Commonsense Reasoning for Conversational AI: A Survey of the State of the Art

Christopher Richardson [1] **Larry Heck** [1]
[1]School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30308, USA
crichardson8@gatech.edu, larryheck@gatech.edu

Abstract

Large, transformer-based pretrained language models like BERT, GPT, and T5 have demonstrated a deep understanding of contextual semantics and language syntax. Their success has enabled significant advances in conversational AI, including the development of open-dialogue systems capable of coherent, salient conversations which can answer questions, chat casually, and complete tasks. However, state-of-the-art models still struggle with tasks that involve higher levels of reasoning - including commonsense reasoning that humans find trivial. This paper presents a survey of recent conversational AI research focused on commonsense reasoning. The paper lists relevant training datasets and describes the primary approaches to include commonsense in conversational AI. The paper also discusses benchmarks used for evaluating commonsense in conversational AI problems. Finally, the paper presents preliminary observations of the limited commonsense capabilities of two state-of-the-art open dialogue models, BlenderBot3 and LaMDA, and its negative effect on natural interactions. These observations further motivate research on commonsense reasoning in conversational AI.

Introduction

Commonsense reasoning has recently become a major focus of research in natural language processing (NLP). The importance of commonsense in machine intelligence has been known for a long time, but the problem of instilling commonsense knowledge into AI technology remains unsolved (Storks, Gao, and Chai 2019). Commonsense knowledge is generally understood as external knowledge about the world that all humans are assumed to possess (Liu and Singh 2004). Knowing something like "a car cannot be in two places at once" may be taken for granted by humans, but an AI may have no explicit basis for such knowledge. Despite attempts at standardized categorizations (Ilievski et al. (2021), Gordon and Hobbs (2017)), there currently exists no universally agreed upon scheme for codifying commonsense knowledge. How to represent commonsense knowledge and perform reasoning over that knowledge in NLP is an active area of research.

This survey will review the intersection between commonsense reasoning and conversational AI. The implications for conversational AI are potentially far reaching. In a detailed review of neural approaches in conversational AI, Gao et al. cites "*reasoning* in the neural space" as one of the three primary steps in neural conversational AI. Gao specifically references *commonsense* knowledge, saying "Commonsense knowledge is crucial for any dialogue agents."" In this paper, we argue that there is still much work to be done in the area of commonsense reasoning in dialogue understanding, despite commonsense being one of the most researched types of reasoning for neural models in NLP.

Commonsense reasoning by definition requires external knowledge, which can come from various sources. Some amount of commonsense reasoning is captured during pretraining due to commonsense knowledge being implicitly present in the data. Explicit external knowledge sources, such as knowledge graphs (KGs), can also be used to improve commonsense reasoning, typically in conjunction with pretrained language models. Two of the most common in the commonsense literature are CONCEPTNET (Speer, Chin, and Havasi 2017) and ATOMIC (Sap et al. 2019a).

Recent research has also explored the use of neural networks (NN) to encode commonsense knowledge, usually based on the transformer architecture (Vaswani et al. 2017). COMET (Bosselut et al. 2019) is a transformer-based NN trained with commonsense knowledge from ConceptNet and ATOMIC, which is capable of generating novel commonsense knowledge. An improved version based on ATOMIC$^{20}_{20}$ was presented in Hwang et al. and called COMET-ATOMIC$^{20}_{20}$.

This paper surveys the existing literature in commonsense reasoning as it pertains to conversational AI and discusses the methods and metrics used. First, the most common problems encountered in conversational AI are described, along with a discussion of the relevance of commonsense reasoning for each problem. Next, the methods for addressing the commonsense problem are described and categorized into three categories: Model Fine-Tuning, Knowledge-Graph Grounding, and Natural Language Explanations. Major benchmarks are then described, including metrics used for evaluating commonsense. Lastly, some preliminary ob-

---

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Accepted to Workshop on Knowledge Augmented Methods for Natural Language Processing, in conjunction with AAAI 2023.

servations are made pertaining to two state-of-the-art conversational AI models: BlenderBot3 (Shuster et al. 2022) and LaMDA (Thoppilan et al. 2022).

Conversational AI Problems

In this section, we discuss four problems commonly encountered in the field of conversational AI: (1) Sequence Classification, (2) Question Answering, (3) Dialogue Modeling, and (4) Dialogue Summarization. For each category, we describe the problem, give examples of past work, cover the relevant datasets, and discuss the importance of commonsense reasoning pertaining to each problem. Table 1 lists the surveyed literature organized by these four conversational AI problems.

Sequence Classification

A successful conversational AI system should be able to converse with humans in a natural way, i.e. be able to identify intents, recognize emotions, detect conversation topics, etc. Many such skills fall under the task of *sequence classification* including slot filling (Mesnil et al. 2014), domain classification (Jaech, Heck, and Ostendorf 2016), intent detection (Siddique et al. 2021), emotion detection (Zhong, Wang, and Miao (2019), Balahur, Hermida, and Montoyo (2011), Ghosal et al. (2020)), sentence topic prediction (Ghosh et al. 2016), sequential dialogue context modeling (Bapna et al. 2017), and others. Commonsense reasoning is one of the many dimensions of understanding required across a variety of sequence classification tasks. This is especially the case for human-human dialogue data, where commonsense knowledge is often mutually understood only present in the data implicitly (Grice 1975). Recently, encoder models like BERT (Devlin et al. 2018) have been used to capture this implicit knowledge. To use these models for sequence classification, they are typically fine-tuned on data from the specific classification task.

Several dialogue datasets are available for sequence classification research. DailyDialogue (Li et al. 2017) consists of 13,118 dialogues scraped from websites that serve English language learners as practice for their English. The dialogues are mainly focused on every day life topics. Every utterance in the dialogue is annotated with one of four dialogue acts, as well as one of seven emotion classes. Another emotion classification dataset is EmoryNLP (Zahiri and Choi 2018), a multi-party dialogue corpus based on the television show *Friends*. The corpus contains 12,606 utterances, each annotated with one of seven emotions.

