Analyzing and Evaluating Pragmatic Knowledge in Open-Domain Dialogue Models

Santi Pornavalai, Menghan Yu, Shengqi Zhu (Group 8)

Overview

- Three parts of the topic & ingredients of our presentation
- An "overview"; no must-reads/core papers
  - please see our slides for a full list of readings/references

Analyzing and Evaluating
Pragmatic Knowledge
in Open-Domain Dialogue Models

Introduction (Shengqi)

We will try to give an preliminary answer to these questions:

- Why do we care about analyzing and evaluating models?
- Why is pragmatics important for analysis?
- Why the open-domain dialog task? (more from Santi!)
- What are the other common practices? (besides dialog)

Why analyzing/evaluating?

- Why do we not focus on performance? Won't that suffice?
  - which are measured by Acc/F1 and BLEU/ROUGE/…?

Why analyzing/evaluating?

- ~~Why do we not focus on performance? Won't that suffice?~~ NO!!
    - ~~which are measured by Acc/F1 and BLEU/ROUGE/…?~~

- Not a good claim for many reasons
    - explanability, robustness, privacy and ethics…

- But there's even another important issue: "Performance" is not intrinsic

  which model is better depends on *how we judge them*

- We'll see some most significant issues for overall scoring

Some tasks aren't meant to be caged

- Comparing the system output with references doesn't help much
  - is made even worse with *word-overlapping metrics* (BLEU/ROUGE/…)
  - Classic examples: chatbots, (abstractive) summarization, …
  - More importantly: ***slightly higher overlap doesn't mean anything***

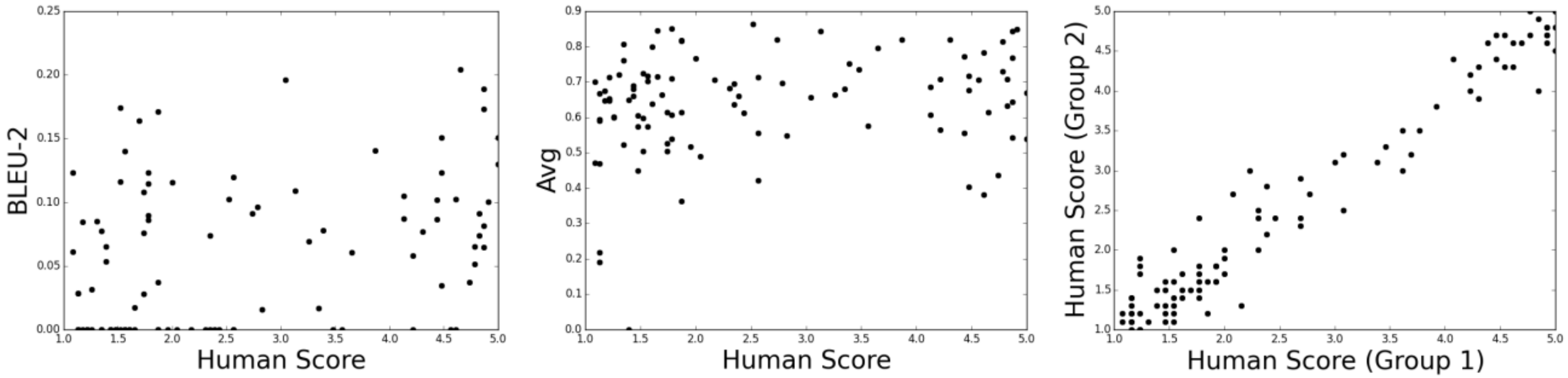
Context of Conversation
Speaker A: Hey John, what do you want to do tonight?
Speaker B: Why don't we go see a movie?
Ground-Truth Response
Nah, I hate that stuff, let's do something active.
Model Response
Oh sure! Heard the film about Turing is out!

Table 1: Example showing the intrinsic diversity of valid responses in a dialogue. The (reasonable) model response would receive a BLEU score of 0.

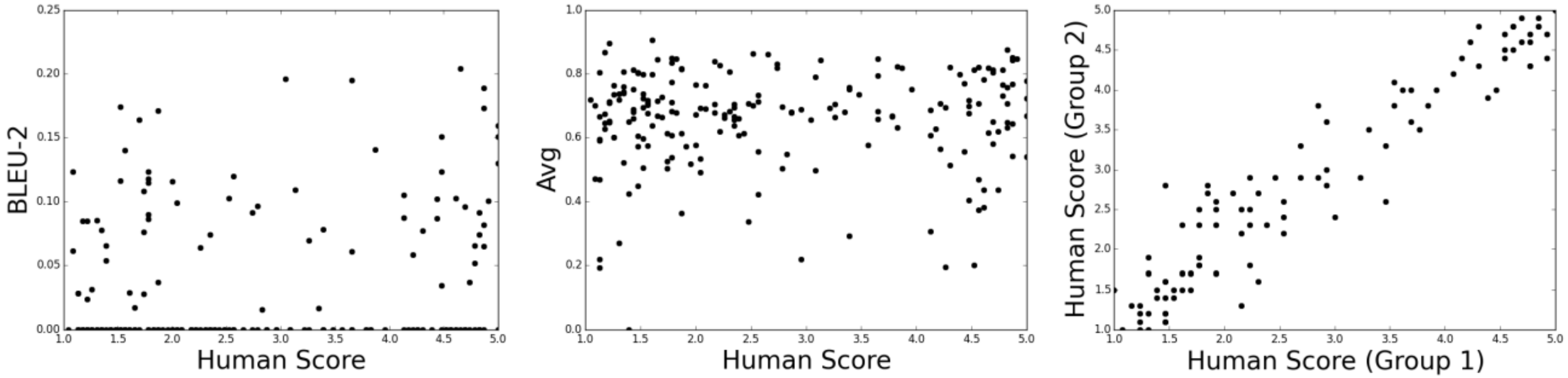
Ref: [How NOT To Evaluate Your Dialogue System](), EMNLP 2016

Does it really work?

