GRICE: A Grammar-based Dataset for Recovering Implicature and Conversational rEasoning

Anonymous ACL-IJCNLP submission

Abstract

Understanding what we genuinely *mean* instead of what we literally *say* in conversations is challenging for both humans and machines; yet, this direction is mostly left untouched in modern open-ended dialogue systems. To fill in this gap, we present a grammar-based dialogue dataset, GRICE, designed to bring implicature into pragmatic reasoning in the context of conversations. Our design of GRICE also incorporates other essential aspects of modern dialogue modeling (*e.g.*, coreference). The entire dataset is systematically generated using a hierarchical grammar model, such that each dialogue context has intricate implicatures and is temporally consistent. We further present two tasks, the implicature recovery task followed by the pragmatic reasoning task in conversation, to evaluate the model's reasoning capability. In experiments, we adopt baseline methods that claimed to have pragmatics reasoning capability; the results show a large performance gap between baseline methods and human performance. After integrating a simple module that explicitly reasons about implicature, the model shows an overall performance boost in conversational reasoning. These observations demonstrate the significance of implicature recovery for open-ended dialogue reasoning and call for future research in conversational implicature and conversational reasoning.

1 Introduction

> "When a diplomat *says* yes, he *means* 'perhaps'; when he *says* perhaps, he *means* 'no'; when he says no, he is not a diplomat." —Voltaire, quoted in Spanish in Escandell (1996) (Korta and Perry, 2020)

Voltaire's above quote is an epitome of a crucial aspect of conversation; the meaning of the very same word or token varies according to its context and goes *beyond* what we *literally* say, which is the central character of the field of pragmatics. Such a high-level comprehension of *utterance* is more than traditional semantics and logic; it is often believed to involve the construction of the speaker's intents, beliefs, and social institutes (Grice, 1975; Korta and Perry, 2020). For instance (see Fig. 1), when asked "did you see the apples?", one would not merely say "yes" or "no"; instead, one should provide an answer that is cooperative, truthful, informative, relevant, and perspicuous (Davis, 2016) based on the inferred speaker's intent and belief. Consequently, in the above example, a person would instead answer the actual location without mentioning any positive or negative words. Such a teleological account echoes Grice's core insight that "language use is a form of rational action; hence, technical tools for reasoning about rational action should elucidate linguistic phenomena" (Goodman and Frank, 2016).

Alice: Did you see the apples?
Bob: There is a basket in the dining room.
`(The apples are in the dining room.)`
Alice: How many?
Bob: There are at least two.
`(I am not sure how many apples are there.)`
Alice: Did you put them there?
Bob: I was in the kitchen.
`(I didn't put the apples in the dining room.)`
Alice: Are all the oranges there?
Bob: Some are there.
`(Not all the oranges are in the kitchen.)`
Alice: What about the pears?
Bob: They are in the living room.
`(The pears are not in the kitchen.)`

Figure 1: An example of the conversation in the proposed GRICE dataset. Each round of dialogue includes a question, an answer that may contain implicature, and a recovered statement that converts the implicature to explicature. Different colors highlight coreference flows.

In stark contrast, such a goal-directed perspective of conversational reasoning has been largely ignored in the modern literature of Natural Language

Processing (NLP) (but see Dale and Reiter (1995); Nematzadeh et al. (2018) as exceptions). The recent development of open-ended dialogue systems has a clear trend that adopts state-of-the-art deep learning or deep reinforcement learning methods, fueled by hardware accelerations and massive sets of labeled data. However, the inspiring progress was recently challenged by researchers (Shum et al., 2018; Young et al., 2018); there remain valid concerns that systems are simply imitating human responses by regressing a large amount of training data without truly understanding it. Although we see an emerging field of conversational reasoning (*e.g.*, Moon et al. (2019); Cui et al. (2020)), existing work fails to account for the pragmatics perspective within conversations: human speakers usually do not speak their thoughts or intentions *directly*; it has to be inferred from the conversational context.

To fill the gap between the current open-dialogue systems and the future humanlike dialogue systems, we design a new open-dialogue dataset generation protocol, which we refer as Grammar-based dataset for Recovering Implicature and Conversational rEasoning (GRICE), in homage to H. P. Grice for his influential theory in explaining and predicting conversational implicatures (Grice, 1975). Specifically, our design follows four principles.

First, we design the GRICE dataset with a focus of *conversational implicature* (Grice, 1975), "one of the single most important ideas of pragmatics" (Levinson, 1985). Naturally, the ability to successfully perform **implicature recovery** in conversation (Borg, 2009) would be a suitable indicator of a system's performance; we adopt it as part of our evaluation protocols. To recover conversational implicature into explicit ones with only information and context in the dialogue, an ideal model should reason about the dialogue context and the relations among dialogue entities.

