A Study of Conversational Implicature in an Out-of-Domain Setting

Wai Ching Leung
Georgetown University
wl607@georgetown.edu

Jessica Lin
Georgetown University
yl1290@georgetown.edu

Abstract

Conversational implicatures are ubiquitous in human conversations but existing dialogue systems have been found to struggle to understand such implicatures. Multiple corpora and models have been developed to automatically identify implicatures from utterances. However, most of the models are tested in an in-domain setting, and it remains a question whether there is a difference between synthetic data vs naturally-occurring in their performance in out-of-domain, naturally occurring data. In this paper, we train BERT-based models using synthetic and/or naturally occurring data, and test them on out-of-domain data collected from real human conversations. We also conduct an error analysis using the Gricean Maxims. Our results show that none of the models generalize well to out-of-domain data, and struggle particularly with implicatures that violate the Maxim of Relevance.

1 Introduction

Introduced by Grice (1975), conversational implicatures are inferences that are derived from the way an utterance is spoken, its context, as well as conversational conventions, going beyond *what* is said. For example, as shown in Figure 1, when speaker A asks 'Are you vegetarian?', and speaker B responds with 'I love burgers too much', speaker B implies he is not vegetarian because burgers usually contains beef. Such an implicature is understood because it is believed that speaker B is providing a relevant response to what speaker A has just asked. In other words, speaker B is complying to the Cooperative Principle (Grice, 1975) in conversations, which governs the understanding of conversational implicatures. See Section 3 for a detailed discussion on the principle.

It has been found that indirect answers are often preferred in real world human conversations as a way to address anticipated follow-up questions (Searle, 1979). While humans have little difficulty

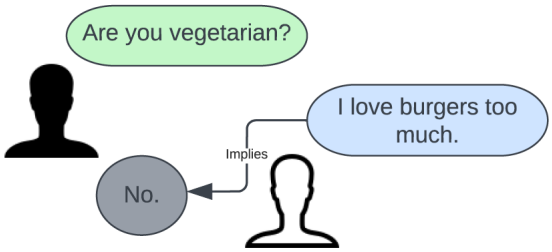

Figure 1: Example of indirect responses

in interpreting implicatures, they pose challenges to computers because implicatures are not tied to surface meanings. Given their ubiquitous nature in human conversations, we believe understanding pragmatic implicatures in natural language is integral to a successful dialogue system as it allows an efficient and natural interaction with humans.

Multiple corpora have been developed from naturally occurring conversations with question-answer pairs (Damgaard et al., 2021; Sanagavarapu et al., 2022; George and Mamidi, 2020) as resources for developing automatic models to determine implicatures in conversations. However, their sizes are relatively small, ranging from around 1,000 to 6,000 question-answer pairs. To address this problem, Louis et al. (2020) create a synthetic corpus with 43,268 pairs of polar questions and indirect responses. Their data are developed via crowdsourcing instead of from natural conversations.

Sanagavarapu et al. (2022) combine the synthetic data from Louis et al. (2020) with naturally occurring data to test a model's performance on determining implicatures in the naturally occurring data. They conclude that synthetic data do not generalize well to naturally occurring data. However, based on their results, one can argue that their results are simply caused by out-of-domain generalization issue, which is a known challenge in machine learning (Koh et al., 2020). In other words, their naturally occurring data might not generalize well to the naturally occurring data from another dataset

either. We believe this is a question worth exploring because it determines whether a model is able to derive implicatures in real world data, i.e data that might have distributions different from training distributions.

To the best of our knowledge, there has not been any research on testing conversational implicature models in a out-of-domain setting. To address this research gap, in this work, we train models using synthetic data and/or naturally occurring data by fine-tuning BERT (Devlin et al., 2018), and test them on naturally occurring data from a different dataset. The main contributions of this study are as follows:

1. We compare the performance of models trained on different types of data (synthetic vs naturally occurring data) on out-of-domain data. This addresses the question whether models trained on naturally occurring data will generalize better than if trained on synthetic data on out-of-domain naturally occurring data

2. We investigate how a model that is trained on both synthetic data and naturally occurring data fair with models only trained on synthetic or naturally occurring data

3. We propose a systematic way to conduct error analysis: we categorize the erroneous predictions made by our models using Gricean Maxims (Grice, 1975), which sheds lights on the strengths and weaknesses of each model in deducing different types of implicatures, and provides understanding of the nature of the naturally occurring data vs synthetic data

2 Related Work

George and Mamidi (2020) create a dataset by transcribing listening comprehension recordings from the English proficiency test TOFEL and dialogues from movies. They manually annotate each indirect response with the implicature 'yes' or 'no'. However, they do not perform any modelling on the data.