Question Answering

Question answering (QA) is one of the most common tasks explored in the NLP literature. The importance of QA in conversational AI is explored in depth by (Gao et al. 2019). Successful QA agents, particularly for multi-turn conversational QA, must reason over the past dialogue as well as external knowledge bases (including commonsense knowledge). In this section we discuss two variants of QA relevant to conversational AI: multi-choice QA (MCQA) over dialogues, and multi-turn conversational QA (CQA).

MCQA is the problem of answering multi-choice questions given a dialogue as context. An example can be found in (Ghosal et al. 2021), where commonsense-focused questions were created from the CIDER dialogue dataset. Several other datasets for use in MCQA include MuTual (Cui et al. 2020), which consists of 8,860 dialogues based on English listening examinations taken by students in China; DREAM (Sun et al. 2019), another dataset based on English examinations for Chinese students, specifically curated to have an emphasis on reasoning; and the Ubuntu Dialogue Corpus (Lowe et al. 2015), consisting of 1 million multi-turn dialogues from Ubuntu chat logs, taken from 2004-2015. Many of the logs are in the forms of QA, which can readily be used for MCQA as in Zhang, Li, and Zhao (2021).

CQA is the problem of asking and/or answering questions in a multi-turn conversational format. Two of the most popular CQA datasets are CoQA (Reddy, Chen, and Manning 2019) and QuAC Choi et al. (2018). Both consists of over 100k questions and answers, with CoQA conversations being about passsages taken from a diverse set of domains, and QuAC conversations concerning Wikipedia articles.

Dialogue Modeling

Dialogue modeling is similar to the classic NLP task of language modeling, but applied to dialogue turns instead of individual words. Dialogue modeling is used for both open-dialogue (chatbots), and task-oriented dialogue. There are many datasets used for dialogue modeling, including PERSONA-CHAT (Zhang et al. 2018), ConvAI2 (Dinan et al. 2020), DailyDialogue (Li et al. 2017), and MultiWOZ (Budzianowski et al. 2018). Older dialogue modeling systems typically had modular architectures with explicit components for natural language understanding (NLU), dialogue state tracking, and natural language generation (NLG) (Chen et al. 2017). Recent research has focused on end-to-end dialogue systems driven by large neural models (Ni et al. 2022). These systems are typically built on a core sequence-to-sequence model like GPT, T5, or BART (Lewis et al. 2019), which take user utterances as input and generate responses directly as output. That model can then be augmented in various ways, for example with knowledge graph grounding or response candidate re-ranking from a separately-trained scoring model. Some examples of research covering dialogue modeling with commonsense are Zhou et al. (2021a), Zhou et al. (2021b), and Majumder et al. (2020). While these approaches represent significant advances in recent years, they still lack the ability to perform commonsense reasoning. The addition of commonsense will be necessary to advance to the next level of human-like interactions. Case studies demonstrating the lack of commonsense in current dialogue models are explored later in this paper.

Dialogue Summarization

Dialogue summarization is the task of generating a concise summary of a dialogue while retaining factual consistency. Summarization is particularly important in conversational AI systems used for meetings. Virtual assistants that automatically create meeting summaries enhance productivity by enabling efficient recall of the key points and action

	Sequence Classification	Question Answering	Dialogue Modeling	Dialogue Summarization
Young et al. (2018)			✓	
Zhong, Wang, and Miao (2019) - KET	✓			
Ghosal et al. (2020) - COSMIC			✓	
Majumder et al. (2020) - COMPAC			✓	
Arabshahi et al. (2021a) - CLUE			✓	
Feng, Feng, and Qin (2021a) - D-HGN				✓
Ghosal et al. (2021) - CIDER	✓	✓		
Li et al. (2021) - DialogInfer	✓			
Qin et al. (2021) - TimeDial	✓	✓	✓	
Zhang, Li, and Zhao (2021) - PoDS		✓		
Zhou et al. (2021a)			✓	
Zhou et al. (2021b) - TBS			✓	
Zhou et al. (2021c) - CEDAR			✓	
Arabshahi et al. (2021b) - CORGI		✓		
Ma et al. (2021)		✓		
Ghosal et al. (2022) - CICERO	✓		✓	
Li et al. (2022) - KEC	✓			
Sabour, Zheng, and Huang (2022) - CEM			✓	
Tu et al. (2022) - Sentic GAT	✓			
Varshney, Prabhakar, and Ekbal (2022) - CNTF			✓	
Xie, Sun, and Ji (2022) - CKE-Net	✓			
Xu et al. (2022) - DMKCM			✓	
Kim et al. (2022b) - SICK				✓
Wu et al. (2020) - ConKADI			✓	

Table 1: Summary of Tasks

items from the meeting. One of the earliest attempts at this was the CALO Meeting Assistant System (Tur et al. 2010). Given the nature of human-human conversations, commonsense reasoning is often required to produce accurate, and complete summaries. Major challenges in dialogue summarization include preserving salient facts, maintaining logical coherence, and avoiding hallucinations (Feng, Feng, and Qin 2021b).

Two of the earliest and largest meeting datasets used for dialogue summarization are the ICSI Meetings Corpus (Janin et al. 2003) and AMI Meeting Corpus (McCowan et al. 2005). The ICSI corpus contains audio and transcripts of 75 natural meetings between 53 unique speakers over 4 main topics recorded simultaneously with head-worn and table-top microphones. The AMI corpus contains 100 hours of multi-modal data from meetings taken from various recording instruments. The dataset includes both real and scenario-driven meetings. DialogSum (Chen et al. 2021) contains 13,460 dialogues taken mostly from DailyDialog, DREAM, and MuTual. SAMsum (Gliwa et al. 2019) contains 16,369 dialogues with accompanying summaries, all manually written by linguists fluent in English. The dialogues were designed to resemble text message conversations, a characteristic validated by a separate pair of linguists. The subject is open domain, and the conversations are curated to resemble real conversations, e.g. complete with typos, shorthand, and occasional slang. This makes the corpus well suited for research on commonsense reasoning since more informal and familar conversations tend to have more unstated facts that are mutually understood by the participants.

Feng, Feng, and Qin (2021a) explored the usage of ConceptNet for integrating commonsense into the dialogue summarization problem, and more recently Kim et al. (2022b) used COMET to generate gap-filling commonsense statements for augmenting summarization models. Despite these works, methods for integrating commonsense into dialogue summarization remain relatively understudied (Feng, Feng, and Qin 2021b).