- Frustratingly low correlation with human judgements and with other metrics



(a) Twitter



Ref: [How NOT To Evaluate Your Dialogue System](), EMNLP 2016

Other issues

- The nature of some tasks needs fine-grained scores on *certain aspects*

- How to convince someone to adopt a new metric?

- Towards the actual use and understanding of language (beyond surface forms)

Ref: Why We Need New Evaluation Metrics for NLG, EMNLP 2017;
A Survey of Evaluation Metrics Used for NLG Systems, ACM Computing Surveys

Why analyzing pragmatics?

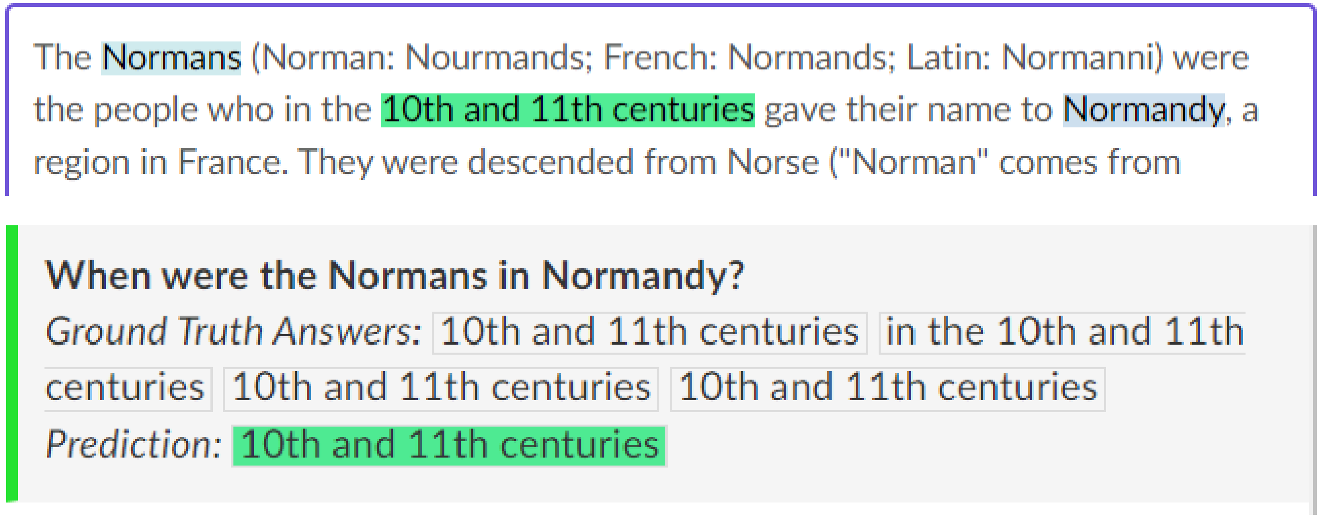
- Pragmatics are subtle and extremely undetectable with overall scores

- Many aspect-level judgments essentially involve pragmatic concerns
  - coherence, relevance, …

- Less studied area in the NN era
  - Paper numbers from ACL 2020 (+Workshops):
  - Syntax/Syntactic: 24
  - Semantic(s): 41
  - Pragmatic(s) + a list of possible key words: 2 🥲

Why open-domain dialog?

- One (mentioned) reason: ample space for “correct” answers
- The opposite case: overall accuracy is what really matters
  - or serves as a fairly good proxy
  - Question Answering (Reading Comprehension) is more like this fashion



Ref: [The Stanford Question Answering Dataset]

Why open-domain dialog?

- Another (trivial yet important) reason:

  Dialogues are the place where many pragmatic theories originate!

- Goal of mutual understanding, The “turn-taking” nature, …
- (no attempts to cover the whole history of pragmatics here XD)

- Santi will give more insights on the properties of the dialog task!