Recently, numerous studies (Louis et al., 2020; Sanagavarapu et al., 2022; Damgaard et al., 2021) have experimented transfer learning with indirect answer datasets. Louis et al. (2020) release a crowdsourced corpus 'Circa' which contains 34,268 polar questions and answer pairs. They train multiple models that experiment with using with and without questions or answers during training. They also train various transfer learning models that are first fine-tuned on other datasets, such as BoolQ (Clark et al., 2019) and MNLI (Williams et al., 2017), followed by their own dataset. Their results show that the model that is first fine-tuned on the MNLI dataset followed by their own data performs best with a F1 score of 88.2 for 4 labels ("yes", "no", "conditional yes" and "in the middle"). However, it is not significantly better than the model only fine-tuned on their own data, with a F1 score of 87.8.

Damgaard et al. (2021) introduce a dataset that contains pairs of polar-questions and answers collected from the Friends TV series. They train various CNN models with either GloVe (Pennington et al., 2014) or BERT word representations, and with and without both questions and answers during training. They conclude that the best model is achieved by a BERT-based CNN model that is trained on both questions and answers, with an F1 score of 55.0 on 4 labels ("yes", "no", "yes, "subject to conditions" and "neither"), significantly higher than the majority baseline F1 score of 16.5. They also develop a model that is trained on both Circa dataset (Louis et al., 2020) and their own data, and conclude the adding the Circa data only hurts the model's performance.

Sanagavarapu et al. (2022) develop a dataset with indirect responses from phone transcripts from the **SwDA** data (Jurafsky et al., 1997). They show combining synthetic data from MNLI (Williams et al., 2017) and Circa (Louis et al., 2020) do not have a significant impact on the performance of their BERT-based models in predicting implicatures on their phone conversation data.

Unlike our study, the above research only tests their models in an "in-domain" setting [1], meaning they train and test their models on data from the same datasets. In our work, we test our models in a out-of-domain setting in order to understand how models trained on different types of data can generalize well to data from different sources.

3 Gricean Maxims

Implicature arises when what a speaker intends to mean differs from sentence meaning (Davis, 2019).

[1] We notice that Sanagavarapu et al. (2022) has also tested their model on an "out-of-domain" dataset, but the dataset is collected and annotated by the same authors, which might bias their results

Grice (1975) develops the *Theory of Conversational Implicatures*, in which he shows that *what is meant* often goes beyond *what is said* in conversations and that this extra meaning is systematic and predictable. To give an example of conversational implicatures, consider the following conversation:

A: Were you out with Jennifer last night?
B: I was out drinking with the boys.

In this conversation, B is *implicating* that he was not with Jennifer last night because he was out with the boys instead. The response does not explicitly answer *no*, nor does its literal meaning suggest the answer *no*. As one can see, what is meant goes beyond what is said. Therefore, conversational implicature arises.

Grice (1975) proposes that interlocutors in a conversation are governed by rules in which they try to achieve rational and cooperative communication. The rules are composed of four categories: *the Maxims of Quality*, *the Maxims of Quantity*, *the Maxims of Relevance*, and *the Maxims of Manner*. In the following, we will discuss each category in detail along with examples.

- **Quality Implicatures**: Make your contribution true; do not convey what you believe is false or unjustified.

  A: Tehran's in Turkey, isn't it, teacher?
  B: And London's in Armenia, I suppose. (implicature: Tehran is not in Turkey just like London is not in Armenia)
  In this example, B is not saying what he believes is true, so he is flouting the maxim of Quality. Conversational implicature arises because the addressee understands the speaker flouted the maxim for a reason and infers further meaning from this breach of convention.