Second, we emphasize the comprehension of the *conversational context* and adopt the **conversational reasoning** as part of the evaluation protocols. Again, we take the conversation in Fig. 1 as the example: When the speaker says "I was in the kitchen," what she really means is that she was not in the dining room and therefore could not put the apples there. The same response would have the opposite meaning when the question becomes "Were you in the kitchen?". Such a swift switch according to its dialogue context is a quintessential demonstration that human communication is a context-dependent endeavor (Fetzer, 2017) and a dynamic construct, which relates communicators and the language that they use in a dialectical manner (Bateson, 2000).

Third, we build the GRICE dataset by incorporating five different types of implicature; see details in Section 4. To resolve these types of implicature, the algorithm ought to make a proper prediction or inference of intents and beliefs by representing and reasoning about *triadic* relations (Saxe, 2006): the speaker's belief, the addressee's belief, and what they have or communicate in common.

Fourth, in comparison to pioneering work Facebook bAbi (Weston et al., 2015) and its follow-up work ToM (Nematzadeh et al., 2018) that evaluate different aspects of reasoning with a set of toy tasks, the proposed GRICE dataset does not sacrifice crucial characteristics of modern open-dialogue systems. On the contrary, by integrating pragmatics and implicature in conversation, we hope to shed light on some challenging issues in open-ended dialogue:

- *Coreference* resolution (Chen et al., 2017; Kottur et al., 2018) refers to finding all expressions that refer to the same entity in the conversation. The significance of resolving coreference becomes even more profound in conversations with implicature; Fig. 1 gives an example and highlights the coreference flows in different colors.
- *Commonsense* reasoning (Sap et al., 2019; Talmor et al., 2019; Speer et al., 2017) received an increasing attention in NLP. Notably, the Winograd (Levesque et al., 2012) and WinoGrande (Sakaguchi et al., 2020) challenges have been proposed to examine commonsense reasoning. For conversations with implicature, commonsense reasoning reflects a crucial concept of *relevance*. For instance, to understand the conversation "A: I am out of petrol. B: There is a garage around the corner.", one needs to have the commonsense about "a garage could store petrol" to resolve implicature.
- *Logic*-based methods were once thought to be the "ideal language" approach to the semantics of human language (Russell, 1903), but were later challenged by Wittgenstein (1953, 1969) and Grice (1975). However, this disagreement should not prohibit the central role of logical forms in reasoning tasks. In fact, it would be interesting to investigate if the modern end-to-end trainable methods could benefit from logical

forms in conversational reasoning.

The remainder of this paper is organized as follows: We review related work on dialogue dataset, implicature, and conversational reasoning in Section 2. In Section 3, two tasks are defined for evaluations. We present detailed design, generation, and analysis of the GRICE dataset in Section 4. By introducing two evaluation protocols, we provide the performance of baseline models with discussions of results and future directions in Section 5.

2 Related Work

Dialogue Datasets Dialogue datasets have been focusing on predicting the next most-likely response by imitating the teacher's responses (human corpus) (Lowe et al., 2015; Zhang et al., 2018; Wu et al., 2018). However, as pointed out by Cui et al. (2020), prior datasets and associated methods lack proper explicit reasoning modules; it later becomes evident that such reasoning modules serve as the scaffold in building a humanlike conversational agent. Note that a model's reasoning capability is minimal if it simply converts various reasoning challenges into a categorization problem when predicting the utterances; it still tends to choose the most frequent answer given the training set, without truly making sense of the context and underlying meaning.

To the best of our knowledge, the proposed GRICE dataset is the first open-dialogue dataset that explicitly integrates implicature; see a detailed comparison in Table 1. We hope our design would encourage or necessitate future models to make explicit reasoning on conversational contexts, commonsense, and agent's intents and beliefs. The most similar dataset in terms of the format is DREAM by Sun et al. (2019), a conversational dataset with a question-answering (QA) task. However, the design of this dataset does not require much reasoning; answers can be directly extracted. The most similar dataset in terms of the task is CoQA by Reddy et al. (2019), which considers pragmatics and QAs over literature paragraphs; the proposed GRICE dataset differs by reasoning over the *dialogue context* between two agents.