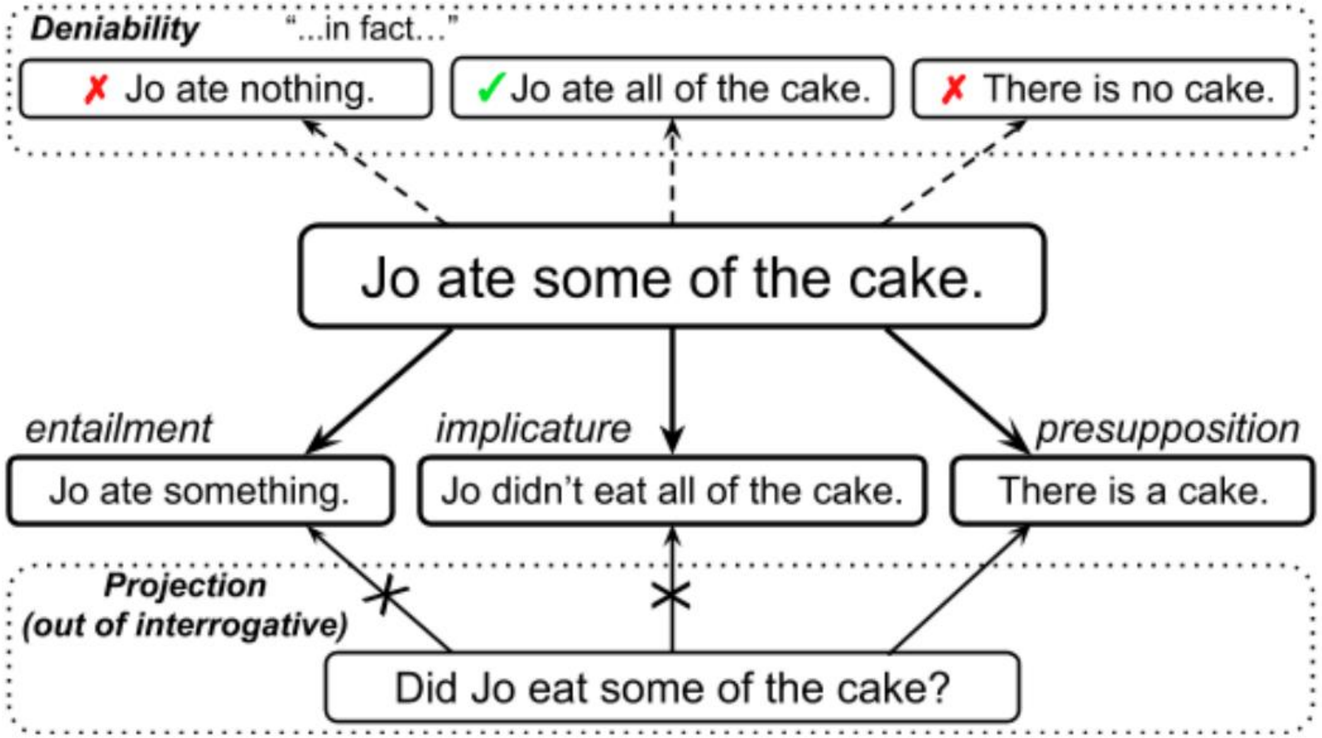
Ref: https://plato.stanford.edu/entries/implicature

But it doesn’t have to be dialogs :)

- Many auxiliary tasks and probing methods can apply
  - that doesn’t rely on dialog (or even not NLU)
  - e.g. Diagnostic Classifiers

- We’ll quickly look at some of the methods

NLI for implicatures and presuppositions



- IMPPRESive dataset
- (similar work: NOPE dataset)

Ref: [Are Natural Language Inference Models IMPPRESsive](), ACL 2020
[NOPE: A Corpus of Naturally-Occurring Presuppositions in English](), CoNLL 2021

NLI for implicatures and presuppositions

Trigger	Affirmative Example	Negative Example	Presupposition
Change of state	A microsecond later, images from his exterior sensors snapped into focus.	A microsecond later, images from his exterior sensors **didn't** snap into focus.	Previously, images from his exterior sensors hadn't been in focus.
Clefts	But it is the horse racing that is just for children.	But it **isn't** the horse racing that is just for children.	There's something that is just for children
Comparatives	That is a bigger problem, than the chairman's claim.	That **isn't** a bigger problem, than the chairman's claim.	The chairman's claim is a problem.
Aspectual verbs	At the age of 55, I began preparing myself to die.	At the age of 55, I **didn't** begin preparing myself to die.	Before age 55, I was not yet preparing to die.
Embedded questions	I fail to see how you can rationalize rewarding illegality.	I **don't** fail to see how you can rationalize rewarding illegality.	You can rationalize rewarding illegality.
Clause-embed. verbs	In 20 years we'll realize that's a mistake.	In 20 years we **won't** realize that's a mistake.	[Pushing people towards pharmaceuticals] is a mistake.
Implicatives	The survivors managed to scramble out through the tiny gap in the rocks.	The survivors **didn't** manage to scramble out through the tiny gap in the rocks.	The survivors made an attempt to scramble out through the tiny gap in the rocks.
Numeric determiners	Both protagonists in the room defy a political force and receive aid from a higher authority.	Both protagonists in the room **do not** defy a political force and receive aid from a higher authority.	There are two protagonists in the room.
"Re-" prefixed verbs	Taoism reconnects aging to the great cycles of nature.	Taoism **doesn't** reconnect aging to the great cycles of nature.	Aging was once connected to the great cycles of nature.
Temporal adverbs	He took them to the NL Championship Series last year before being swept by the Atlanta Braves.	He **didn't** take them to the NL Championship Series last year before being swept by the Atlanta Braves.	Johnson was swept by the Atlanta Braves.

Ref: Are Natural Language Inference Models IMPPRESsive, ACL 2020
NOPE: A Corpus of Naturally-Occurring Presuppositions in English, CoNLL 2021

Prompting Language Models

- Example: prompting *reporting bias* of color in LMs

- Reporting bias: people chose not to say obvious things (Gricean maxim)

- What colors are bananas?
- In real life: yellow bananas >> green/red/blue bananas >>> other colors
- In LMs: “green banana” = 332% “yellow bananas”!

Ref: [The World of an Octopus: How Reporting Bias Influences a Language Model’s Perception of Color](), EMNLP 2021

Prompting Language Models

- Prompting: "Most bananas are ___ ."