- **Quantity Implicatures**: Be as informative as required.

  A: Do you have a sufficient amount of fruit and vegetables for the child?
  B: She only eats milk and cereal. (implicature: She's not having a sufficient amount of fruit and vegetables because she only eats milk and cereal)
  In this example, B does not follow the Maxim of Quantity by providing "redundant" information that the child only eats milk and cereal, instead of just *yes* or *no*. Conversational implicature still arises because A knows B is giving more information than needed on purpose, thus inferring from B's response that the child is not having a sufficient amount of fruit and vegetables.

- **Relevance Implicatures**: Be relevant.

  A: Do you think you got an A on the test?
  B: Do chickens have lips? (implicature: No, it's impossible for me to get an A on the test just like it's impossible for chickens to have lips)
  In this example, B is flouting the maxim of Relevance by not responding something relevant to A's question. Implicature still arises because A assumes B is being relevant and cooperative. Thus, A infers that B is relating chicken having lips to getting an A on the test, which is both impossible.

- **Manner Implicatures**: Be perspicuous; avoid ambiguity. Be brief and orderly.

  A: You got authorization from Aunt Ginny?
  B: I gave her a call like you asked. Very nice woman, we talked for about an hour. (implicature: No, I did not get authorization from her)

  In this example, B is violating the maxim of Manner by giving unnecessary details about the phone call with Aunt Ginny - a simple yes or no would answer the question. Implicature arises in that A knows B is giving too much detail on purpose because the sentence with less detail is false.

In this paper, we categorize all the erroneous predictions made by the models based on Grice's Theory of Conversational Implicatures. More specifically, we look into how each error violates the four maxims. This provides a systematic way to investigate the strengths and weaknesses of each model, and the distinct features of the naturally occurring data and synthetic data.

4 Experiments

4.1 Task

To determine whether a model is able to derive conversational implicature, we formulate the task as determining whether the model is able to predict the underlying direct answer ("yes" or "no"), given a polar question and its indirect response. For example, given the question "Are you vegan?" and

the response “I love burgers too much" as shown in Figure 1, we evaluate if the model is able to predict the direct answer "No".

This task is related to but different from the task of Natural Language Inference (NLI) (MacCartney and Manning, 2008). For the task of NLI, given a premise and a hypothesis, one must determine the relation of the two sentences: *entailment*, *contradiction* and *neutral*. For example, the premise "Two boys are playing in the garden." entails the hypothesis "Two people are outside" because "two boys" entails "two people" and "in the garden" entails "outside". In other words, an inference relation is a logical relation tied to the words and phrases used in the two sentences. On the other hand, conversational implicatures arise from contextual factors and conventions in conversations. For example, given the question "you made pancakes?", the response "grab a plate" implies the answer "yes" (Damgaard et al., 2021). One can only understand such an implicature if they understand the social meaning of grabbing a plate in this scenario. Thus, it is potentially a more challenging task because it requires understanding beyond surface meaning.

4.2 Data

To train our models, we use Circa dataset (Louis et al., 2020) and Friends QIA dataset (Damgaard et al., 2021). For evaluation, we use the Conversational Implicature dataset (George and Mamidi, 2020).

All the data has a triple annotation: *question*, *indirect answer*, and *label*. An example of QA pairs from Friends QIA dataset looks like this:

Monica: You still work at the multiplex? (question)

Chip: Oh, like I'd give up that job! Free popcorn and candy, anytime I want. I can get you free posters for your room. (indirect answer)

label: Yes

Circa dataset (Louis et al., 2020) is a crowdsourced, synthetic dataset that contains 34,268 question-answer pairs with indirect answers to questions written by annotators. Given a dialogue scenario (e.g.,talking to a friend about music preferences), crowdworkers were asked to write yes/no questions. Another set of crowdworkers were then asked to provide indirect answers to the questions. Therefore, this dataset does not include naturally occurring utterances and was created under specific, topic-restricted settings. The data are labeled with a strict scheme with 8 labels as well as a relaxed scheme with 4 labels (*Yes, No, Conditional Yes, Middle*). In our study, we collapse the relaxed scheme into binary labels *Yes, No* by merging the two categories *Yes* and *Conditional Yes*, and removing *Middle*. As a result, there are a 32,044 pairs, with 19,211 *Yes*, and 12,833 *No*.

