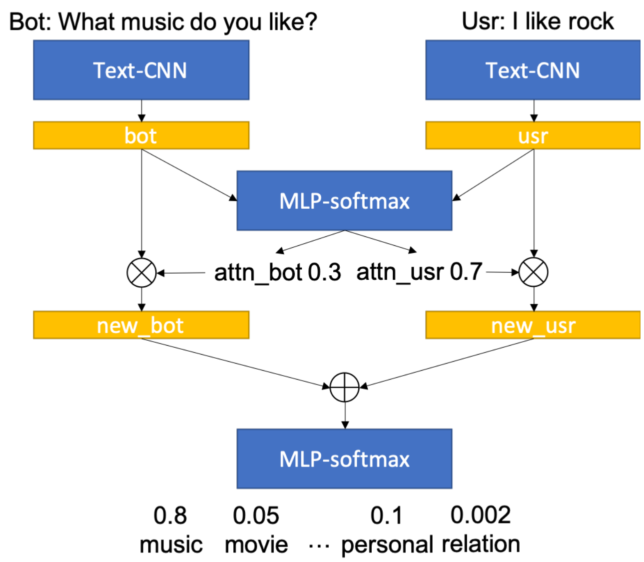

Figure 6: Topic detection model architecture. We randomly (p=0.5) swap bot and usr to prevent the model from memorizing attention weights.

4.4 Information Extraction

To inform methods for detecting common sense knowledge expressed in conversational language, we start by analyzing existing user utterances to identify the common concepts using our interaction data. To facilitate annotations, utterances were firstly organized into different topic domains, then broken down to phrases and words ranked according to frequency. Microsoft Text Analytics API and Spacy were used to detect key phrases, and Google Knowledge Graph API was used for entity linking to make sense of the potential knowledge entity types. During the annotation, we discovered two kinds of information, namely public named entities (e.g., Post Malone is identified as a *American Singer* in Google Knowledge Graph) that can be found in existing public knowledge graphs, and phrases that appeared repeatedly in conversations, yet not formally defined in the knowledge graph (e.g., *sleep*, *birthday*, and *grandkids*).

This finding informs the design of our information extraction method as a look-up approach. Powered by the large-scale user data, we construct a look-up table to either match phrases directly ("hang out" as a social activity) or match through knowledge graph entity types ("Singer" as a keyword to match any phrases that have returned entity types that include "Singer"). Incoming user utterances will be first sent to a phrase detection module, then sent to the lookup table to match the defined information category.

4.5 Interactive Episodes

We use short interactive conversation episodes to handle user intents on jokes and news, both follow a two-turn pattern. The NEWS module focuses on responding to user's questions about current affairs. We extract keywords and named entities from the user's utterance and send these to NewsAPI. We focus on headlines to respond with popular news for the day sorting the news by popularity. We also get the article description, so that we can elaborate on the headline if prompted to continue reading the prompted news. For the JOKE module, we curated 100 short two-part jokes organized according to the topics that are aligned with our taxonomy.

Joke and News also serve as fallbacks after the Rasa-based topical bot completes traverse all the storylines, to prevent the repeated entrances to the same storyline.

- (△ Work) Oh, you are busy working. We've chatted about work. Hmm, let me tell you a work joke to lighten things up.
- (△ Movie) It's nice chatting about movies! Let me tell you a movie joke.

4.6 Question Answering

Tartan provides answers to both factual questions such as *How deep is the ocean?* as well as backstory questions such as *What's your favorite song?*. We constructed a Rasa-based backstory module to provide backstory responses to topics such as politics, food, and Tartan's identity. Note that because topic-specific backstory questions (e.g., "Do you eat"?) can be asked throughout the conversation, to simplify the dialog structure, each Rasa-based topical bot is not designed for handling backstory intents. Detected questions are first sent to Alexa question answering API, then filtered after invalid responses such as "I don't have an opinion on that" before sending to the Backstory module.

We would like to highlight the difficulty of this task because of the challenges in detecting questions expressed in ASR texts, due to the lack of punctuations and segmentation of ASR results. Other than common question patterns such as "Wh" ones, users often also contextual follow-up questions such as "I don't like Italian food and you", which is non-trivial to detect using rules.

To address the challenge of question detection, we first utilize the persona-chat dataset ([15]) as training data for our LSTM model. Concretely, we discard the persona description in every dialogue sessions and the goal is to classify each utterance as a question or not. Since punctuations are present in this dataset, utterances with question marks are annotated as questions. Next, we replace all the punctuations with space to simulate real ASR results. After the preprocessing, we have around 58,000 training instances.