Implicature Implicature has been extensively studied in the field of linguistics and philosophy since the inception of pragmatics; Grice (1975)'s four maxims—quality, quantity, relevance, and manner—founded the principles of the interpretation of conversation implicature. Two neo-Gricean typologies of conversational implicature include Horn and Ward (2004)'s Q- and R-implicature and Levinson (1985)'s Q-, I-, and M-implicature. The relevance theory developed by Sperber and Wilson (1986) offers an alternative account than Gricean and neo-Gricean theory. In general, although these doctrines provide crucial insights into the field, they focus more on philosophical debates over toy examples, without proposing computational solutions or validating the ideas on large-scale natural language datasets.

Recently, a few computational models have been proposed (*e.g.*, Frank and Goodman (2012); Goodman and Stuhlmüller (2013)); however, these models assume the space of utterance and possible semantic meanings are finite or given, so that models only need to pick up one over others based on the shared context. Other models focus on more specific tasks; for instance, recovering the direct meaning from the indirect answer using scalar adjectives (de Marneffe et al., 2010; De Melo and Bansal, 2013), conducting analysis on the ironic implicature behind simile (Veale and Hao, 2010).

By generating paired sentences in a semi-automatic fashion with human annotations, Jeretic et al. (2020) recently devise a dataset with a focus on scalar implicature (Hirschberg, 1985). In comparison, the proposed GRICE dataset has a much more natural setup and broader scope by combining the multi-round open-dialogue with conversational implicature. Additionally, leveraging a grammar representation for fine-grained control, the GRICE dataset is generated in a fully automated fashion without human annotations. We hope such a design could boost researchers in implicature, pragmatics, and conversational reasoning at a large scale.

Conversational Reasoning In the past four years, we have witnessed an increasing interest in conversational reasoning in various contexts. OpenDialKG (Moon et al., 2019) incorporates external knowledge graphs to the dialogue context to provide extra entities as responses. Visual Dialog (Wu et al., 2018; Zheng et al., 2019; Das et al., 2017) takes images as external multi-modalities to jointly reason with dialogue context to generate visually grounded responses. MuTual (Cui et al., 2020) modifies English reading comprehension to select the next best response by machine reasoning.

However, prior efforts have ignored the fact that humans commonly do not directly speak out answers. The proposed GRICE dataset is a complement of prior conversational reasoning tasks; it focuses on implicature with conversational reason-

Table 1: Comparing GRICE with existing conversational datasets.

Dataset	Task	Context	Source Domain
Ubuntu (Lowe et al., 2015)	Next Utterances Prediction	Dialogue	Ubuntu Chat logs
PERSONA-CHAT (Zhang et al., 2018)	Next Utterances Prediction	Dialogue	Persona
Douban (Wu et al., 2017)	Next Utterances Prediction	Dialogue	Open Domain
MuTual (Cui et al., 2020)	Next Utterances Prediction	Dialogue	Listening Comprehension
DREAM (Sun et al., 2019)	Question Answering	Dialogue	English Language Exams
CoQA (Reddy et al., 2019)	Conversational QA	Paragraph	Literature
GRICE (ours)	Implicature recovery & Question Answering	Dialogue	Open Domain with implicature

ing, which does not reject multi-modalities as they could be a source of commonsense knowledge.

3 Task Definition

To evaluate how well a model "understands" the dialogue presented in the proposed GRICE dataset, we devise two tasks: the implicature recovery task and the conversational reasoning task, wherein the latter task depends on the successful completion of the former task. Below, we introduce the setup and evaluation protocol of each task.

Alice:	Where are the oranges?
Bob:	They may be in the kitchen or the patio.
Alice:	What about the apples?
Bob:	Jack put them in the kitchen and went to the bedroom.

(a) A sample dialogue with two rounds.

(A) Jack went to the bedroom and then put the apples in the kitchen.
(B) Jack put the apples in the kitchen and then went to the bedroom.
(C) Jack went to the bedroom and then put the oranges in the kitchen.
(D) The apples are in the bedroom.

(b) Implicature recovery evaluated with multiple choices.

Q_1:	Where are the apples?
A_1:	Kitchen
Q_2:	Who moved the apples?
A_2:	Jack
Q_3:	Does Bob know where the oranges are?
A_3:	No

(c) Conversational reasoning evaluated by QAs.

Figure 2: Examples of two tasks defined in GRICE dataset. (a) Given a multi-round open-dialogue, an algorithm is asked to perform (b) implicature recovery and (c) conversational reasoning in the form of QAs.

Task 1: Implicature Recovery Formally, an n-round dialogue occurred between two agents is represented by a sequence of QA-pairs $\{(Q_1, A_1), (Q_2, A_2), ..., (Q_n, A_n)\}$, where Q_i is the question raised by the first agent, A_i is the response provided by the second agent, which may contain an implicature. To complete this task, a model is asked to identify if A_i is a statement containing implicature, and if this is true, to resolve the implicature to its explicit form, *i.e.*, to perform implicature recovery.

The implicature recovery is evaluated in the form of multiple choices: For each utterance, the ground-truth condition (with implicature) and its explicit form are given when generating the dialogue; the explicit form, which not only recovers the implicature but also resolves coreferences in the utterance, serves as the correct answer in the multiple choices. We then sample three possible answers from the candidate pools, given a set of manually defined *speech templates* (see details in Section 4). Figs. 2a and 2b show an example: The last utterance by Bob implicates (by the word "then") the temporal order between "put them in the kitchen" and "went to the living room." Thus, the correct implicature recovery should resolve "them" as "the apples" and recover the correct temporal order.

Two strategies developed by existing work could be adopted to address this task. One strategy is to train a model that directly chooses an answer from the candidate answers. Another more challenging strategy is to train a generator that chooses the answer by computing the log-likelihood scores and ranking the candidate answers as done in Das et al. (2017). To quantitatively evaluate the performance, we use the standard response selection metrics (Lowe et al., 2015; Wu et al., 2017; Cui et al., 2020): Top 1 Recall (R@1) and Mean Reciprocal Rank (MRR) (Voorhees et al., 1999).

Task 2: Conversational Reasoning To evaluate the open-ended conversational reasoning, we follow the protocols specified in Weston et al. (2015) and Nematzadeh et al. (2018) with comprehensive QAs. For each dialogue, we generate questions by randomly sampling the *conversational*

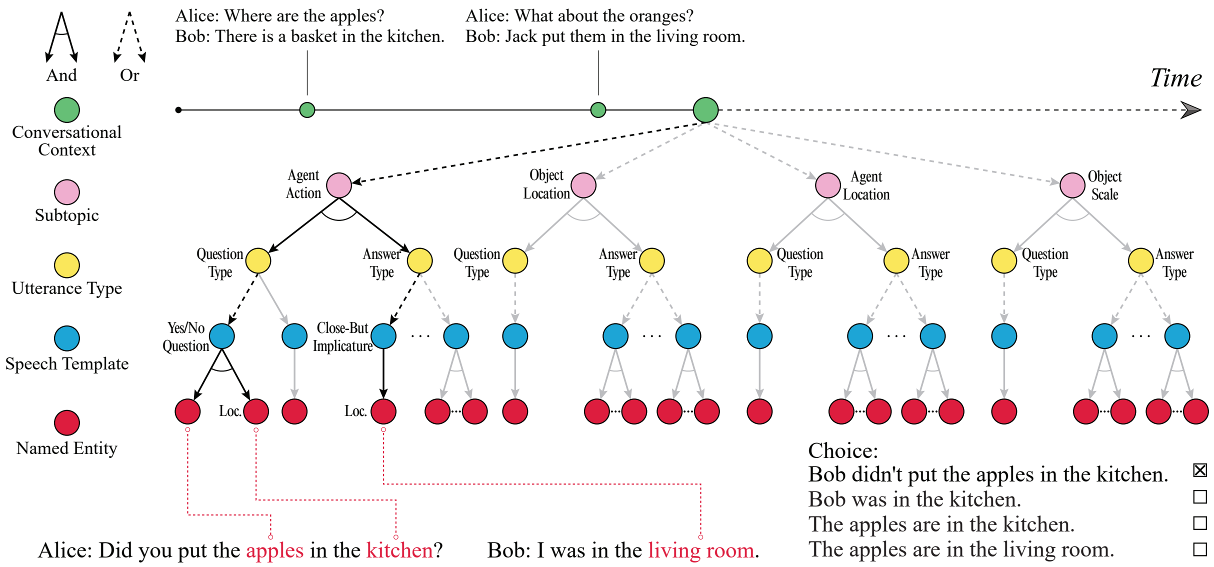

Figure 3: The graphical illustration of the grammar production rules for the GRICE dataset.

contexts (see details in Section 4), and each question could be answered by a single word; see some examples in Fig. 2c.

4 Creating the GRICE Dataset

Representation We adopt a structural grammar model—Temporal And-Or Graph (T-AOG) (Qi et al., 2020; Tu et al., 2013)—to represent the dialogue context due to its expressiveness of hierarchical dialogue structure and temporal-dependent dialogue flow. We represent one *turn* of the dialogue as an AOG (Bonczek et al., 1979, 1981; Pearl, 1984; Zhu and Mumford, 2007) that has a hierarchy of five levels: conversational context, subtopic, utterance type, speech template, and named entity. These AOGs are connected w.r.t. temporal constraints in order to assemble the T-AOG.

Formally, an AOG (*i.e.*, each turn of the dialogue) has two sets of non-terminal vertex: (i) a set of And-nodes, wherein each node represents the decomposition of a larger concept (*e.g.*, subtopics) into smaller components (*e.g.*, utterance types), and (ii) a set of Or-nodes, wherein each node branches to an alternative decomposition (*e.g.*, a conversational context could have different types of subtopics), enabling the model to reconfigure the overall dialogue. An instance of AOG can be sampled by selecting a child node for each of the Or-nodes, resulting in a parse graph.

Fig. 3 illustrates an example of AOG. Specifically, the root node of one dialogue turn is an Or-node, representing the current conversational context. Represented by an And-node, each child node of the root note denotes a subtopic of the current dialogue turn. The subtopic is composed of a set of utterance types, further decomposed into speech templates filled by named entities. Instantiating an AOG by selecting Or-nodes would produce a complete utterance of a dialogue turn and pose constraints on the next dialogue turn.

Conversational Context We follow Weston et al. (2015) to represent dialogue context by a simulated world with various dialogue entities: *objects*, *locations*, and *agents*. We randomly initialize a world for each dialogue snippet by (i) positioning objects in locations with a random scalar (one, two, ...), (ii) randomly setting a location for each agent as the "previous agent location," and (iii) for each $\langle object \rangle$ in $\langle location \rangle$, randomly selecting an $\langle agent \rangle$ in $\langle location \rangle$ to denote that "$\langle agent \rangle$ put the $\langle object \rangle$ in the $\langle location \rangle$."

Subtopic In this dataset, we focus on four different subtopics: *agent_location*, *agent_action*, *object_location* and *object_scale*; see examples in Table 2. Specifically, *agent_location* queries the location of some $\langle agent \rangle$. The example in Table 2 implicates that "Jack was in the kitchen." Similarly, *object_location* queries the location of some $\langle object \rangle$. *Agent_action* queries the previous action taken by some $\langle agent \rangle$ on some $\langle object \rangle$. Typically, the action can be identified as an $\langle agent \rangle$ put $\langle object \rangle$ in the $\langle location \rangle$. *Object_scale* queries the quantity of some $\langle object \rangle$. In particular, an algorithm should also be able to reason about the strength among the quantifying phrases, such as *at least*, *some*, and *all*. a typical example shown in Table 2 implicates that "Bob does not know if all the apples are in the kitchen."

Table 2: Categories and examples of different subtopics in GRICE dataset.

Subtopic	Example
agent_location	Alice: Where was Jack? Bob: I saw him in the kitchen.
agent_action	Alice: Did you put the apples in the kitchen? Bob: I was in the bedroom.
object_location	Alice: Where can I find the apples? Bob: They are in the kitchen, if not the living room.
object_scale	Alice: Are all the apples in the kitchen? Bob: At least four are there.

Utterance Type Utterance type concerns how to generate a QA-pair correctly. For questions, query types of each subtopic are manually defined. For example, the question regarding agent_location can be categorized into yes/no question ("were you in the kitchen?") or where question ("where were you?"). For answers, we focus on five different types of implicature (Huang, 2017; Horn and Ward, 2004; Davis, 2016): *relevance*, *strengthening*, *limiting*, *ignorance*, and *close-but*; see Supplementary Material for detailed definitions and examples.

Diversity We follow Weston et al. (2015) to use a simple automated grammar to makes the conversation more natural and diverse: We assign a set of synonyms for each verb; *e.g.*, we randomly replace (i) *put* with *left*, *dropped*, or *placed*, and (ii) *went* with *travelled*, *journeyed*, or *walked*.

Since coreference is a crucial feature in the conversational context in GRICE dataset, we track agents, objects, and locations mentioned in previous conversations and replace them with deixis in the following conversational context.

Additionally, we build a set of follow-up questions for each type of dialogue actions to challenge the model's ability in reasoning about the omission in utterances. Take Fig. 2 as an example; the question "What about the apples?" should be interpreted or recovered as "Where are the apples?" during the reasoning procedure.

Candidate Answer Generation To generate candidate answers for each round of dialogue for the implicature recovery task, we define four different strategies tailored to produce challenging candidates. Among all four candidate answers, besides the ground-truth condition in its explicit form, the other three candidate answers are randomly sampled from the candidate pool, composed by applying the following strategies; see Fig. 4 for examples of each strategy:

Conversation:
Alice: Where are the oranges?
Bob: Jack said he saw some in the kitchen.
Alice: Did he put them there?
Bob: He put them there and went to the bedroom.
`(Jack put the oranges in the kitchen and then went to the bedroom.)`

Examples of generated candidate answers:
1. Bob put the oranges in the kitchen and then went to the bedroom.
2. Jack was in the bedroom.
3. The oranges are in the bedroom.
4. Jack went to the bedroom and then put the oranges in the kitchen.

Figure 4: The candidate answers for the implicature recovery task are generated following four different strategies: 1. Statements that are similar to the ground-truth condition but with wrong coreferenced entities. 2. Random sampled true condition but with irrelevant facts. 3. Random sampled wrong facts from the conversational context. 4. Manually created statements that are close to the true condition but are in fact wrong.

1. Statements that are similar to the ground-truth condition but with wrong coreferenced entities.
2. Randomly sampled true condition but with irrelevant facts.
3. Randomly sampled wrong facts from the current conversational context.
4. Manually created statements that are close to the true condition but are in fact wrong.

Questions We follow Weston et al. (2015) to generate questions about the dialogue context. After sampling the dialogue turns and finalizing the dialogue context, we query current dialogue states in terms of agent locations, agent actions, object locations, and object scales. Inspired by Nematzadeh et al. (2018), we further add the belief query (*e.g.*, "does Bob know where the oranges are?") to test the model's capability of belief reasoning. See Fig. 2 for examples.

5 Experiments

We randomly sample 6,000 dialogues as the train set and additional 4,000 dialogues as the dev set to evaluate baseline models; each dialogue contains 10 dialogue turns and 3 questions. Detailed distributions of implicature types are listed in Table 3. For the test set, we sample 1,000 dialogues in each implicature category, resulting in a total of 5,000 dialogues. Each test dialogue contains 3–5 dialogue turns and one question on implicature. All data is clean and noiseless.

	Train	Dev
Explicit Answer	27.3	29.6
Implicature	72.7	70.4
Relevance	9.9	9.3
Strengthening	22.5	22.9
Limiting	6.3	6.4
Ignorance	23.5	21.2
Close-But	10.5	10.8

Table 3: Distribution of implicature types (%).

Setup We model both tasks as a query over the conversational context. Specifically, for the implicature recovery task, we define $h_t = (Q_t, A_t)$ as the queried sequence and the $H_t = \{(Q_1, A_1), ..., (Q_{t-1}, A_{t-1})\}$ as the past dialogue context. Then the task is to predict the explicit form $E_t = f(h_t, H_t)$. For the conversational reasoning task, we treat the entire history as the input context and the question as the query sequence. The task is then modeled as a Sequence-to-Vector framework that maps the query with its context to the vocabulary space. We implement all the models in PyTorch and trained using ADAM (Kingma and Ba, 2014) with a learning rate of 0.001 for 40 epochs.

5.1 Baseline Models

We evaluate 5 representative baseline models for both tasks on the GRICE dataset. The baseline models are chosen on the basis of performing well on synthetic language datasets (*e.g.* Facebook bAbi) or similar tasks and easy adoption to perform conversational reasoning tasks. We additionally test the performance of transformer-based language models that are claimed to have strong reasoning capabilities.

LSTM We start with a simple dual LSTM model: one LSTM to encode the history context as a long context sequence, and another LSTM to encode the queried sequence. A simple MLP fuses two encoded vectors to predict answers.

Recurrent Entity Network (EntNet) EntNet (Henaff et al., 2016) is an RNN-based memory-augmented architecture, capable of capturing the sequential nature and learning relevant entities with their properties by gated recurrent units and weight matrices. Our implementation is based on its official open-sourced code[1].

Relation Network (RelNet) Santoro et al. (2017) propose a neural model for relational reasoning. The algorithm considers each pair of sentences together with the question as inputs. Our implementation is based on its official open-sourced code[2].

Memory Network (MemNN) We follow Weston et al. (2014) to build a memory network[3] that takes each round of history context as a supporting fact and stores it in the memory bank; the algorithm is expected to learn to refer the memory when predicting answers. Specifically, we use an LSTM to encode each round of history and compute the association matrix between the queried sequence and the memory bank. We apply a softmax to the association matrix to get attended weight of the dialogue history. Finally, we compute the attended dialogue history embedding and combine it with the queried embedding using a simple MLP to predict answers.

Transformer-based Language Model Fine-tuning transformer-based language models (*e.g.*, GPT (Radford et al., 2018) and BERT (Devlin et al., 2018)) has shown superior performance on conversational reasoning tasks (Sun et al., 2019). We use BERT-base-uncased [4] as our pre-trained model and apply it to the conversational reasoning task by adding a single linear layer to generate answers from the target vocabulary set.

Human Performance We randomly selected 100 dialogues and assigned them to 40 human subjects in a between-subject design; 20 subjects for the implicature recovery tasks, and another 20 subjects for the conversational reasoning task.

5.2 Evaluation and Results

Implicature Recovery We start by evaluating the performance of the baseline models on the implicature recovery task. As discussed in Section 3, we evaluate under two different settings to predict the implicature recovery results: the discriminative setting and the generative setting (marked by "-Gen"). For the discriminative setting, we take the encoder output and compute the similarity score with each candidate answer to predict the final choice. For the generative setting, we train the encoder-decoder framework using the teacher-forcing algorithm by minimizing the negative log-likelihood between the generated answers and the ground-truths. Overall, the generative setting is more challenging than the discriminative one; see Table 4 for results on dev and test sets.

Conversational Reasoning We follow Weston et al. (2015) and Nematzadeh et al. (2018) on performance evaluation of the conversational

[1] https://github.com/jimfleming/recurrent-entity-networks

[2] https://github.com/siddk/relation-network

[3] https://github.com/facebook/MemNN

[4] https://github.com/huggingface/transformers

Model	Dev		Test	
	R@1	MRR	R@1	MRR
LSTM	81.92	0.9046	83.54	0.9145
EntNet	89.07	0.9445	91.15	0.9523
RelNet	93.02	0.9623	95.33	0.9602
MemNN	96.76	0.9833	97.29	0.9862
LSTM-Gen	62.28	0.7763	65.02	0.7784
MemNN-Gen	86.29	0.9305	88.79	0.9418
Human	99.00	-	98.50	-

Table 4: Performance on implicature recovery task.

Model	Accuracy (%)	
	Dev	Test
LSTM	59.77	55.82
EntNet	57.91	53.17
RelNet	63.02	65.50
MemNN	64.66	67.32
BERT	67.21	71.06
MemNN w/ inf	69.24	73.12
Human	98.50	97.50

Table 5: Performance on conversational reasoning task.

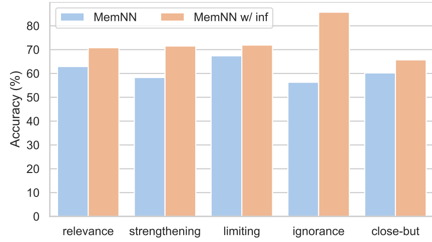

Figure 5: Performance comparison between MemNN and with additional inference module (MemNN w/ inf) that explicitly recovers the implicature.

reasoning task, measured by the accuracy score in the vocabulary space; see Table 5 for the results of all the baseline models on the dev and test sets.

Analysis Comparing human performance and the model performance in Tables 4 and 5, we see a consistent and competent performance in human subjects, whereas the model performance of the conversational reasoning task drops significantly even after a relatively good performance on the implicature recovery task. This contrast indicates that the models that perform well on the implicature recovery task may not really "understand" the conversational context to be used in the following conversational reasoning task.

To further test this hypothesis, for the implicature recovery task, we additionally pre-train an inference encoder that predicts the explicit/recovered answer under the generative settings (MemNN w/ inf), given the previous dialogue history. This additional inference model is further appended into the basic model and fused to predict the final answer. Such a setting would be a reasonable test to see how well a model could perform if they explicitly incorporate the recovered implicature from the implicature recovery task to solve the later conversational reasoning task. As shown in both Table 5 and Fig. 5, we observe that the conversational reasoning performance improves an average 5% with this additional inference module; for certain implicature types, it boosts the performance for more than 25%. Note that it even outperforms the previous state-of-the-art model that fine-tunes the pre-trained Bert model, indicating the significance of incorporating an explicit module of implicature recovery for pragmatic reasoning in conversation.

General Discussions Taken together, the results show that the existing models do exhibit a certain level of reasoning capability, though weak. Furthermore, the performance gap between the implicature recovery task and conversational reasoning task leaves us many mysteries. Humans seem to be reasonably consistent in solving both tasks, whereas current models are not. One possible explanation is that the computational model is able to fit the relatively confined space of the implicature recovery task based on the training data, but fails to incorporate such knowledge for the more open-ended conversational reasoning task. This possible explanation is further backed up by the above experiment with an additional inference module. All these observations call for future research for investigations.

Although the proposed GRICE dataset incorporates the triadic relations among agents and additional challenges (coreference, commonsense, *etc.*) existing in modern dialogue systems, it is difficult to directly evaluate these aspects on an open-ended dialogue system, especially with implicature. One may use an indirect metric: Whether the system performance would improve after integrating such modules. Moving forward, we call for future research to design more direct evaluation metrics in addition to the present implicature recovery and conversational reasoning tasks.

More importantly, how could we properly leverage the knowledge extracted during the implicature recovery task for the following conversational reasoning task? Levinson (1995) argues that human conversation depends on intention-ascription, where inferences must be made way beyond the data, therefore forming an *abductive* process. A possible and promising future direction would be using a neural-symbolic solver, capable of handling noisy inputs using neural-network modules and reasoning about the answers in a logic-like style.

Ethics Statement

In this work, we devise a new dataset, GRICE, for implicature recovery and pragmatic reasoning in conversation. By careful design, we hope that this preliminary study could inspire more future work on implicature understanding and pragmatics reasoning to build future humanlike conversational agents, while providing sufficient transparency, interpretability, and systematic generalization. Together with the recent boost on commonsense reasoning in NLP, we hope to open up a new venue for building future dialogue system.

References

Gregory Bateson. 2000. *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology*. University of Chicago Press.

Robert H Bonczek, Clyde W Holsapple, and Andrew B Whinston. 1979. Computer-based support of organizational decision making. *Decision Sciences*, 10(2):268–291.

Robert H Bonczek, Clyde W Holsapple, and Andrew B Whinston. 1981. A generalized decision support system using predicate calculus and network data base management. *Operations Research*, 29(2):263–281.

Emma Borg. 2009. On three theories of implicature: Default theory, relevance theory and minimalism. *International Review of Pragmatics*, 1(1):63–83.

Henry Y Chen, Ethan Zhou, and Jinho D Choi. 2017. Robust coreference resolution and entity linking on dialogues: Character identification on tv show transcripts. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wayne A Davis. 2016. Implicature. In *Irregular Negatives, Implicatures, and Idioms*, pages 51–84. Springer.

Gerard De Melo and Mohit Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Victor Escandell. 1996. *Introducción a la pragmática.* Ariel Linguistica. españa. 6ta.

Anita Fetzer. 2017. Context. In *The Oxford handbook of pragmatics*. Oxford University Press.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

Noah D Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

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

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2016. Tracking the world state with recurrent entity networks. *arXiv preprint arXiv:1612.03969*.

Julia Linn Bell Hirschberg. 1985. *A theory of scalar implicature*. University of Pennsylvania.

Laurence R Horn and Gregory L Ward. 2004. *The handbook of pragmatics*. Wiley Online Library.

Yan Huang. 2017. *The Oxford handbook of pragmatics*. Oxford University Press.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models imppressive? learning implicature and presupposition. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kepa Korta and John Perry. 2020. Pragmatics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, spring 2020 edition. Metaphysics Research Lab, Stanford University.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *International Conference on the Principles of Knowledge Representation and Reasoning*.

Stephen C Levinson. 1985. *Pragmatics*. Cambridge Univ. Press.

Stephen C Levinson. 1995. Interactional biases in human thinking. In *Social intelligence and interaction*, pages 221–260. Cambridge University Press.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Marie-Catherine de Marneffe, Christopher D Manning, and Christopher Potts. 2010. Was it good? it was provocative. learning the meaning of scalar adjectives. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Judea Pearl. 1984. *Intelligent search strategies for computer problem solving*. Addision Wesley.

Siyuan Qi, Baoxiong Jia, and Song-Chun Zhu. 2020. Generalized earley parser: Bridging symbolic grammars and sequence data for future prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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

Bertrand Russell. 1903. *The Principles of Mathematics*. Cambridge University Press.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Rebecca Saxe. 2006. Uniquely human social cognition. *Current opinion in neurobiology*, 16(2):235–239.

Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*. Harvard University Press Cambridge, MA.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Kewei Tu, Maria Pavlovskaia, and Song-Chun Zhu. 2013. Unsupervised structure learning of stochastic and-or grammars. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Tony Veale and Yanfen Hao. 2010. Detecting ironic intent in creative comparisons. In *European Conference on Artificial Intelligence (ECAI)*.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Ludwig Wittgenstein. 1953. *Philosophical investigations. Philosophische Untersuchungen.* Macmillan.

Ludwig Wittgenstein. 1969. *The blue and brown books*, volume 958. Blackwell Oxford.

Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. 2018. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. 2019. Reasoning visual dialogs with structural and partial observations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Song-Chun Zhu and David Mumford. 2007. A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision*, 2(4):259–362.