Friends QIA dataset (Damgaard et al., 2021) contains transcripts of 5,930 question-answer pairs, which involve indirect responses to yes-no questions from the Friends TV episodes. Different from Circa (Louis et al., 2020), this dataset is obtained from a more open domain (TV show) and includes a broader context. The labels consist of six classes (*Yes, No, Conditional Yes, Neither Yes nor No, Other, NA*). In our final training data, we collapse the 6-class labels into binary labels *Yes, No* by dropping examples from the *Neither Yes nor No, Other, NA* classes and merging the *Conditional Yes* class with the *Yes* class. The final dataset contains a total of 4,580 question-answer pairs, with 3045 *Yes* and 1535 *No*.

Conversational Implicature dataset (George and Mamidi, 2020) contains 1,001 dialogue snippets collected from listening comprehension tasks from English proficiency tasks and dialogues from movies. As a preprocessing step, we removed data that does not contain polar question, i.e., wh-questions or non-questions. The final test set contains a total of 733 question-answer pairs, with 380 *Yes* and 353 *No*.

4.3 Approaches

In our experiment, we develop the following three models by fine-tuning BERT:

1. Circa Model: fine-tuning BERT using only the Circa dataset (Louis et al., 2020).

2. Friends Model: fine-tuning BERT using only the Friends dataset (Damgaard et al., 2021).

3. Transfer Model: first fine-tuned on the Circa data (Louis et al., 2020), followed by the Friends data (Damgaard et al., 2021).

In order to compare and evaluate how well the models can generalize to out-of-domain data, we first validate the models using validation data from the same datasets, i.e., Freinds model is trained on the train-split of the Friends data, and validated on the validation split of the same dataset, and then

test the models on the Conversational Implicature dataset. (George and Mamidi, 2020).

5 Results and Analyses

5.1 Results

As shown in Table 1, all the models significantly perform better on the in-domain validation datasets than on the out-of-domain test data. The Circa model shows a 50% improvement from the baseline accuracy of 0.6, whereas the Friends model improves by 18% from the baseline accuracy of 0.67. The Transfer model's accuracy only improves from the baseline by around 6%.

A potential reason that can account for the considerably higher performance of the Circa model on its own validation data is that the Circa data might show more consistencies because it was developed via crowd-sourcing. In real world conversations, there might be more variations in how people respond to polar questions indirectly. This also explains why the Friends model has relatively less improvement in validation due to higher level of inconsistency in training data.

Similar to findings from previous research (Louis et al., 2020; Sanagavarapu et al., 2022), combining data from different sources might not enhance performance in predicting conversational implicature. The Transfer model was first trained on synthetic data (Louis et al., 2020), then on naturally occurring data (Damgaard et al., 2021), and tested on another subset of the naturally data for validation. This only hurts the model's performance since the accuracy decreases from 0.79 to 0.71, and F1 from 0.75 to 0.64.

The majority baseline accuracy for the test dataset (George and Mamidi, 2020) is 0.52, and all the models' performance is similar to the majority baseline. Although both data from Friends (Damgaard et al., 2021) and test data from George and Mamidi (2020) are collected from naturally occurring conversations, the Friends model does not perform significantly better than the other two models. Therefore, models trained on naturally occurring data might not necessarily generalize better to naturally occurring data from another dataset.

However, the considerably smaller size of Friends data (4,580 pairs) vs Circa data (32,044 pairs) could also account for the results since neural network requires large amount of data for proper training. In this case, there might not be sufficient training data for the model to learn the features of conversational implicatures in real conversations, hence struggling to make correct predictions on the test data. It would be worth exploring in future research to investigate if a model trained on sufficient data from naturally occurring conversations can generalize well to natural conversations from another dataset.