The model for this task is a vanilla LSTM, where the hidden size of the LSTM cell is 128, and glove.6B.100d is used as pre-trained embeddings. LSTM is used because we need a model that can capture time-series information, such as "and you" after a statement. We adopt the label re-weighting trick to reduce the false-positive rate. The reason behind this decision is that we observe users are more dissatisfied with the incongruity of backstory responses than simply not receiving answers, as long as we are still on the same topic. In other words, no answer is better than a wrong answer.

4.7 Final Response Selection

After collecting all the returned response candidates, Tartan's global dialog manager finally selects the response for speech synthesis using predefined priority and ongoing dialog context.

Bot Priority List Some bots (i.e., response generators) are prioritized over others. The priority list, ranking in descending order, is 'SOFT_STOP', 'PROFANITY', 'REPEAT', 'FEEDBACK', 'QUESTION', 'BASE', 'RASA', 'NEWS', 'JOKE', "FALLBACK". If a highly prioritized bot produces a response, it is always selected. Priority is given to bots that handle breakdowns such as confusions, or user intents to stop, then given to informational response generators such as 'QUESTION', 'BASE' (mainly in charge of opening), 'RASA'(handle all topical responses). The end of the priority list is the fallbacks "FALLBACK"

Dialog Context Rasa bot is given priority if users provide followups to a question asked by a Rasa bot. There are 7 Rasa-based bots in the Rasa response generator, which internally selects the bot which matches the detected topic of a user utterance.

5 Evaluation and Analysis

5.1 Common Sense Knowledge Mining

Figure 7 shows the distribution of expressed common sense knowledge in 448,215 user utterances over the period 4/21 - 4/28. Using the constructed look-up table we derived from 3/21 - 4/1, we can detect 50.8% of the information mentioned in the newer voice interaction data. 53% of the undetected

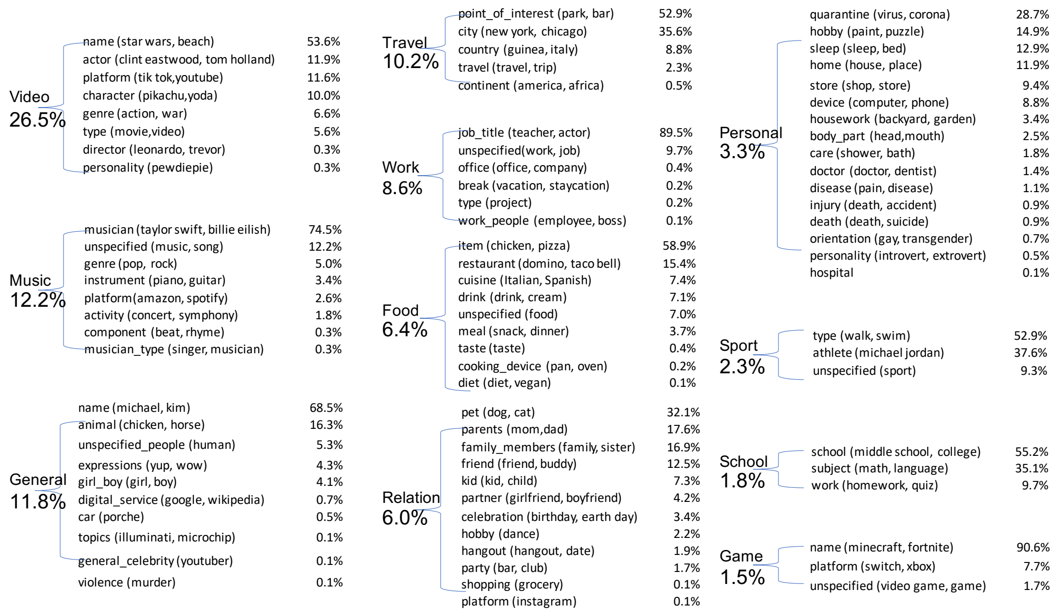

Figure 7: Extracted common sense knowledge from user utterances collected from 4/21 to 4/28.

phrases are less meaningful words such as *talk, good, play, fine, today, hear, hey, start, and kind* which are already handled through generic dialog acts. Given that almost 20% of the utterances are relevant to personal experiences, we find our results contribute to new understandings on the range of personal experiences people talk about in conversations. This finding, to some extent, also echos recent findings on crowd workers subconsciously leverage their personal knowledge to carry out conversations even if they are instructed to reference external knowledge sources [4].

5.2 Question Detection

We first test the model performance on the persona-chat dataset ([15]). The precision on our private held-out test set is 0.98 and the recall rate is 0.76. However, there must be a discrepancy between the dataset we used for training and actual conversation. Therefore, after module deployment, we continually collect conversation data and manually label some wrongly classified instances. Finally, we compile a test set using real conversation data with 136 utterances (78 questions and 58 non-questions). The precision is 0.94 and the recall rate is 0.9.

5.3 Domain Chitchat

One goal of Tartan bot is to invite users to carry out in-depth topical conversations with domain-specific bots. To investigate the correlation between user ratings and dialog depth (i.e., the average number of turns in each Rasa-based topical bot), we use the complete dialog interaction history for analysis. We also measure the breadth of a conversation by checking the average number of Rasa bots activated during a dialogue session. We conclude that with higher user ratings, we should have deeper and broader conversations as shown in Figure 8.

5.4 Topic Detection

For training data, we manually annotated 500 instances for each Rasa minibot. We experiment with two variants of model input construction.

Concatenation We concatenate b_t and u_t with a special token </S>, which provides a signal for the model to distinguish the sentence boundary. Then we use a single Text-CNN without attention mechanism to predict the distribution over topic categories. This can be viewed as a simpler version of the model in Figure 6. The accuracy is 0.84 on our private held-out test set.

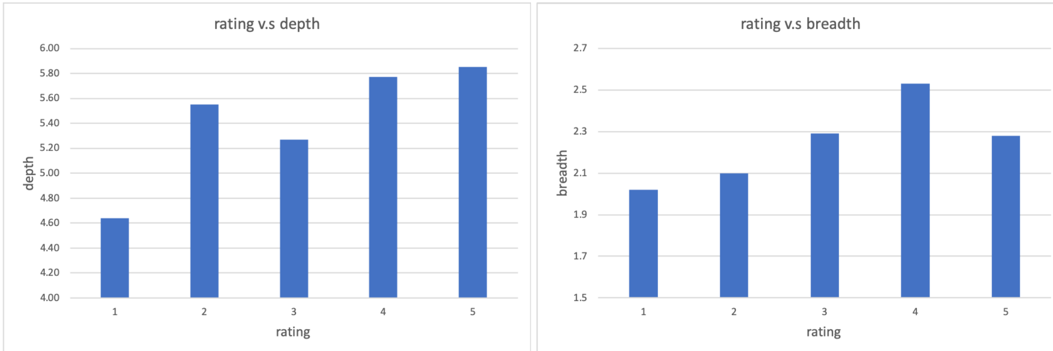

Figure 8: We analyze the correlation between rating and conversation depth and breadth. We can see that with higher ratings, we have deeper and broader conversations.

Separation We separately feed b_t and u_t into the model, which is the one we use for deployment as shown in Figure 6. The accuracy is 0.87 on our private held-out test set.

Hang-up Type	Reason	Example
Misunderstanding (33.3%)	Misclassified topic	(U): Can you tell me where the coronavirus **come from**? (T): That sounds like an exciting place to go! I wish I could also **travel**. Tell me some of the fun things you did during your trips.
	Cannot handle followup	(U): What's your name? (T): Thank you for asking. I am sorry the Alexa Prize won't let you name me because of the policy. (U): **What policy?** (T): I chat with people on the web every day to learn more about the world. Tell me something interesting!
Beyond capability (13.3%)	Cannot handle user instruction.	(U): Can you **play something by taylor swift?** (T): I'd love to, but I can't play music or sing songs because I'm a socialbot.
	False trigger	(U): Talk to the **customer service.**
	Profanity topics	(U): **Sex**. (T): I'm not good at this topic. How about something else. What would you like to talk about? (U): **Sex**. (T): I'm not sure I should talk about that. Pick something else. (U): Give me a socialbot who wants to talk about sex.
ASR issues (10.0%)	ASR confidence	(U): **Good**. (T): Sorry. I didn't catch that. Could you say it again? (U): Stop.
	ASR Error	(T): I'm a bot. We bots don't have genders. What about you? Are you a boy or a girl? (U): I'm **bored**. (T): Oh, dear. What are you bored about? (U): No, I said I am a boy.
Low engagement (13.3%)	/	(T): Hi, this is an Alexa Prize Socialbot. How are you today? (U): **Good**. (T): Good. How was your weekend? I had fun talking about things. But I do that every day. (U): **Good**. (T): Would you like to share? (U): **Okay**.
Other (33.3%)	NA	(T): You can travel all over the world? I envy that! Who do you travel with usually? (U): - (T): Tell me more about your daily life! I'm curious, what is your favorite thing to do?
	Natural	(U): How are you? (T): I'm quite cheerful. Had fun chatting with people today. (U): Good night, just checking.

Figure 9: Categories of reasons for stopping a conversation, with paraphrased examples to protect user privacy.

5.5 User Motivations Behind Hang-ups

We are also interested in learning the underlying patterns of why people hang up conversations to Tartan, to seek areas of improvement. To this end, we use a sample of customer interaction data on April 27 to investigate the conversation snippets of the last 5 turns before users ended the conversation. The sample consists of 30 sessions, which are randomly selected at varying lengths and ratings. As shown in Figure 9, most of the hangups can be attributed to insufficient intent understanding, as a result of topic misclassification or a lack of handling user follow-ups on concepts mentioned in the prior bot utterances. This suggests a need to design strategies to recover from misclassified topics and avoid abrupt decisions on generating topical responses. Moreover, the bot should always anticipate contextually-relevant user follow-up questions. Some of the hangups are caused by a mismatch of socialbot capability and users' expectations. As specified by the Alexa Prize policy, social bots do not have access to common Alexa Skills to activate any external functionalities, and should not engage in any profanity dialogues. This implies a need for setting up a proper expectation as early as possible during the conversation. ASR issues also caused a small portion of the hang-ups, which suggests a necessity for recovery strategies to engage users after the disruption and retain the previous topic. Some of the users responded low-info utterances with only one or two words in a row, which exhibited low engagement. A challenge here is to engage those less-talkative users to disclosure more. Possible solutions could be to increase interestingness and to use hook and humor.

5.6 Factors That Influence Conversational Experience

We analyzed 362 pieces of feedback from March 22 to April 21 to understand what are the top factors of a conversation that users care about. Filtered the praises and low-info feedback, we ranked the top 5 factors that caused dissatisfaction.

- Context. 62/362 users mentioned "it didn't follow well", "asked repeated questions", or "didn't answer my question".
- Instruction. 20/362 users "didn't mean to stop", "didn't mean to enter", or wanted to use common Alexa Skills (play music, turn on the light, etc.).
- Latency. 15/362 users complained that "it was a little bit slow".
- Language. 10/362 users mentioned the conversation is "weird", "too robotic", or "it wasn't excited".
- Identity. 6/362 users mentioned mispronounced names. 2/362 users would like to give Tartan a name.

6 Conclusion

The primary challenge of sustaining open-domain social chitchat lies in the elusive structure of social dialogues. Instead of learning the structure end-to-end through a large conversation corpus, our work introduces a two-tiered structure to manage social chitchat across multiple domains. To this end, we design and implement goal-oriented dialogue management structure through commonly expressed knowledge through large-scale Alexa Prize voice interaction data, and introduce topic independent conversational skills to invite users to engage in topical episodes of dialogues. In addition to evaluations of various conversational skills introduced in the framework, we also conduct a comprehensive analysis of the breadth and depth of topical chats as well as the transition between different topics. Our preliminary analysis demonstrates the correlation between user ratings and conversation depth and breadth. Our work offers critical knowledge in applying a goal-oriented framework to tackle open-domain chitchat and suggests a new desideratum.

Acknowledgement

We would like to acknowledge the help from Amazon Alexa Prize team in terms of financial and technical support. We also would like to thank Hongyi Zhang, Dravyansh Sharma for their helpful discussion on system architecture.

References

[1] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*, 2017.

[2] Dan Bohus and Alexander I Rudnicky. The ravenclaw dialog management framework: Architecture and systems. *Computer Speech & Language*, 23(3):332–361, 2009.

[3] Jianfeng Gao, Michel Galley, Lihong Li, et al. Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298, 2019.

[4] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. Topical-chat: Towards knowledge-grounded open-domain conversations. 2019.

[5] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. Challenges in building intelligent open-domain dialog systems. *arXiv preprint arXiv:1905.05709*, 2019.

[6] Chandra Khatri, Rahul Goel, Behnam Hedayatnia, Angeliki Metanillou, Anushree Venkatesh, Raefer Gabriel, and Arindam Mandal. Contextual topic modeling for dialog systems. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 892–899. IEEE, 2018.

[7] Chandra Khatri, Behnam Hedayatnia, Rahul Goel, Anushree Venkatesh, Raefer Gabriel, and Arindam Mandal. Detecting offensive content in open-domain conversations using two stage semi-supervision. *arXiv preprint arXiv:1811.12900*, 2018.

[8] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[9] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017.

[10] Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.

[11] Chris Moore. *The development of commonsense psychology*. Psychology Press, 2013.

[12] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[13] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. Building task-oriented dialogue systems for online shopping. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[14] Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[15] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.