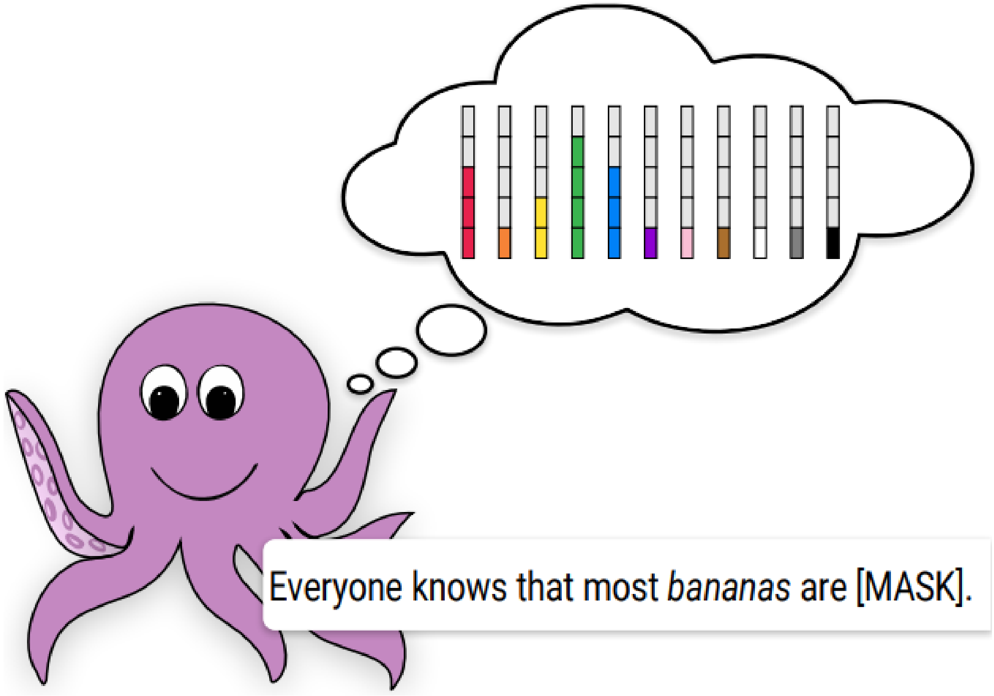
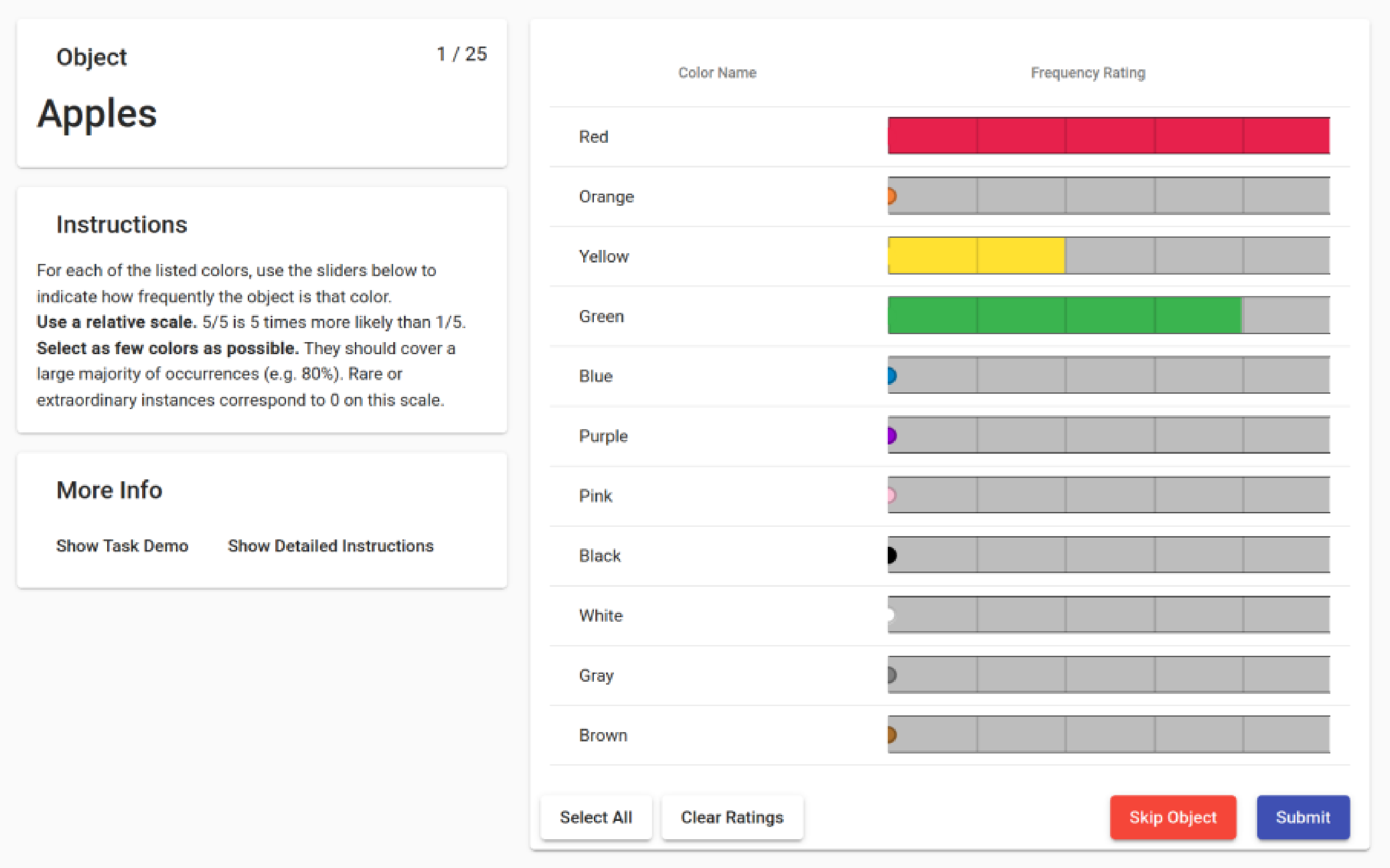


Figure 1: An example prompt from CoDa.

Ref: [The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color](), EMNLP 2021

Prompting Language Models

- Compare with human annotations of colors



Ref: [The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color](), EMNLP 2021

Auxiliary Training Losses for Pragmatics

- Introduce Auxiliary training objectives and losses for NLU models
- e.g. relevance requirement

We optimize the ranking log likelihood

$$L_{\text{rel}} = \sum_{\substack{(\mathbf{x},\mathbf{y}_g)\in D, \\ \mathbf{y}_r \sim D_{\mathbf{y}}}} \log \sigma(s_{\text{rel}}(\mathbf{x},\mathbf{y}_g) - s_{\text{rel}}(\mathbf{x},\mathbf{y}_r)), \tag{10}$$

where $\mathbf{y}_g$ is the gold ending and $\mathbf{y}_r$ is a randomly sampled ending.

$$a = \text{maxpool}(\text{conv}_a(e(\mathbf{x}))),$$
$$b = \text{maxpool}(\text{conv}_b(e(\mathbf{y}))).$$

The scoring function is then defined as

$$s_{\text{rel}} = \mathbf{w}_l^T \cdot (a \circ b),$$

Ref: Learning to Write with Cooperative Discriminators, ACL 2018

Thanks!

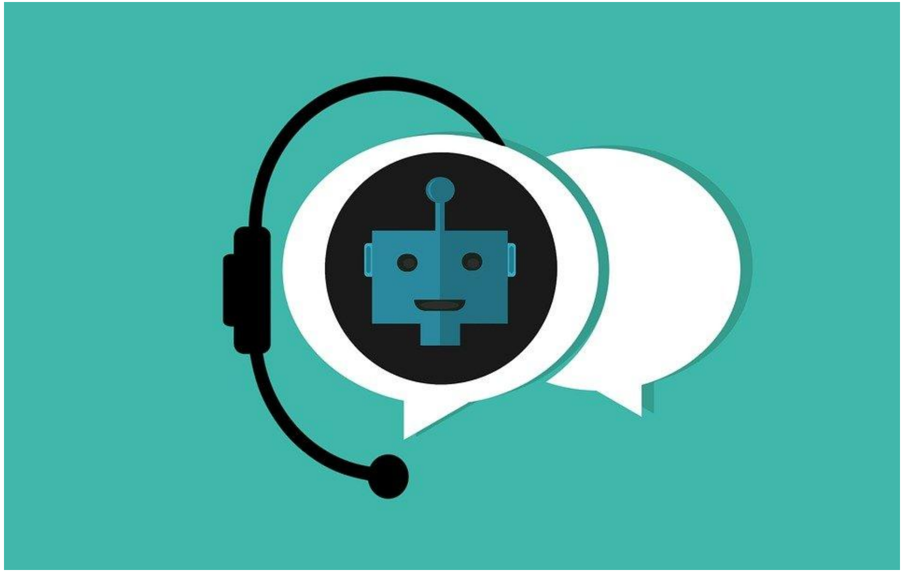
Our Experiment Designs

- (still under construction)
- Overall question: which parts of the input are most *relevant/important* ?
- Adversarial sets:
- (1) equally plausible (or similar) answers without certain components
- (2) same dialog acts, without certain components
- (3) with certain components, but non-relevant
- Off-the-shelf models/fine-tune on the pragmatic task
- Different metrics (Semantic Similarity vs. Interpretability)

Dialogue Systems

Types of Dialogue Systems

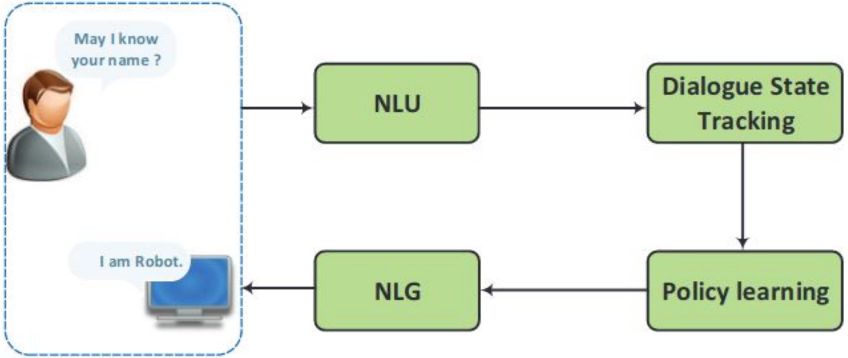
- Task/Goal-Oriented
  - Assist users in completing a task
  - Typically with pre-defined goals
  - Virtual assistants, find restaurants
- Chatbots
  - Mimic unstructured conversations
  - Open-ended
  - Often combined with Task-Oriented



Next few slides paraphrased from Chen et al (2017) and J&M 3rd ed.

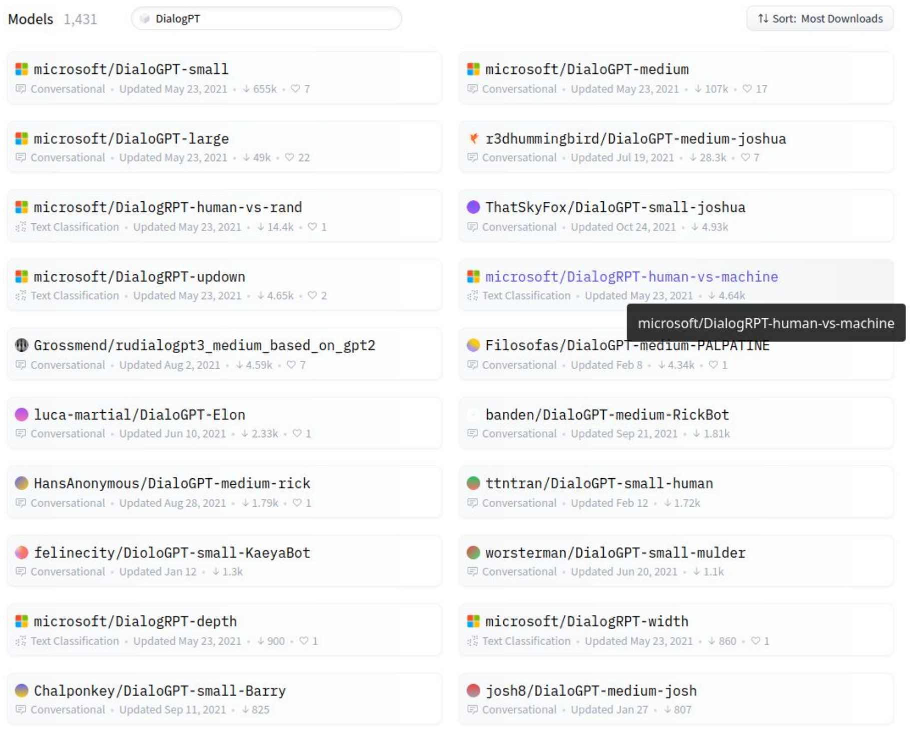
Architectures

- Pipeline
  - Most common for Task-oriented
- Retrieval
  - Treat dialogue as an IR problem
- Rule-based
- End-to-end Generation



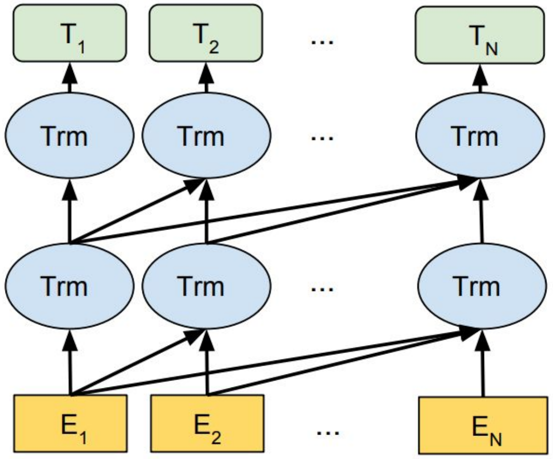
Chatbots

- Rule-Based & Retrieval methods were the norm
- LLMs enable end-to-end generative systems
- Pre-training/ Fine-tuning paradigm



DialoGPT (Zhang et al, 2020)

- Based on GPT-2 Architecture
- Dialogue as (causal) language modelling
- Pre-trained on a corpus of *Reddit* threads
- Best results obtained when continuing from the standard GPT-2
- Uses a separate model to score and rank informative responses



$$p(T|S) = \prod_{n=m+1}^{N} p(x_n|x_1, \cdots, x_{n-1})$$

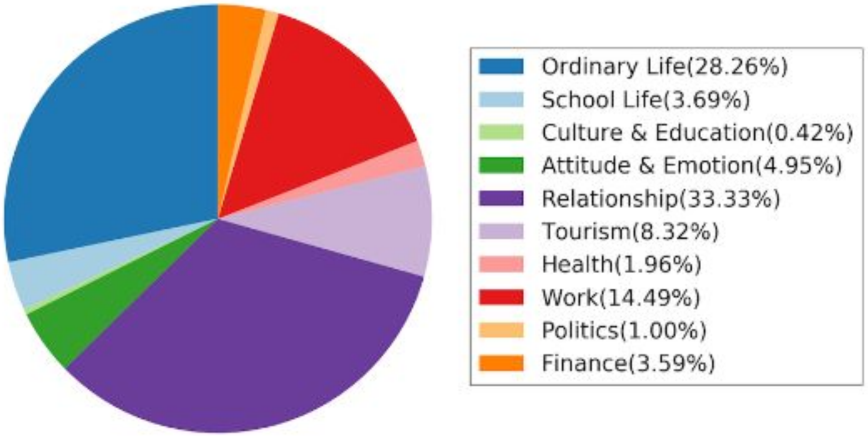
$$p(T_i|T_1, \cdots, T_{i-1})$$

Pre-Training Data

- Reddit reply chains from 2005-2017
- Filtered out replies with:
  - URLs and Markup
  - word repetitions of at least three words
  - non-English
  - Longer than 200 words
  - Offensive language (World filter)
  - “Uninformative” content
- Ca. 150 Million Dialogue instances, 1.8B tokens

DailyDialog (Li et al, 2017)

- Is social media data representative?
- Open-Domain, mix of task-oriented and chit-chat
- Dialogues are crawled from English learning websites
- Relatively short dialogues compared to social media datasets
- Annotated for dialogue acts and emotion



DailyDialog++ (Sai et al, 2020)

- Adversarial Dataset for DailyDialog
- Intended for evaluating retrieval-based methods and BertScore-type metrics
- Added added alternate responses,random negatives and adversarial examples.
- Adversarial examples were created by tasking the annotators to generate irrelevant responses given a number of words from the context.

Reference

- Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A Survey on Dialogue Systems: Recent Advances and New Frontiers. *ArXiv, abs/1711.01731*.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Analyzing pragmatics in dialogs

Group 8

Measuring the 'I don't know' Problem through the Lens of Gricean Quantity, 2021 NAACL;
GRICE: A Grammar-based Dataset for Recovering Implicature and Conversational rEasoning,
2021 ACL

Recap of Gricean Maxims

Maxim	Definition	Violated by...	Prompt: What color is grass?
QUANTITY	Be informative.	not answering a question (fully), or giving too much information.	I don't know.
QUALITY	Be truthful.	lying, or saying something without evidence.	Grass is purple.
RELATION	Be relevant.	off-topic responses.	I like pizza.
MANNER	Be clear, brief, and orderly.	disfluent responses	is green grass usually.

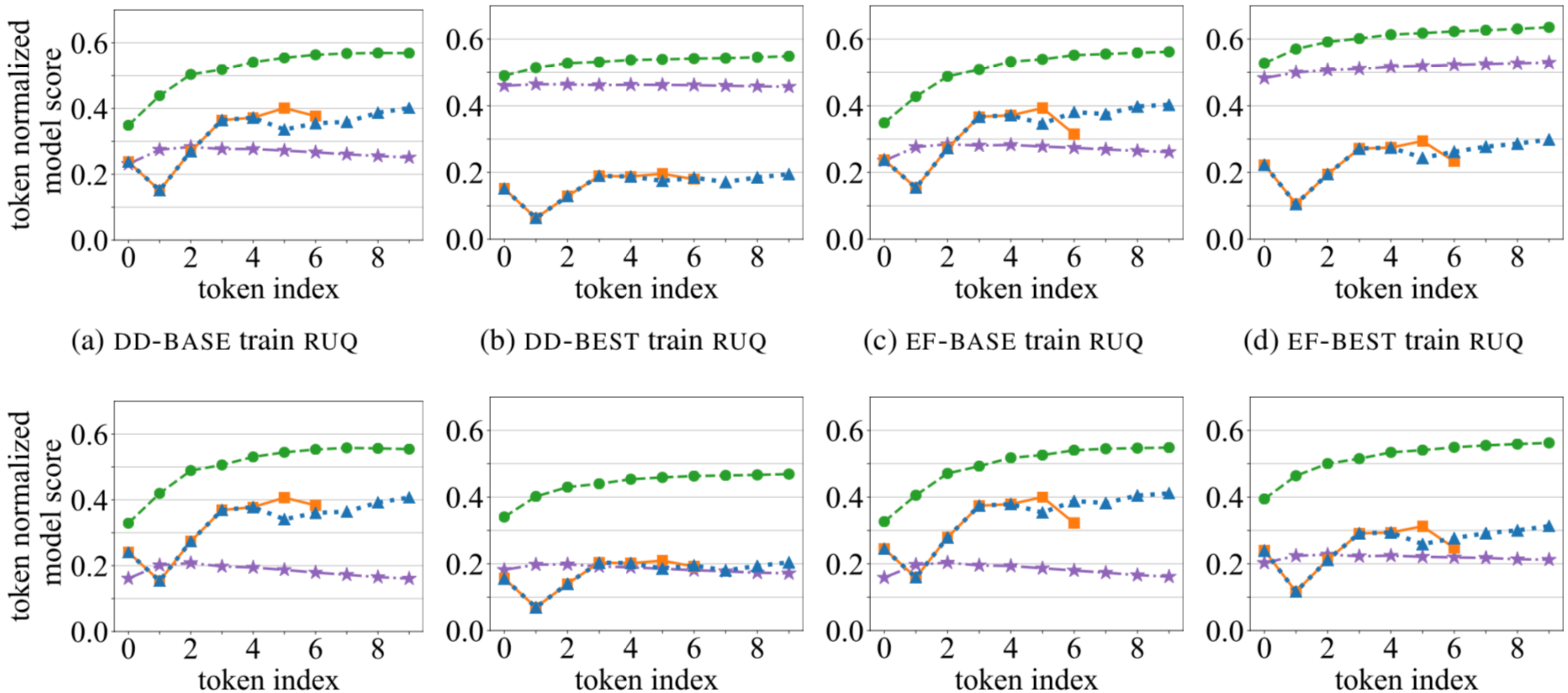
Measuring ‘I don’t know’ Problem

- Violation of the Gricean maxim of Quantity
- Measuring method: Relative Utterance Quantity (RUQ)

Methods – Relative Utterance Quantity

- Model: Transformer chatbos in FAIRSEQ using parameters from the FLORES benchmark for low-resource MT
- Plot the average model score for each token across sentences.
- compare the original reference, beam search output, and two ‘I don’t know’ (IDK) variants: ‘I don’t know.’ and ‘I don’t know what to do’.
- compute the (length normalized) model score for ‘I don’t know.’ and the reference of each training prompt, and count how many times the reference is preferred. (RUQ score)

Experiment Result



(a) DD-BASE train RUQ (b) DD-BEST train RUQ (c) EF-BASE train RUQ (d) EF-BEST train RUQ

(e) DD-BASE test RUQ (f) DD-BEST test RUQ (g) EF-BASE test RUQ (h) EF-BEST test RUQ

Experiment Result

training data	BASE	BEST
DAILYDIALOG	28.5%	95.3%
ENTROPY-FILTERED	37.9%	89.2%

Table 3: Training data RUQ scores. Entropy filtering improves how often the reference is preferred to 'I don't know.', but by less than the hyperparameter sweeps (which are denoted BEST).

Gricean Maxims – Related work

- Niels Ole Bernsen, Hans Dybkjær, and Laila Dybkjær. 1996. Cooperativity in human-machine and human- human spoken dialogue
- Sanda Harabagiu, Dan Moldovan, and Takashi Yukawa. 1996. Testing gricean constraints on a wordnet-based coherence evaluation system. In *Working Notes of the AAAI-96 Spring Symposium on Computational Approaches to Interpreting and Generating Conversational Implicature*, pages 31–38.
- Prathyusha Jwalapuram. 2017. Evaluating dialogs based on Grice's maxims. In *Proceedings of the Student Research Workshop Associated with RANLP 2017*, pages 17–24, Varna. INCOMA Ltd.
- Mohammed R. H. Qwaider, Abed Alhakim Freihat, and Fausto Giunchiglia. 2017. TrentoTeam at SemEval-2017 task 3: An application of Grice maxims in ranking community question answers. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 271–274, Vancouver, Canada. Association for Computational Linguistics.

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

- Motivation:
  bring implicature into pragmatic reasoning in the context of conversations

- Grice dataset: systematically generated using a hierarchical grammar model

- Result: Model shows an overall performance boost in conversational reasoning

Motivation

- *“Language is a form of rational action”*

- *Current* open-ended dialogue systems
  - imitate human responses by regressing large amount of training data
  - fail to account for pragmatics perspective

- *“ Human speakers usually do not speak their thoughts or intentions directly “*
  -> Conversational implicature

Sample Dataset:

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.

Task Definition (How well a model “understands”)

- Implicature recovery
- Convasational reasoning evaluated by QAs

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.

Grammar Production Rules

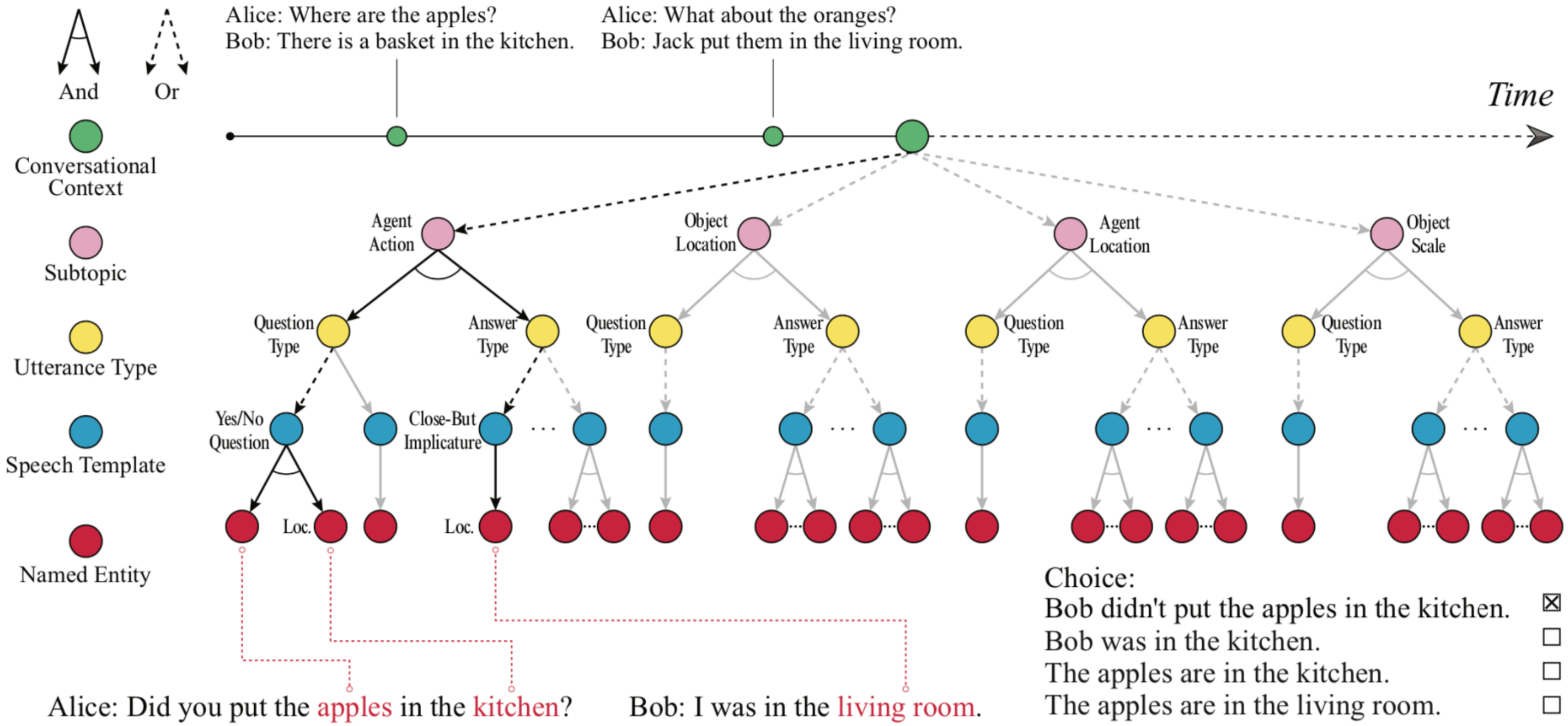


Figure 3: The graphical illustration of the grammar production rules for the GRICE dataset.

Subtopics

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.

five types of implicature

Category	Definition	Example
Relevance	Implicating the answer to an expressed or implied question by stating something related to the answer by implication or explanation.	Alice: Where did you see the apples? Bob: There is a basket in the kitchen. `(The apples are in the kitchen.)`
Strengthening	Implicating a stronger proposition S^+ when not understatement.	Alice: Are some of the apples in the kitchen? Bob: All of them are there. `(Not just some, but all of the apples are in the kitchen.)`
Limiting	Implicating the **denial** of S^+.	Alice: Are all the apples in the kitchen? Bob: Some are. (**Not all** apples are in the kitchen.)
Ignorance	Implicating that one **does not know** whether S^+ is true (or that S^+ may or may not be true).	Alice: Where did you see Jack? Bob: He was in the kitchen or the bedroom. `(I am not sure where Jack was.)`
Close-But	Implicating a **negative** answer to a question by affirming something close to a positive answer in contextually salient respects.	Alice: Did you put the apples in the kitchen? Bob: I was in the living room. `(I did not put the apples in the kitchen since I was in somewhere else.)`

Examples of generating answers

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.

Main take-aways

- Evaluation metric comparing generic answers e.g. “I don’t know” vs. “reference answer” (maxims of quantity)
- Generate conversation templates
    - generate plausible answers using implicatures
    - generate adversarial examples (4 methods mentioned for implicature recovery tasks)

Thank you!

Group 8