5.2 Error Analysis

We conduct a manual qualitative analysis of 200 randomly picked errors made by our models by categorizing them using the Gricean Maxims of Conversational Implicature discussed in Section 3. Each error is categorized into one of the four maxims that the conversation violates. In addition to the four maxims, we also include the implicature types (metaphor, hyperbole, idiom) of the cases in which the models have difficulty predicting. These are the implicature types that are found to be prevalent in the errors. Furthermore, we look into the type(s) of maxims that each model struggles with most. This will provide insights into the strengths and weaknesses of each model as well as the differences between naturally occurring data and synthetic data.

First, Figure 2 shows the distribution of errors made by at least one model in our experiment. The most common maxim that the models have difficulty predicting is the *Maxim of Relevance*. As shown in (1), the indirect response is irrelevant to the question, thus it violates the *Maxim of Relevance*.

(1) Q: Did he pass the exam?
A: Does a monkey build a house?
(Labels: No; Prediction: Yes)

Those cases that violate the *Maxim of Relevance* are particularly hard for the models because (i) the lack of relevance between questions and answers potentially leads to a lack of association between linguistic cues and the sentences. Since "irrelevant" answers can potentially be in any domains, it can be challenging for models to find a consistent distribution for responses in this category, which is often required for models to make inferences. (ii) Models might need world knowledge to infer from the responses. In this example, the models need to understand the common knowledge that "it is not easy for a monkey to build a house".

As can be seen in Figure 3, following the *Maxim of Relevance*, implicatures in the *Other* category

Models	Baseline Results		Validation Results		Test Results	
	Validation Accuracy	Test Accuracy	Accuracy	F1	Accuracy	F1
Circa	0.60		0.91	0.90	0.50	**0.50**
Friends	0.67	0.52	0.79	0.75	**0.53**	0.43
Transfer	0.67		0.71	0.64	0.52	0.46

Table 1: Model results on the validation and test datasets. For each model, the training and validation data come from the same dataset. For example, the Friends model was trained on 90% and 10% of the data from Damgaard et al. (2021). All the models are tested on a out-of-domian dataset from George and Mamidi (2020).

are also found to be challenging, particularly for the Circa model (Figure 3a) and the Transfer model (Figure 3c). (2) shows a typical example in the *Other* category:

(2) Q: Would you like some milk in your coffee?
A: Please
(Label: Yes; Prediction: No)

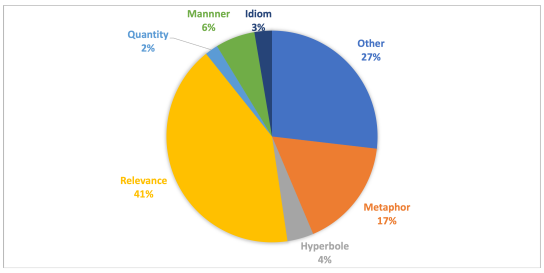

Figure 2: The distribution of errors made by at least one model

This response in (2) does not violate any of the Gricean Maxims nor does it belong to any of the implicature types. Rather, it shows an example of *Generalized conversational implicature* (Grice, 1975), where an implicature does not depend on a specific context. Humans would have little difficulty understanding that “please" implies “yes" to a yes-no question, regardless of the context. However, machines seem to lack this kind of commonsense knowledge, resulting in their inability to correctly identify the implicature behind it. It is also found that models trained on the Circa dataset (the Circa model and the Transfer model) particularly find it hard to identify this type of implicature. A potential reason that can account for this result is that this type of implicatures is prevalent in real human conversations, but not in the synthetic ones.

Another typical case in the Other category is shown in (3):

(3) Q: Did you call Carl to the concert?
A: I tried several times.
(Label: Yes; Prediction: No)

(3) belongs to the category of *particularized conversational implicature* (Grice, 1975). Unlike *Generalized conversational implicature*, this type of implicature depends on both the context and the utterance itself. In (3), the action of “trying several times" does not ordinarily convey anything about “calling Carl", so the implicature is highly dependent on the context and the previous utterance “Did you call Carl to the concert?" Models trained on synthetic data find it hard to identify this type of implicature, although it is extremely common in human conversations.