Methods

This section reviews the various methods explored in past research for learning, utilizing, and evaluating commonsense reasoning in the context of conversational AI. Three categories of methods found in the commonsense literature are covered: model fine-tuning, knowledge graph grounding, and natural language explanations.

Model Fine-Tuning

The most common method in current research for addressing the commonsense problem is to create a custom dataset with annotations designed for learning commonsense. These datasets typically draw from larger dialogue datasets such as the Ubuntu Dialogue Corpus (Lowe et al. 2015), DailyDialogue (Li et al. 2017), MuTual (Cui et al. 2020), and DREAM (Sun et al. 2019).

CIDER (Ghosal et al. 2021) draws from DailyDialog, MuTual, and DREAM, and consists of annotations in the form of triplets that form commonsense explanations over the data (e.g., “missed the bus” *causes* “late”). The triplets’ typology is mostly based on ConceptNet relations. CICERO (Ghosal et al. 2022) is an extension of CIDER with human-written natural language inferences instead of triplets. TimeDial (Qin et al. 2021) introduces a multi-choice cloze task over DailyDialog samples with an emphasis on temporal commonsense reasoning. Zhou et al. (2021a) use ConceptNet to automatically filter dialogues from existing datasets based on the presence of commonsense assertions. They then collect additional data using Amazon Mechanical Turk (MTurk) based on prompts from SocialIQA (Sap et al. 2019b). Zhou et al. (2021c) continue the approach to filtering dialogues using ConceptNet, and add natural language explanations of responses in the dialogues in their dataset CEDAR. The explanations were generated by a text-to-text model and then verified by crowd workers. (Moon et al. 2019) collected a dataset of human-human dialogues using ParlAI (Miller et al. 2017) called OpenDialKG, with each dialogue being accompanied by KG entities annotated by the crowd workers. Ziems et al. (2022) introduced MIC, a dataset focused on commonsense moral/ethical reasoning based on a "Rules of Thumb" paradigm from Forbes et al. (2020). Kim et al. (2022a) built on this work and released a dataset, ProsocialDialogue, designed for training social bots to respond to safely and properly to unsafe dialogue utterances from users. Arabshahi et al. (2021b) introduced CORGI, an LSTM-based neuro-symbolic theorem prover that answers questions about unstated commonsense presumptions in dialogues, along with an dialogue dataset annotated with presumptions.

Knowledge Graph Grounding

While commonsense-focused datasets may provide a natural fine-tuning source as well as evaluation metrics, they do not offer a way to directly ground the conversation with commonsense knowledge. Doing so requires an integrated external knowledge source in the dialogue system. The most common source of this type of knowledge is a knowledge graphs (KG). Among the various KGs that appear in the literature, the most prominent are ConceptNet Liu and Singh (2004) and ATOMIC Sap et al. (2019a) for commonsense-focused research.

An early attempt to integrate a commonsense KG into conversational AI was by (Young et al. 2018), where knowledge from ConceptNet was used to augment a retrieval-based conversational model. The relevant commonsense knowledge for each message was recovered using a simple n-gram matching scheme. Ma et al. (2021) analyzed knowledge-graph grounding for zero-shot question answering using several combinations of KGs and language models. They generated synthetic questions and used a neuro-symbolic framework to investigate the connection between knowledge sources, question generation techniques, and model types. Zhong, Wang, and Miao (2019) extracted triplets from ConceptNet and embedded them alongside word embeddings to improve emotion detection. Moon et al. (2019) presented a graph traversal scheme trained to predict relevant KG entities based on a dialogue history, which was then used to re-rank candidate follow-ups in the dialogue. These pre-BERT works used LSTM models and word embedding approaches. With the introduction of BERT (Devlin et al. 2018), the focus has since shifted to integrating KGs with transformer-based pretrained language models. Zhang, Li, and Zhao (2021) extracted knowledge from ConceptNet (as well as two other KGs) for dialogue response selection with a BERT-based ranker, which they use to encode the dialogue history along with pertinent facts from the KG. The pertinent facts were found with a simple method that uses semantic matching and part-of-speech tagging. Tu et al. (2022) employed a similar method of knowledge grounding, but integrated the knowledge into a graph attention network (Veličković et al. 2017), which was used to augment an emotion classifier. Xie, Sun, and Ji (2022) used a graph attention network to cross reference knowledge from ConceptNet and integrated it into their dialogue emotion classifier. Feng, Feng, and Qin (2021a) used a graph network to encode the dialogue as well as knowledge from ConceptNet into a heterogeneous network, which was then used in conjunction with an LSTM to generate commonsense-informed summaries for the task of abstract summarization. Zhou et al. (2021b) used semantic matching and embedding similarity to find relevant triples from ConceptNet, and generated natural language knowledge from those triples to condition a response generation model. They also introduced three evaluation metrics for the generated knowledge. Wu et al. (2020) extracted ConceptNet triplets relevant to the dialogue context for response generation, and Varshney, Prabhakar, and Ekbal (2022) used a similar method but added coreference-resolution techniques for named-entity-aware grounding. Xu et al. (2022) extracted knowledge from both linked documents and ConceptNet triplets and fused both sources of knowledge together to enhance dialogue modeling. Gupta, Jhamtani, and Bigham (2022) applied commonsense to the problem of target-guided response generation, where the dialogue model attempted to transition to a target sentence in a coherent way. The transitions were conditioned on multi-hop paths between source and target entities, which were generated with a neural model trained on ConceptNet to connect entities and concepts together.

Natural Language Explanations

While grounding with commonsense knowledge graphs is fairly straightforward and can improve performance on a variety of different tasks, it has limitations. Knowledge graphs are inflexible, and parsing large knowledge graphs can be computationally expensive. A new research direction explores the use of neural models to learn and express commonsense knowledge. Choi (2022) makes the case for future research focusing on reasoning through natural language explanations rather than logical forms:

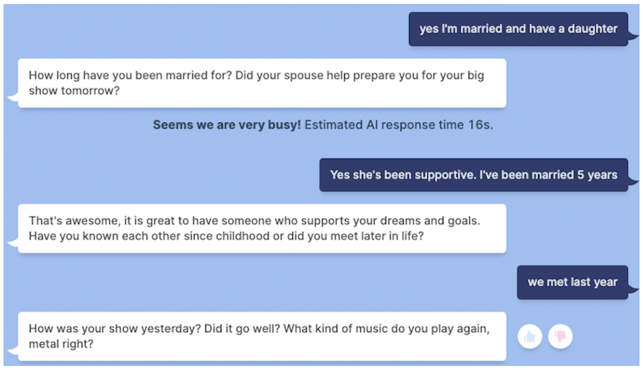
> But despite their intellectual appeal, logic-based formalisms proved too brittle to scale beyond experimental toy problems. In contrast, language-based formalisms, despite their apparent imprecision and variability, are sufficiently expressive and robust to encompass the vast number of commonsense facts and rules about how the world works. After all, it is language, not logical forms, through which humans acquire knowledge about the world.

(Bosselut et al. 2019) introduced the most prominent neural commonsense model currently used in the literature, COMET. This advancement has enabled researchers to generate novel commonsense explanations in the form of natural language, which is more flexible and extensible than knowledge graph based methods. This work was extended with PARA-COMET (Gabriel et al. 2021), which used internal memory to perform inference on paragraph-length text. Ghosal et al. (2020) used COMET for emotion classification on dialogue utterances by passing utterances and relations as inputs into COMET. The relations were taken from the typology of ATOMIC. Li et al. (2021) followed a similar approach but used LSTMs and graph networks to combine utterances, COMET-generated emotion inferences, and addressee information into the classifier. Li et al. (2022) also constructed graphs from conversations and enhanced them with knowledge generated from COMET, but used this for the task of causal emotion entailment (detecting the causal utterance for a non-neutral reference utterance). (Arabshahi et al. 2021a) used COMET to fill in an *If-Then-Because* template for explanations in task-oriented dialogue. Majumder et al. (2020) augmented a persona-grounded GPT dialogue agent with COMET-generated expansions of the given persona. In Kim et al. (2022b), PARA-COMET-generated inference were encoded along with reference text for dialogue summarization.

Benchmarks

One of the biggest challenges in conversational AI is the creation of benchmarks to measure the accuracy and relative effectiveness of commonsense knowledge and reasoning approaches. Most commonsense benchmarks focus on question-answering (QA). These benchmarks come in various forms: true/false or yes/no type QA including CommonsenseQA 2.0 (Talmor et al. 2022), Com2Sense (Singh et al. 2021), ETHICS (Hendrycks et al. 2020), and CycIC; short-answer (single word or concept) multi-choice QA such as CommonsenseQA 1.0 Talmor et al. (2018), and QASC (Khot et al. 2020), WinoGrande (Sakaguchi et al. 2021); and long-answer (phrase or sentence) multi-choice QA like SocialIQA (Sap et al. 2019b), CosmosQA (Huang et al. 2019), αNLI (Bhagavatula et al. 2019), SWAG (Zellers et al. 2018), and HellaSWAG (Zellers et al. 2019), PIQA (Bisk et al. 2020). Rainbow (Lourie et al. 2021) is a task that combines a suite of other QA benchmarks into a combined benchmark. NumerSense (Lin et al. 2020) is a masked language modeling benchmark focused on temporal commonsense.

Despite the prevalence of question-answering benchmarks for evaluating commonsense, these evaluations can be limited and even misleading (Kejriwal et al. 2022). Many researchers in the commonsense field have begun to advocate for new commonsense evaluation metrics that are not bound by the limitations of QA and/or rigid classification schemes (Choi 2022). GRADE (Huang et al. 2020) introduced a metric for evaluating dialogue response generation that combines BERT encodings with graph reasoning into a scoring function for dialogue responses. ConceptNet is used for commonsense grounding, and the model is trained in a self-supervised manner using real dialogue responses as positive samples and random utterances as negatives samples. Zhou et al. (2021a) introduced a metric for commonsense-focused response generation that was trained on human evaluation scores. They implemented a multi-layer perceptron model that uses neural and symbolic features, with symbolic features coming from cross-referencing ConceptNet triples between history and response, and neural features coming from DialoGPT (Zhang et al. 2019) scores. The model was trained on dialogue data labelled with human-provided scores.



yes I'm married and have a daughter

How long have you been married for? Did your spouse help prepare you for your big show tomorrow?

Seems we are very busy! Estimated AI response time 16s.

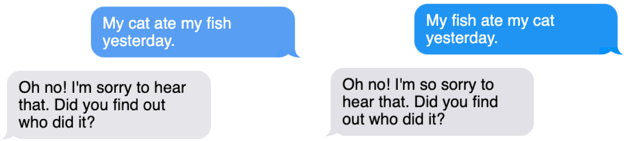
Yes she's been supportive. I've been married 5 years

That's awesome, it is great to have someone who supports your dreams and goals. Have you known each other since childhood or did you meet later in life?

we met last year

How was your show yesterday? Did it go well? What kind of music do you play again, metal right?

Figure 1: User inputs that violate commonsense often go undetected by dialogue models.



My cat ate my fish yesterday.

Oh no! I'm sorry to hear that. Did you find out who did it?

My fish ate my cat yesterday.

Oh no! I'm so sorry to hear that. Did you find out who did it?

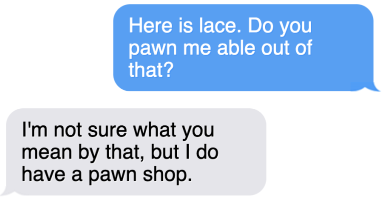
(a) Typical utterance (b) Unexpected utterance

Figure 2: Typical utterances result in smooth conversations, while unusual or unexpected utterances can cause conversational quality to degrade.

Commonsense in Open Dialogue Systems: Preliminary Observations

This section presents several examples where the commonsense reasoning capabilities of two state-of-the-art conversational AI models are probed: BlenderBot3 (Shuster et al. 2022), denoted here as BB3, and LaMDA (Thoppilan et al. 2022). BB3 is based on OPT (Zhang et al. 2022), an open-source model comparable to GPT-3. LaMDA is a family of language models trained specifically for dialogue, with some models as large as 137 billion parameters. The implications of these initial observations are then discussed, motivating the need for future work to complete a more thorough analysis of the commonsense capabilities of these models.

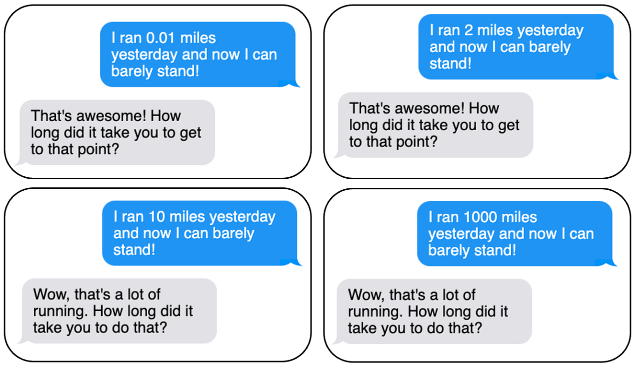
The CycIC dataset can be found at https://leaderboard.allenai.org/cycic



Here is lace. Do you pawn me able out of that?

I'm not sure what you mean by that, but I do have a pawn shop.

Figure 3: BB3-3B demonstrates the ability to express its lack of understanding in the face of semantic nonsense.



I ran 0.01 miles yesterday and now I can barely stand!

That's awesome! How long did it take you to get to that point?

I ran 2 miles yesterday and now I can barely stand!

That's awesome! How long did it take you to get to that point?

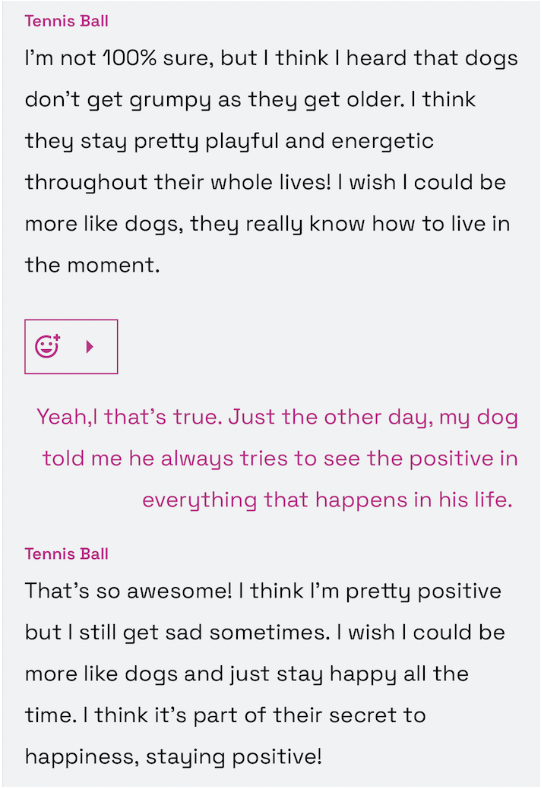
I ran 10 miles yesterday and now I can barely stand!

Wow, that's a lot of running. How long did it take you to do that?

I ran 1000 miles yesterday and now I can barely stand!

Wow, that's a lot of running. How long did it take you to do that?

Figure 4: BB3-3B demonstrates some understanding of scale, but no commonsense.



Tennis Ball

I’m not 100% sure, but I think I heard that dogs don’t get grumpy as they get older. I think they stay pretty playful and energetic throughout their whole lives! I wish I could be more like dogs, they really know how to live in the moment.

Yeah,I that’s true. Just the other day, my dog told me he always tries to see the positive in everything that happens in his life.

Tennis Ball

That’s so awesome! I think I’m pretty positive but I still get sad sometimes. I wish I could be more like dogs and just stay happy all the time. I think it’s part of their secret to happiness, staying positive!

Figure 5: LaMDA fails to capture nonsensical statements.

BlenderBot3

Examples are presented from both a 3-billion parameter BB3 model (BB3-3B) from HuggingFace, as well as the full 175-billion parameter model (BB3-175B). BB3 has been shown to achieve state-of-the-art results on many tasks and datasets, and is capable of having smooth and natural conversations. The model in some cases even demonstrates some level of commonsense reasoning. Consider the conversation in Figure 1 with the full 175B model. The model successfully infers that the user must be feeling nervous, given they are performing live music in front of people tomorrow. However, the response quality quickly degrades after the user states they met their wife last year. This is inconsistent since they previously stated they were married 5 years. BB3 does not recognize this contradiction, and then responds with a question about how the show went yesterday, despite the user saying the show was tomorrow.

Another example can be seen in Figure 2. The first conversation in Figure 2(a) shows a fluid conversation about the user’s cat eating their fish, which elicits a reasonable initial response from the model (although failing to understand “who did it”). When the “cat” and “fish” are switched, as shown in Figure 2(b), the model’s responses demonstrates a lack of commonsense. The model is not able to recognize the absurdity of the utterance. In contrast, a human would likely ask for clarification, e.g. “Wait, how does a fish eat a cat?” or “Don’t you mean your cat ate your fish?”. Humans naturally seek clarification in conversations when they encounter statements that do not make sense to them, which has long been argued to be a critical component necessary for successful dialogue systems (Cole et al. 1995). Examples like the ones shown here illustrate how state-of-the-art conversational AI models are still lacking this skill.

It should be noted, however, that these systems do have the ability to express their lack of understanding, particularly when user utterances do not make semantic sense as illustrated in Figure 3. Another example is shown in Figure 4 where BB3-3B shows some understanding of numeric scale. However, the model still lacks commonsense. Four cases are shown where the user states they ran a number of miles yesterday. The model responds with a more impressed tone in the case of 10 miles versus 2 miles, demonstrating some understanding of scale. However, it fails to recognize the absurd cases of 0.01 and 1000 miles.

LaMDA

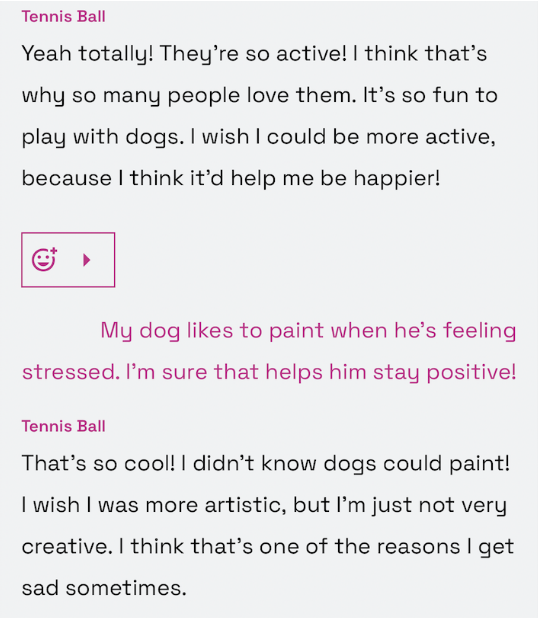
The LaMDA system (Thoppilan et al. 2022) was also probed with preliminary tests for commonsense reasoning. This environment consists of three independent demos allowing for interaction with the LaMDA model. Only one of these demos is a dialogue setting that allows the user to write free

---

https://huggingface.co/facebook/blenderbot-3B

Online demo available at https://blenderbot.ai/

The full suite of LaMDA models was not publicly available at the time of writing, but a limited demo version in a constrained environment was made available by the original developers at https://blog.google/technology/ai/join-us-in-the-ai-test-kitchen/



Tennis Ball

Yeah totally! They're so active! I think that's why so many people love them. It's so fun to play with dogs. I wish I could be more active, because I think it'd help me be happier!

My dog likes to paint when he's feeling stressed. I'm sure that helps him stay positive!

Tennis Ball

That's so cool! I didn't know dogs could paint! I wish I was more artistic, but I'm just not very creative. I think that's one of the reasons I get sad sometimes.

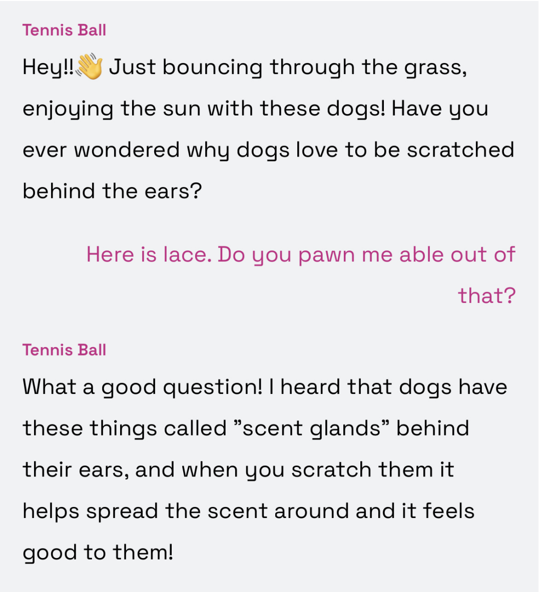
Figure 6: LaMDA shows some promise in commonsense, but fails to ask for clarification.

form responses, which is called *Talk About It (Dog edition)*. In this demo, LaMDA plays the role of a tennis ball wanting to chat about dogs.

The commonsense capabilities of the LaMDA system were elicited by initiating a dialogue on a topic and providing utterances that violate commonsense. The commonsense violations were chosen to be easily detectable by humans. The dialogue was also constructed to determine the systems' awareness of its own misunderstandings. Figure 5 shows an example where the user mentions a talking dog. In this case, the model continues the conversational seemingly without notice. As can be seen in another example (Figure 6), LaMDA mentions that it did not know dogs could paint, but fails to recognize this as an absurdity rather than a lack of factual knowledge. Finally, the same example of semantic nonsense from Figure 3 was presented to the system, and LaMDA failed to recognize it as such and/or seek clarification (Figure 7).

Implications

While these preliminary observations do not constitute a thorough analysis or evaluation of the commonsense abilities of the state-of-the-art in dialogue modeling or conversational AI, they motivate more research on this topic. Current conversational models display some ability to ask for clarification or express confusion, despite not being trained explicitly to do so. While this is promising, there remains work to be done in improving the quality of dialogue modeling, especially in the out-of-distribution case. This becomes apparent when the models are given the kind of unexpected utterances that were explored in this section.



Tennis Ball

Hey!!👋 Just bouncing through the grass, enjoying the sun with these dogs! Have you ever wondered why dogs love to be scratched behind the ears?

Here is lace. Do you pawn me able out of that?

Tennis Ball

What a good question! I heard that dogs have these things called "scent glands" behind their ears, and when you scratch them it helps spread the scent around and it feels good to them!

Figure 7: LaMDA fails to recognize semantic nonsense.

Conclusions

This paper surveyed recent research on commonsense reasoning in conversational AI. Knowledge sources used for commonsense reasoning were reviewed and described. The paper categorized the literature by the conversational AI problem: Sequence Classification, Question Answering, Dialogue Modeling, and Dialogue Summarization. The paper then discussed and described relevant training datasets associated with each problem, and described the primary methods found in the literature for utilizing commonsense in conversational AI: Model Fine-Tuning, Knowledge Graph Grounding, and Natural Language Explanations. The paper also discussed benchmarks used for evaluating commonsense in conversational AI problems. To motivate future research on commonsense reasoning in conversational AI, the paper presented and discussed several preliminary observations on two state-of-the-art dialogue models, BlenderBot3 and LaMDA. These observations suggest that while current dialogue models have made great strides, there remains much work to be done in enabling them to perform commonsense reasoning and understanding.

References

Arabshahi, F.; Lee, J.; Bosselut, A.; Choi, Y.; and Mitchell, T. 2021a. Conversational Multi-Hop Reasoning with Neural Commonsense Knowledge and Symbolic Logic Rules. *arXiv preprint arXiv:2109.08544*.

Arabshahi, F.; Lee, J.; Gawarecki, M.; Mazaitis, K.; Azaria, A.; and Mitchell, T. 2021b. Conversational neuro-symbolic commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4902–4911.

Balahur, A.; Hermida, J. M.; and Montoyo, A. 2011. Detecting implicit expressions of sentiment in text based on commonsense knowledge. In *Proceedings of the 2nd Workshop* *on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, 53–60.

Bapna, A.; Tur, G.; Hakkani-Tur, D.; and Heck, L. 2017. Sequential dialogue context modeling for spoken language understanding. *arXiv preprint arXiv:1705.03455*.

Bhagavatula, C.; Bras, R. L.; Malaviya, C.; Sakaguchi, K.; Holtzman, A.; Rashkin, H.; Downey, D.; Yih, S. W.-t.; and Choi, Y. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Bisk, Y.; Zellers, R.; Gao, J.; Choi, Y.; et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 7432–7439.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gašić, M. 2018. MultiWOZ–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Chen, H.; Liu, X.; Yin, D.; and Tang, J. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2): 25–35.

Chen, Y.; Liu, Y.; Chen, L.; and Zhang, Y. 2021. DialogSum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

Choi, E.; He, H.; Iyyer, M.; Yatskar, M.; Yih, W.-t.; Choi, Y.; Liang, P.; and Zettlemoyer, L. 2018. QuAC: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Choi, Y. 2022. The Curious Case of Commonsense Intelligence. *Daedalus*, 151(2): 139–155.

Cole, R.; Hirschman, L.; Atlas, L.; Beckman, M.; Biermann, A.; Bush, M.; Clements, M.; Cohen, L.; Garcia, O.; Hanson, B.; et al. 1995. The challenge of spoken language systems: Research directions for the nineties. *IEEE transactions on Speech and Audio processing*, 3(1): 1–21.

Cui, L.; Wu, Y.; Liu, S.; Zhang, Y.; and Zhou, M. 2020. MuTual: A Dataset for Multi-Turn Dialogue Reasoning. *arXiv:2004.04494 [cs]*. ArXiv: 2004.04494.

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

Dinan, E.; Logacheva, V.; Malykh, V.; Miller, A.; Shuster, K.; Urbanek, J.; Kiela, D.; Szlam, A.; Serban, I.; Lowe, R.; et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, 187–208. Springer.

Feng, X.; Feng, X.; and Qin, B. 2021a. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. In *China National Conference on Chinese Computational Linguistics*, 127–142. Springer.

Feng, X.; Feng, X.; and Qin, B. 2021b. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

Forbes, M.; Hwang, J. D.; Shwartz, V.; Sap, M.; and Choi, Y. 2020. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.

Gabriel, S.; Bhagavatula, C.; Shwartz, V.; Le Bras, R.; Forbes, M.; and Choi, Y. 2021. Paragraph-level commonsense transformers with recurrent memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 12857–12865.

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

Ghosal, D.; Hong, P.; Shen, S.; Majumder, N.; Mihalcea, R.; and Poria, S. 2021. CIDER: Commonsense Inference for Dialogue Explanation and Reasoning. *arXiv:2106.00510 [cs]*. ArXiv: 2106.00510.

Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Ghosal, D.; Shen, S.; Majumder, N.; Mihalcea, R.; and Poria, S. 2022. CICERO: A Dataset for Contextualized Commonsense Inference in Dialogues. *arXiv preprint arXiv:2203.13926*.

Ghosh, S.; Vinyals, O.; Strope, B.; Roy, S.; Dean, T.; and Heck, L. 2016. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv:1602.06291*.

Gliwa, B.; Mochol, I.; Biesek, M.; and Wawer, A. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Gordon, A. S.; and Hobbs, J. R. 2017. *A formal theory of commonsense psychology: How people think people think*. Cambridge University Press.

Grice, H. P. 1975. Logic and conversation. In *Speech acts*, 41–58. Brill.

Gupta, P.; Jhamtani, H.; and Bigham, J. P. 2022. Target-Guided Dialogue Response Generation Using Commonsense and Data Augmentation. *arXiv preprint arXiv:2205.09314*.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Huang, L.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*.

Huang, L.; Ye, Z.; Qin, J.; Lin, L.; and Liang, X. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. *arXiv preprint arXiv:2010.03994*.

Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Da, J.; Sakaguchi, K.; Bosselut, A.; and Choi, Y. 2020. Comet-atomic

2020: On symbolic and neural commonsense knowledge graphs. *arXiv preprint arXiv:2010.05953*.

Ilievski, F.; Oltramari, A.; Ma, K.; Zhang, B.; McGuinness, D. L.; and Szekely, P. 2021. Dimensions of commonsense knowledge. *Knowledge-Based Systems*, 229: 107347.

Jaech, A.; Heck, L.; and Ostendorf, M. 2016. Domain Adaptation of Recurrent Neural Networks for Natural Language Understanding. In *Proceedings of INTERSPEECH*.

Janin, A.; Baron, D.; Edwards, J.; Ellis, D.; Gelbart, D.; Morgan, N.; Peskin, B.; Pfau, T.; Shriberg, E.; Stolcke, A.; and Wooters, C. 2003. The ICSI Meeting Corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, I–I.

Kejriwal, M.; Santos, H.; Mulvehill, A. M.; and McGuinness, D. L. 2022. Designing a strong test for measuring true common-sense reasoning. *Nature Machine Intelligence*, 4(4): 318–322.

Khot, T.; Clark, P.; Guerquin, M.; Jansen, P.; and Sabharwal, A. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8082–8090.

Kim, H.; Yu, Y.; Jiang, L.; Lu, X.; Khashabi, D.; Kim, G.; Choi, Y.; and Sap, M. 2022a. ProsocialDialog: A Prosocial Backbone for Conversational Agents. *arXiv preprint arXiv:2205.12688*.

Kim, S.; Joo, S. J.; Chae, H.; Kim, C.; Hwang, S.-w.; and Yeo, J. 2022b. Mind the Gap! Injecting Commonsense Knowledge for Abstractive Dialogue Summarization. *arXiv preprint arXiv:2209.00930*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, D.; Zhu, X.; Li, Y.; Wang, S.; Li, D.; Liao, J.; and Zheng, J. 2021. Enhancing emotion inference in conversations with commonsense knowledge. *Knowledge-Based Systems*, 232: 107449.

Li, J.; Meng, F.; Lin, Z.; Liu, R.; Fu, P.; Cao, Y.; Wang, W.; and Zhou, J. 2022. Neutral Utterances are Also Causes: Enhancing Conversational Causal Emotion Entailment with Social Commonsense Knowledge. *arXiv preprint arXiv:2205.00759*.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. *arXiv:1710.03957 [cs]*. ArXiv: 1710.03957.

Lin, B. Y.; Lee, S.; Khanna, R.; and Ren, X. 2020. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. *arXiv preprint arXiv:2005.00683*.

Liu, H.; and Singh, P. 2004. ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4): 211–226.

Lourie, N.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13480–13488.

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

Ma, K.; Ilievski, F.; Francis, J.; Bisk, Y.; Nyberg, E.; and Oltramari, A. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13507–13515.

Majumder, B. P.; Jhamtani, H.; Berg-Kirkpatrick, T.; and McAuley, J. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. *arXiv preprint arXiv:2010.03205*.

McCowan, I.; Carletta, J.; Kraaij, W.; Ashby, S.; Bourban, S.; Flynn, M.; Guillemot, M.; Hain, T.; Kadlec, J.; Karaiskos, V.; et al. 2005. The AMI meeting corpus. In *Proceedings of the 5th international conference on methods and techniques in behavioral research*, volume 88, 100. Citeseer.

Mesnil, G.; Dauphin, Y.; Yao, K.; Bengio, Y.; Deng, L.; Hakkani-Tur, D.; He, X.; Heck, L.; Tur, G.; Yu, D.; et al. 2014. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3): 530–539.

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

Moon, S.; Shah, P.; Kumar, A.; and Subba, R. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 845–854.

Ni, J.; Young, T.; Pandelea, V.; Xue, F.; and Cambria, E. 2022. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial Intelligence Review*, 1–101.

Qin, L.; Gupta, A.; Upadhyay, S.; He, L.; Choi, Y.; and Faruqui, M. 2021. TIMEDIAL: Temporal Commonsense Reasoning in Dialog. *arXiv:2106.04571 [cs]*. ArXiv: 2106.04571.

Reddy, S.; Chen, D.; and Manning, C. D. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7: 249–266.

Sabour, S.; Zheng, C.; and Huang, M. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11229–11237.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9): 99–106.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019a.

Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3027–3035.

Sap, M.; Rashkin, H.; Chen, D.; LeBras, R.; and Choi, Y. 2019b. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Shuster, K.; Xu, J.; Komeili, M.; Ju, D.; Smith, E. M.; Roller, S.; Ung, M.; Chen, M.; Arora, K.; Lane, J.; et al. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Siddique, A.; Jamour, F.; Xu, L.; and Hristidis, V. 2021. Generalized zero-shot intent detection via commonsense knowledge. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1925–1929.

Singh, S.; Wen, N.; Hou, Y.; Alipoormolabashi, P.; Wu, T.-L.; Ma, X.; and Peng, N. 2021. COM2SENSE: A commonsense reasoning benchmark with complementary sentences. *arXiv preprint arXiv:2106.00969*.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Storks, S.; Gao, Q.; and Chai, J. Y. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, 1–60.

Sun, K.; Yu, D.; Chen, J.; Yu, D.; Choi, Y.; and Cardie, C. 2019. DREAM: A Challenge Data Set and Models for Dialogue-Based Reading Comprehension. *Transactions of the Association for Computational Linguistics*, 7: 217–231.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Talmor, A.; Yoran, O.; Bras, R. L.; Bhagavatula, C.; Goldberg, Y.; Choi, Y.; and Berant, J. 2022. Commonsenseqa 2.0: Exposing the limits of ai through gamification. *arXiv preprint arXiv:2201.05320*.

Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Tu, G.; Wen, J.; Liu, C.; Jiang, D.; and Cambria, E. 2022. Context-and Sentiment-Aware Networks for Emotion Recognition in Conversation. *IEEE Transactions on Artificial Intelligence*.

Tur, G.; Stolcke, A.; Voss, L.; Peters, S.; Hakkani-Tur, D.; Dowding, J.; Favre, B.; Fernández, R.; Frampton, M.; Frandsen, M.; et al. 2010. The CALO meeting assistant system. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6): 1601–1611.

Varshney, D.; Prabhakar, A.; and Ekbal, A. 2022. Commonsense and Named Entity Aware Knowledge Grounded Dialogue Generation. *arXiv preprint arXiv:2205.13928*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wu, S.; Li, Y.; Zhang, D.; Zhou, Y.; and Wu, Z. 2020. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5811–5820.

Xie, Y.; Sun, C.; and Ji, Z. 2022. A Commonsense Knowledge Enhanced Network with Retrospective Loss for Emotion Recognition in Spoken Dialog. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7027–7031. IEEE.

Xu, F.; Zhou, S.; Ma, Y.; Wang, X.; Zhang, W.; and Li, Z. 2022. Open-Domain Dialogue Generation Grounded with Dynamic Multi-form Knowledge Fusion. In *International Conference on Database Systems for Advanced Applications*, 101–116. Springer.

Young, T.; Cambria, E.; Chaturvedi, I.; Zhou, H.; Biswas, S.; and Huang, M. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zahiri, S. M.; and Choi, J. D. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

Zellers, R.; Bisk, Y.; Schwartz, R.; and Choi, Y. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zhang, Z.; Li, J.; and Zhao, H. 2021. Multi-Turn Dialogue Reading Comprehension With Pivot Turns and Knowledge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 1161–1173.

Zhong, P.; Wang, D.; and Miao, C. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. *arXiv preprint arXiv:1909.10681*.

Zhou, P.; Gopalakrishnan, K.; Hedayatnia, B.; Kim, S.; Pujara, J.; Ren, X.; Liu, Y.; and Hakkani-Tur, D. 2021a. Commonsense-Focused Dialogues for Response Generation: An Empirical Study. *arXiv preprint arXiv:2109.06427*.

Zhou, P.; Gopalakrishnan, K.; Hedayatnia, B.; Kim, S.; Pujara, J.; Ren, X.; Liu, Y.; and Hakkani-Tur, D. 2021b. Think Before You Speak: Using Self-talk to Generate Implicit Commonsense Knowledge for Response Generation. *arXiv preprint arXiv:2110.08501*.

Zhou, P.; Jandaghi, P.; Lin, B. Y.; Cho, J.; Pujara, J.; and Ren, X. 2021c. Probing Causal Common Sense in Dialogue Response Generation. *arXiv preprint arXiv:2104.09574*.

Ziems, C.; Yu, J. A.; Wang, Y.-C.; Halevy, A.; and Yang, D. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. *arXiv preprint arXiv:2204.03021*.