As shown in Figure 3b and 3d, metaphor is also a common type of implicature that all our models have a hard time identifying. An example of metaphor is shown in (4):

(4) Q: Are you able to carry the box?
A: It is as light as a feather.
(Label: Yes; Prediction: No)

To correctly identify the implicature in (4), models have to first understand the metaphor “as light as feathers". Then, the model can infer from the response that the speaker is able to carry the box because it is light. This result (see Figure 3d) aligns with previous research results that show understanding dialogues with figurative language (e.g., irony, idiom, hyperbole) is still a challenging task for current dialogue systems (Jhamtani et al., 2021). Even for the Friends model, which is trained on real human conversations, it does not perform well on this type of implicature(see Figure 3b). We hope this result can motivate more research on dialogue systems to look into figurative language in conversations.

To summarize, we have found that all the models

(a) Error type distribution for the Circa model

(b) Error type distribution for the Friends model

(c) Error type distribution for the Transfer model

(d) Error type distribution for all models

Figure 3: Error type distribution for each model

struggle to predict implicatures from the *Maxim of Relevance* and *Metaphor* categories. Additionally, our results show that the model trained on the Circa dataset (synthetic dataset) has difficulty in identifying generalized and particularized conversational implicatures, while all three models, especially the one trained on naturally occurring dataset find it particularly hard to understand cases that involve metaphor. Future research on how current dialogue systems can deal with these difficult cases would be beneficial.

6 Conclusion

As indirect answers are prevalent in human conversations, understanding implicature arising from these conversations is crucial for a successful dialogue system. In this paper, we trained three models using two types of data (synthetic vs naturally occurring data) by fine-tuning BERT, and tested them in an out-of-domain setting.

Our results showed that there is a significant degradation when testing the models on an out-of-domain data. Although the out-of-domain test data is collected from naturally occurring conversations, the Friends model (which is also trained on natural conversations) does not show a better performance than the Circa model that is trained on synthetic data. Therefore, models trained on naturally occurring data do not necessarily generalize better to naturally occurring data. Similar to previous findings, we have also found that training a model on combined data (synthetic and naturally occurring) does not improve its performance.

We have also found that the Circa model shows significantly better on validation, which we speculate could be due to the higher consistencies that exist in the synthetic data than in real world conversations.

However, we acknowledge that the Friends dataset is considerably smaller than the Circa dataset (4,580 instances vs 32,044 instances). Therefore, it might not be a completely fair comparison as a way to compare the performance of synthetic data vs naturally occurring data. In our future research, we will experiment with synthetic and naturally occurring datasets with similar sizes for a more robust comparison.

Finally, we found that the *Maxim of Relevance* is the most challenging category for all the models. In general, the synthetic model struggles more with Generalized/particularized conversational implicature, and the naturally occurring model more with implicatures that involve metaphor. Future research on how dialogue systems can deal with these difficult cases in conversations would be beneficial.

References

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Cathrine Damgaard, Paulina Toborek, Trine Eriksen, and Barbara Plank. 2021. "i'll be there for you": The one with understanding indirect answers. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 1–11.

Wayne Davis. 2019. Implicature. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2019 edition. Metaphysics Research Lab, Stanford University.

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*.

Elizabeth Jasmi George and Radhika Mamidi. 2020. Conversational implicatures in english dialogue: Annotated dataset. *Procedia Computer Science*, 171:2316–2323.

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

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs. *arXiv preprint arXiv:2110.00687*.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth

Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 88–95. IEEE.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2020. Wilds: A benchmark of in-the-wild distribution shifts.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. " i'd rather just go to bed": Understanding indirect answers. *arXiv preprint arXiv:2010.03450*.

Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK. Coling 2008 Organizing Committee.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Krishna Sanagavarapu, Jathin Singaraju, Anusha Kakileti, Anirudh Kaza, Aaron Mathews, Helen Li, Nathan Brito, and Eduardo Blanco. 2022. Disentangling indirect answers to yes-no questions in real conversations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4677–4695.

John R. Searle. 1979. *Indirect speech acts*, page 30–57. Cambridge Universit y Press.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference.