Disentangling Indirect Answers to Yes-No Questions in Real Conversations

Krishna C. Sanagavarapu[1], **Jathin P. Singaraju**[2], **Anusha Kakileti**[2],
Anirudh Kaza[2], **Aaron A. Mathews**[2], **Helen Li**[2], **Nathan R. Brito**[2], **Eduardo Blanco**[1]
[1]Arizona State University, [2]University of North Texas,
{ksanagav, eduardo.blanco}@asu.edu, {pran.singaraju, arkaza820}@gmail.com,
{akakileti, AaronMathews, helenli, Nathanbrito}@unt.edu

Abstract

In this paper, we explore the task of determining indirect answers to yes-no questions in real conversations. We work with transcripts of phone conversations in the Switchboard Dialog Act (SwDA) corpus and create SwDA-IndirectAnswers (SwDA-IA), a subset of SwDA consisting of all conversations containing a yes-no question with an indirect answer. We annotate the underlying direct answers to the yes-no questions (yes, probably yes, middle, probably no, or no). We show that doing so requires taking into account conversation context: the indirect answer alone is insufficient to determine the ground truth. Experimental results also show that taking into account context is beneficial. More importantly, our results demonstrate that existing corpora with synthetic indirect answers to yes-no questions are not beneficial when working with real conversations. Our best models outperform the majority baseline by a substantial margin, but the task remains a challenge (F1: 0.46).

1 Introduction

Dialogue systems have become a reality enabled by large datasets and deep neural networks. Task-oriented (Wen et al., 2017) and open-domain (Tang et al., 2019) dialogue systems are commonplace. Neural approaches are popular (Zhang et al., 2018) although systems based on logical inference and rules outperform networks in open-domain dialogue (Finch et al., 2021). State-of-the-art systems face challenges keeping track of a conversation and avoiding inconsistencies (Welleck et al., 2019). Evaluation is also an open issue as automated metrics have several drawbacks (Liu et al., 2016). An alternative is to evaluate dialogue systems based on whether they can collaborate with humans to solve a problem (Lewis et al., 2017) or elicit some action such as donating to a cause (Wang et al., 2019).

People do not explicitly say what they mean when they speak to each other yet they seamlessly carry on conversations. For example, customers asking *Where are the $1 cups?* reveal to the sales associate that they want to buy a (cheap) cup (Tatu, 2005). In this paper, we investigate the underlying direct answers to yes-no questions. A yes-no question is a question that expects a *yes* or *no* for an answer. As we shall see, we work with direct yes-no questions (e.g., Did you drive yourself to the airport) and indirect ones (e.g., I am not sure if you drove yourself to the airport).

A: Do you work outside of the home?
B: No, I am not working currently.
Underlying direct answer: No

A: Do you work outside of the home?
B: Uh, last month I left my company.
Underlying direct answer: No

A: Do you work outside of the home?
B: Uh, last month I left my company.
A: What happened? Stress?
B: But now I work for a marketing firm where I travel a lot.
Underlying direct answer: Yes

Figure 1: Conversation snippets with a yes-no question (first turn by Sparker A). In the first example, the underlying direct answer is obvious given the answer by Speaker B. In the second example, determining the underlying direct answer requires commonsense knowledge. In the third example, it requires not only commonsense but also taking into account more than the turn by Speaker B immediately following the question.

Consider the conversation snippets in Figure 1. As shown in the first one, the conversation turn following a yes-no question may be a direct answer (i.e., a turn including *yes*, *no*, *obviously*, *never* or similar keywords). Indirect answers (e.g., second snippet) are more common. Speaker B *leaving his company a month ago* does not entail that he is jobless (and thus not working—at home or the office). Given this indirect answer alone, however, it is reasonable to conclude so thus the underlying direct answer is *no*. The broader conversation context often provides a more complete picture and the

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4677 - 4695
July 10-15, 2022 ©2022 Association for Computational Linguistics

true underlying direct answer (e.g., third snippet). After a brief interjection by Speaker A, Speaker B states that he changed jobs and now travels a lot. Thus, the underlying direct answer is *yes*.

The focus of this paper is to determine the underlying direct answers to yes-no questions. Unlike previous efforts, we work with transcripts of *real conversation* as opposed to yes-no questions and indirect answers written by annotators on demand. The main contributions are:[1]

1. We present SwDA-IA, a corpus consisting of the 2,544 yes-no questions in Switchboard (Jurafsky et al., 1997, SwDA) with *I*ndirect *A*nswers and their underlying direct answers.
2. We show that determining underlying direct answers requires context beyond the yes-no question and the next conversation turn. Indeed, the ground truth changes depending on whether we show annotators context.
3. We demonstrate that transfer learning with related tasks as well as synthetic yes-no questions and synthetic indirect answers only bring modest improvements. Specifically, synthetic data do not transfer to real conversations.
4. We provide insights into the most common errors made by our best performing model.

The work presented here provides evidence that determining the underlying direct answer to yes-no questions *in real conversations* is much more challenging than in synthetic data. Our best model obtains 0.49 F1 even when using only the context handpicked as important by annotators.

2 Previous Work

Yes-no questions and indirect answers have been studied for decades (Green and Carberry, 1999). Hockey et al. (1997) report that 27% of all questions in 18 hours of spontaneous speech are yes-no questions with indirect answers (Rossen-Knill et al., 1997). Indirect answers are often used to ask follow-up questions or provide explanations for negative answers (Stenström, 1984), prevent wrong interpretations of direct answers (Hirschberg, 1985) or show politeness (Brown and Levinson, 1978).

Yes-no questions have received considerable attention recently. Outside of the dialogue genre, Clark et al. (2019) present BoolQ, a corpus of 16,000 yes-no questions submitted to a search engine (e.g., *Does France have a Prime Minister and a President?*) along with Wikipedia articles that may contain an answer. Two recent corpora consist of dialogues generated by crowdworkers on demand after being given some text to talk about: QuAC (Choi et al., 2018) and CoQA (Reddy et al., 2019). Unlike the work presented here, the dialogues in these corpora are limited to questions and answers about the text provided to crowdworkers. More importantly, we work with unrestricted, natural conversations rather than dialogues between paid annotators who are asked to have a conversation that resembles a questionnaire geared towards checking for understanding of a piece of text.

Yes-no questions in real conversations have been studied before at a small scale. de Marneffe et al. (2010) work with 224 questions-answer pairs involving gradable adjectives. We borrow from them the five labels to annotate underlying direct answers. de Marneffe et al. (2009) present a typology for 623 yes-no questions from SwDA. Our corpus is four times larger. Unlike them, we show that context is crucial to determine underlying direct answers and present experimental results.

The work by Louis et al. (2020) is the closest to ours. They present Circa, 34,268 yes-no questions and indirect answers written by crowdworkers given one of ten scenarios (e.g., scenario: Talking to a friend about music preferences, Q: Do you like guitars? A: I practice playing every weekend). Unlike us, Circa assumes that the turn following a yes-no question is enough to determine the underlying direct answer. Furthermore, we show that (a) determining underlying direct answers to yes-no questions is much harder in *real conversations* and (b) their synthetic data only bring minor improvements when working with real conversations.

3 SwDA-IA: A Corpus of Yes-No Questions and Indirect Answers

We present SwDA-IA, a corpus consisting of (a) the 2,544 yes-no questions in Switchboard with *I*ndirect *A*nswers and (b) manual annotations indicating the underlying direct answers. Three characteristics set our work apart. First, we work with real conversations as opposed to artificial, synthetic ones (Section 2). Second, we show that determining underlying direct answers requires taking into account context beyond the yes-no question and the next turn. Third, annotators pinpoint which turns within the context around a yes-no-question are useful to determine the underlying direct answer.

[1] Corpora available at `https://github.com/krishna-chaitanya-sanagavarapu/SwDA-IA`.

3.1 Collecting Yes-No Questions with Indirect Answers

While we could work with any dialogue corpora, we chose SwDA (Jurafsky et al., 1997) because it contains unrestricted conversations and the dialogue acts annotations simplify the process of identifying yes-no questions with indirect answers. SwDA contains transcripts of 1,155 five-minute phone conversations. In these conversations, two people talk about topics such as childhood, recycling, and news media. The conversations are unrestricted and often divert from the initial topic. 440 speakers participated in the corpus creation resulting in 122,646 utterances. We use the distribution in ConvoKit (Chang et al., 2020) for convenience.

After manually examining all the dialogue acts, we select all conversation turns containing a dialogue act that indicates a yes-no question (see the list in the supplementary materials). This step results in 5,846 yes-no question including indirect ones (e.g., "I don't know if you are familiar with that issue." is an indirect version of "Are you familiar with that issue?"). After selecting yes-no questions, we discard those that are followed by a turn containing a direct answer. To do so, we discard turns containing a dialogue act indicating direct answers (see list in the supplementary materials). After manually examining the 2,542 questions that are not filtered, we observed that some have a direct answer. In order to avoid them, we also consider a direct answer any turn that contains *yes*, *yea*, *yeah*, *no way*, *nope*, *never*, *sure*, *right*, *you bet*, *of course*, *certainly*, *maybe*, *definitely*, or *uh huh*.

These steps select 2,376 turns with 2,544 yes-no questions followed by an indirect answer (168 turns have more than one yes-no question). Only 4.77% of all utterances in SwDA are a yes-no question. However, 43.52% of all yes-no questions have an indirect answer (2,544 out of 5,846). More importantly, 77.4% of the 5-minute conversations in SwDA contain at least one yes-no question with an indirect answer. The supplementary materials provide an analysis of the yes-no questions we work with. For example, 41.4% are indirect (e.g., do not include a question mark).

3.2 Annotating Underlying Direct Answers

In order to determine the underlying direct answers to the 2,544 yes-no questions, we manually annotated them. Our label set consists of five options: *(definitely) Yes*, *Probably Yes*, *(in the) Middle*, *Probably No* and *(definitely) No*. *Middle* is chosen when annotators do not lean towards any of the other four labels (e.g., "A: Do you have kids? B: Do I have kids?". These labels are heavily inspired by de Marneffe et al. (2010). Louis et al. (2020) include a few more options (e.g., "I am not sure", "In the middle, neither yes or no" and "Other") but they report very low frequencies (0.2–1.9%). These three labels are included in our "Middle."

We found two common scenarios that require additional explanation in order to be consistent:

- If the answer is *yes under certain conditions* or *sometimes yes*, annotators are instructed to choose *Probably Yes*. For example, the correct label for "A: Do you travel a lot for pleasure? B: We try to make one trip per year if I find a good sale" is *Probably yes*.
- If the yes-no question contains a negation, *Yes* and *Probably Yes* have their meaning reversed. For example, the correct label for "A: You didn't move to Alaska, right? B: I have been in Alaska for 13 years now" is *No* (*Yes* would be correct if B had not moved to Alaska).

Annotation Process and Quality The annotation process took place in two independent phases in order to investigate the role of context in determining underlying direct answers. In the first phase, annotators were only shown the yes-no question and the conversation turn immediately after (i.e., the indirect answer). In the second phase, annotators were shown three turns before and after the yes-no question as context. Different annotators annotated the same question in each phase to avoid potential biases (i.e., recalling answers from the previous phase). In both phases, the interface showed conversation turns after brief delays in order to encourage annotators to read the conversation snippet in order (either two turns or seven turns). In addition to selecting the underlying direct answer, in the second phase annotators also tagged the conversation turns that help them determine it. We hosted the annotation interface internally and recruited in-house annotators. Seven native English speakers participated in the annotations.

Inter-Annotator Agreement After refining the explanation for each label and group discussions, we conducted a pilot with 200 questions and the seven annotators. The average weighted Fleiss' Kappa (Plewis and Unit, 1982) between all pairs of annotators was 0.80. Given the high agree-

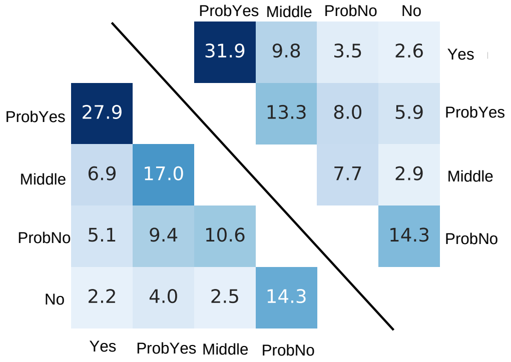

Figure 2: Heatmaps of the inter-annotator disagreements when annotators have access to (a) only the yes-no question and indirect answer (bottom left) and (b) conversation context. Numbers indicate percentages. Most disagreement are minor: between *ProbYes* and either *Yes* or *Middle*, or *ProbNo* and either *No* or *Middle*.

ment (Landis and Koch, 1977), the remaining 2,344 questions were divided into twelve batches, and two annotators annotated each batch. The average Weighted Fleiss' Kappa score for all the batches was 0.81 when not showing context to annotators (first phase) and 0.78 when showing context (second phase). Disagreements in both phases were adjudicated after manual examination. The "true" underlying direct answers are the one annotated when context is shown to annotators, and those are the ones we conduct experiments with (Section 5).

Figure 2 shows the percentage of all disagreements prior to the adjudication step when context is not and is shown to annotators (top right and bottom left). As expected given the high Kappa coefficients, most disagreements are minor: between (a) *Yes* (or *No*) and *Probably Yes* (or *Probably No*) or (b) *Middle* and *Probably Yes* or *Probably No*.

4 Corpus Analysis: the Role of Context

The annotations we collected in both phases differ substantially. Thus context beyond the yes-no question and indirect answer (i.e., the following turn) is needed to determine the ground truth in real conversations. To our knowledge, previous work does not take into account context (Section 2).

Table 1 presents the distribution of each label when the interface shows and does not show context to annotators. The major change is that almost half of the questions annotated with *Middle* without context are annotated with one of the other four labels with context. Note that out of the five labels, *Middle* is arguably the one that reveals the least information about the speakers.

	% without context	% with context
Yes	35.7	42.4
ProbYes	14.9	19.5
Middle	28.0	14.9
ProbNo	7.7	9.9
No	13.6	13.3

Table 1: Label distribution when annotators have access to (a) only the yes-no question and indirect answer (without context), and (b) conversational context (3 turns before and after the question, with context).

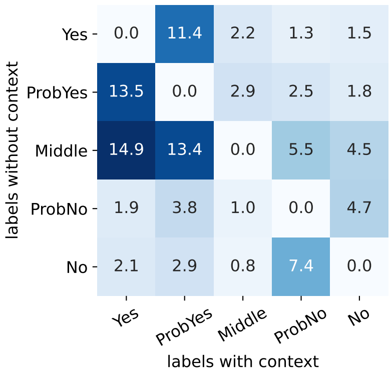

Figure 3: Heatmap of the changes in ground truth depending on whether annotators have access to context in addition to the question and indirect answer. Numbers indicate percentages. Context is crucial to determine the underlying direct answer to a yes-no question.

Figure 3 shows how the ground truth changes when we show context to annotators. The most common change is from *Middle* to either *Yes* (14.9%) or *ProbYes* (13.4%). In other words, context allows annotators to select an underlying direct answer that is more meaningful. There are also many changes from *ProbYes* to *Yes* (13.5%) and vice versa (13.5%), suggesting that context provides further details (a clarification, condition, etc.) to determine the underlying direct answer.

We show examples of changes in ground truth when annotators do not have and have access to context in Table 2. In the first example (top left), the indirect answer (t_1) repeats the question and does not provide any clue about the underlying direct answer. The next turn by A (t_3), however, leaves no doubt: the underlying direct answer is *Yes*. The second example (top right) shows a similar pattern, but it exemplifies negation in the question and uncertainty. Speaker A does not commit to a *Yes* (Yes, It didn't kill anything) but rather sharing that to his knowledge it has not (*Probably Yes*).

The third example (bottom left) shows how follow-up questions in context can clarify the under-

t_{-3}, A: Which is like twenty minutes away.
t_{-2}, B: Right.
t_{-1}, A: But, uh, we don't have any fast foods here in this small city.
t_0, B: *That is probably very fortunate. Do you have kids?*
t_1, A: *Do I have kids?*
t_2, B: Yeah.
t_3, A: Well, I have a son, but he's grown up.

Underlying direct answer ...
without context: Middle; *with context*: Yes

t_{-3}, A: And it seem
t_{-2}, B: I bank at NCNB and they have a number that you can call. I always call in once every other week and check what checks have cleared
t_{-1}, A: Yeah.
t_0, B: *Do you do that?*
t_1, A: *We have the same thing.*
t_2, B: You do too?
t_3, A: Yeah.

Underlying direct answer ...
without context: ProbYes; *with context*: Yes

t_{-3}, A: Um, not bad.
t_{-2}, B: Yeah
t_{-1}, A: Occasionally you'll have a mound pop up, but that is expected.
t_0, B: *What else did it, it didn't kill anything right?*
t_1, A: *Um.*
t_2, B: Not really.
t_3, A: As far as I know it hasn't killed anything. Even the area of the grass that was underneath and around the mounds.

Underlying direct answer ...
without context: Middle; *with context*: ProbYes

t_{-3}, A: So, uh, you know, you need to go to a school that handles whatever it is you want to do.
t_{-2}, B: Yeah. Where did you go to school?
t_{-1}, A: Uh, University of Mississippi.
t_0, B: *Oh. Was that local*
t_1, A: *Uh, well, it was, well, within the state.*
t_2, B: Uh huh.
t_3, A: But it, it was not necessarily local.

Underlying direct answer ...
without context: Yes; *with context*: No

Table 2: Examples of yes-no questions (t_0) with indirect answers (t_1). Determining the underlying direct answers requires taking into account more conversation context than the question and indirect answer: the ground truth annotated by humans changes depending on whether they also have access to context (t_{-3}, t_{-2}, t_{-1}, t_2, and t_3).

lying direct answer of a yes-no question (t_0; t_2 is also a yes-no question). Given t_1 alone, annotators selected *Probably Yes* as there is some uncertainty about what *same thing* refers to. The following two turns (t_2, t_3) make it clear that the underlying direct answer is *Yes*. This example also shows how context *before* the yes-no question (t_{-2}) is sometimes beneficial. The fourth example (bottom right) shows how context can flip the underlying direct answer. The definition of *local* is open to discussion (Is anything within the state local?), but after reading the t_3 it is clear that according to Speaker A the university he went to is not local.

Which turns around a yes-no question are the most important? We show which turns annotators selected as useful to determine the underlying direct answers in Figure 4. Unsurprisingly, the turn immediately after the yes-no question (after$_1$) is almost always useful (96.4%). The next turns by either speaker are often useful, including the follow-up turn by the speaker asking the yes-no question (after$_2$, 20.9%). The turns before the yes-no question (prev$_3$–prev$_1$) are less often important. We also found that the turn *before* the yes-no question is the most useful (prev$_1$, 12.5%) out of all the turns before the question.

Note that annotators almost always (96.4%) deem useful the turn immediately following the question (after$_1$, the indirect answer). This indicates that the indirect answer is rarely an interjection or filler (e.g., *Uhm*, *Really?*). Additionally, annotators selected at least one of the other turns in the context in 50% of questions, indicating that it is often the case that at least two turns are beneficial to determine the underlying direct answer.

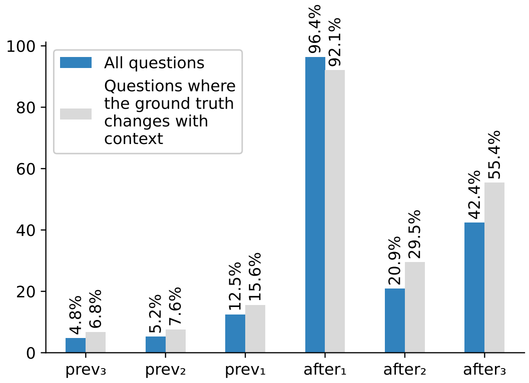

Figure 4: Percentage of times annotators rely in the context to determine the underlying direct answer to a yes-no question (three turns before and after; after$_1$ refers to the indirect answer).

5 Experiments and Discussion

We split SwDA-IA into training, development and test splits (70/15/15) randomly but making sure no split contains yes-no questions from the same conversation. We consider each conversation snippet (yes-no question, three turns before, and three turns after) an instance, and build models to predict the underlying direct answer. Following previous work on (short) text classification and specifically on the same task with synthetic data (Louis et al., 2020), we build transformer-based classifiers fine-tuned using several strategies. We experimented with three transformers: BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and TOD-BERT (Wu et al., 2020) released by HuggingFace (Wolf et al., 2020). Note that the first two are pretrained with general-purpose English while the third is fine-tuned for task-oriented dialogues and includes an attention mechanism designed to keep track of dialogues. While the three transformers obtained roughly the same results (within 0.02 F1), RoBERTa outperformed the other two and we only report results with RoBERTa.

The ground truth depends on context beyond the yes-no question and indirect answer (Section 4). A RoBERTa-based classifier, however, may not benefit from context. In order to experiment with context, we conduct three types of experiments: feeding to the models (a) only the question and the turn immediately following the question (i.e., indirect answer), (b) the full context (i.e., seven turns), and (c) only the annotator-selected turns during the annotation process. We use a separator token to indicate a new turn in the input sequence.

Fine-Tuning RoBERTa We fine-tuned the RoBERTa-based classifier using corpora for related tasks and SwDA-IA, our corpus. Specifically, we used the following corpora in our experiments:

MNLI is a corpus for natural language inference (Williams et al., 2018). It contains premise-hypothesis pairs where the premise entails, is neutral with respect to, or contradicts the hypothesis. It contains 392k pairs for training, 9k for development, and 9k for test. The fine-tuned RoBERTa classifier obtains 83% development accuracy. Following Louis et al. (2020), we rewrite questions into declarative statements and map *entailment* to *Yes*, *neutral* to *Middle*, and *contradiction* to *No*.

BoolQ is a corpus for yes-no question answering (Clark et al., 2019). Questions were submitted to a search engine and Wikipedia articles containing and not containing the answer are included for each question. It contains 9.4k questions for training, 3.2k for development, and 3.2k for test. The fine-tuned RoBERTa classifier with BoolQ obtains 73% development accuracy. We map their *Yes* and *No* (correct and incorrect) to our *Yes* and *No*.

Circa is a corpus with yes-no questions and indirect answers (Louis et al., 2020). Unlike SwDA-IA, which includes questions and answers from real conversations, Circa collected questions and answers from crowdworkers who were given one of ten scenarios (Section 2). In other words, Circa does not include naturally occurring questions and answers. Additionally, Circa does not consider context as all answers are a single turn. Circa contains 3,431 yes-no questions and up to 10 indirect answers for each (total: 34,268 question-answer pairs). The fine-tuned RoBERTa classifier with Circa obtains 78% development accuracy. We map their *Middle* and *I am not sure* to our *Middle*.

Our Corpus: SWDA-IA We consider three versions of fine-tuning with our corpus: taking into account only the *Q*uestion, only the *A*nswer, or both (*QA*). Intuitively, the indirect answer should be the most useful to determine the underlying direct answer but the question may also help.

5.1 Results

Table 3 presents results fine-tuning with each corpus and the best performing combinations of two or more corpora. The supplementary materials provide results with all combinations and detailed results (P, R, and F1) for each label.

Training with the full context around the yes-no question (three turns before and three turns after) yields worse results than only training with the yes-no question and turn immediately after (top block vs. middle block). This may seem surprising, however, it is known that keeping track of a conversation across several turns is challenging (Kim et al., 2020). Regardless of whether the model considers context, we observe that MNLI does not transfer to our task (0.13 and 0.12 with and without context, lower than the majority baseline) and Circa is the most useful out of the three previous corpora (MNLI, BooLQ, and Circa; 0.32 and 0.27 F1).

Fine-tuning with combinations of previous corpora does not surpass the 0.32 F1 obtained with Circa. Fine-tuning with SwDA-IA, however, brings substantial improvements. In particular, fine-tuning with SWDA-IA_QA in addition to combinations of previous corpora yields statistically significant higher results (indicated with †; best: 0.46 F1).

These results lead to the conclusion that determining indirect answers to yes-no questions in real

	All labels			Yes	ProbYes	Middle	ProbNo	No
	P	R	F1	F1	F1	F1	F1	F1
Majority Baseline	0.18	0.42	0.25	0.59	0.00	0.00	0.00	0.00
RoBERTa without context and tuned with …								
MNLI	0.24	0.20	0.12∗	0.32	0.00	0.14	0.00	0.00
BoolQ	0.20	0.35	0.25	0.50	0.00	0.00	0.00	0.00
Circa	0.31	0.34	0.32	0.58	0.04	0.05	0.03	0.35
MNLI+Circa	0.20	0.23	0.21	0.42	0.09	0.00	0.04	0.14
MNLI+BoolQ+Circa	0.26	0.29	0.24	0.50	0.00	0.20	0.04	0.03
SwDA-IA_Q	0.18	0.43	0.26	0.45	0.18	0.22	0.22	0.18
SwDA-IA_A	0.42	0.45	0.43∗	0.60	0.29	0.40	0.09	0.31
SwDA-IA_QA	0.44	0.45	0.44∗	0.62	0.31	0.33	0.08	**0.44**
MNLI+SwDA-IA_QA	0.46	0.43	0.45∗†	0.58	0.25	0.44	0.16	0.36
BoolQ+Circa+SwDA-IA_QA	0.46	0.48	**0.46∗†**	**0.64**	0.22	**0.45**	**0.24**	0.42
MNLI+BoolQ+Circa+SwDA-IA_QA	0.41	0.40	0.40∗†	0.54	0.30	0.34	0.20	0.34
RoBERTa with full context and tuned with …								
MNLI	0.19	0.16	0.13∗	0.12	0.00	0.20	0.00	0.00
BoolQ	0.20	0.18	0.10∗	0.32	0.00	0.00	0.00	0.00
Circa	0.24	0.28	0.27	0.55	0.03	0.00	0.00	0.20
BoolQ+Circa	0.24	0.33	0.28	0.50	0.24	0.00	0.16	0.07
MNLI+BoolQ+Circa	0.21	0.26	0.23	0.34	0.28	0.08	0.00	0.10
SwDA-IA_Q	0.32	0.39	0.33	0.57	0.08	0.26	0.02	0.24
SwDA-IA_A	0.38	0.40	0.38	0.53	0.22	0.26	0.12	0.38
SwDA-IA_QA	0.41	0.43	**0.43∗**	**0.59**	0.24	**0.32**	0.14	**0.39**
MNLI-SwDA-IA_QA	0.44	0.42	0.42∗†	0.54	**0.38**	0.16	0.14	0.38
BoolQ+Circa+SwDA-IA_QA	0.39	0.42	0.40∗†	0.58	0.24	0.26	0.14	0.28
MNLI+BoolQ+Circa+SwDA-IA_QA	0.34	0.40	0.36∗†	0.56	0.19	0.20	**0.16**	0.18
RoBERTa with annotator-selected context and tuned with …								
SwDA-IA_QA	0.43	0.47	0.45∗	0.62	0.24	0.40	0.12	0.42
BoolQ+SwDA-IA_QA	0.42	0.44	0.43∗†‡	0.58	0.30	0.35	0.14	0.41
MNLI+Circa+SwDA-IA_QA	0.52	0.48	**0.49∗†‡**	**0.64**	**0.40**	**0.42**	0.16	0.42
MNLI+BoolQ+Circa+SwDA-IA_QA	0.44	0.46	0.44∗†‡	0.64	0.30	0.37	0.15	0.41

Table 3: Results obtained with the test set. We present results with a RoBERTa-based classifier fine-tuned with two related corpora (MNLI, BoolQ), synthetic data for the same problem (Circa), and our corpus (SwDA-IA; only *Q*uestions, only *A*nswers, or both). We only show the best performing combinations; the supplementary materials provide all of them. We indicate statistical significance (McNemar's test (McNemar, 1947) with $p < 0.05$) as follows: ∗ with respect to the baseline, † with respect to the same fine-tuning except SwDA-IA, and ‡ with respect to the model with full context. Training with real conversation (SwDA-IA) is crucial.

conversation requires fine-tuning with real conversations, as F1 scores jump over 40% when doing so (without context: 0.32 vs. 0.46, with context: 0.28 vs. 0.43). Further, the task barely benefits from the synthetic examples in Circa, MNLI, or BoolQ (SwDA-IA_QA F1: 0.44 and 0.45, best model fine-tuning with additional corpora: 0.46 and 0.49).

Full vs. Annotator-Selected Context Since the ground truth depends on context (Section 2) and our model obtains worse results with full context, it is reasonable to conclude that our RoBERTa-based classifier is unable to extract the required information from context. The bottom block of Table 3, however, shows that context is not useless. Indeed, feeding only the conversation turns handpicked as useful by annotators (out of all the context we consider) brings statistically significant better results than feeding the full context. These results are unrealistic as they require manual annotations, but lead to the conclusion that models that better understand context are worth exploring.

5.1.1 Out-of-Domain Evaluation

In order to investigate whether models trained with SwDA-IA can determine underlying direct answers to yes-no questions from not only SwDA-IA (in-domain evaluation) but also other real conversations, we conducted an out-of-domain evaluation. Specifically, we annotated 200 yes-no questions from MRDA (Shriberg et al., 2004), a corpus of meeting transcripts, with the same procedure from Section 3. The results (Table 11 in the supple-

Long sentences
A: He is still living by himself in a little farmhouse. My grandmother died a couple of years ago. But he doesn't want to move away.
B: Uh huh.
A: And he recently had to have an operation but he just really doesn't want to go to a nursing home.
B: *Is he able to, uh, still do everything himself pretty well?*
A: *Well, he was until this operation. He has arthritis.*
B: Oh, yeah.
A: And now I don't really think he's doing that well. I have one aunt that really looks after him a lot.
Gold: ProbNo; *Predicted*: Yes

Explicit *yes* in context
A: Oh how neat.
B: So, she'd always dreamed of doing that too.
A: Yeah, that's great.
B: *So. Yeah. So is there any place you would talk me into?*
A: *Uh.*
B: It sounds like we've been to some of the same places.
Gold: Middle; *Predicted*: Yes

Negation in question or indirect Answer
A: They're not quite elderly, huh.
B: No. So I haven't, uh, really been in that situation although they are thinking about my grandmother but, uh
A: Uh huh.
B: *But that's really about it. How about you? Have you been in that situation yet?*
A: *Uh, no not for my parents.* I was around, uh, two sets of grandparents, uh, quite a bit.
B: Uh huh.
A: we, we put one, I put one grandfather in a rest home.
Gold: Yes; *Predicted*: No

Negation in context
A: Uh, I work out with free weights.
B: No, uh, I mean the running.
A: Oh, uh, yeah. It is really the, uh, aerobic work out part.
B: *You do it, you do a mile in about eight minutes or less?*
A: *Uh, about seven minutes.*
B: Uh huh. Yeah. Then you don't get, uh, out of breath.
A: Uh,no I do.
Gold: Yes; *Predicted*: No

Table 4: Examples of the most frequent error types by the best model. The yes-no question and indirect answer are in italics. We show context as shown to annotators during the corpus creation process.

mentary materials) show (unsurprisingly) worse results across all models. Interestingly, the results also show a similar trend than the in-domain evaluation: (a) SwDA-IA_QA is crucial and transfer learning with other corpora yields small benefits, and (b) annotator-selected corpus is beneficial.

6 Error Analysis

We conduct a manual qualitative analysis of 200 errors made by our best model (second row from the bottom in Table 3). This analysis sheds lights into what kind of yes-no questions, indirect answers, and context are the hardest. Table 12 in the supplementary materials presents the most common error types, and Table 4 presents examples.

We observe **long sentences** (over 20 tokens) in most errors (62.18%). It is worth noting that long sentences do not lead to many catastrophic mistakes: only 3.80% are between *Probably No* and *Yes*, 3.80% between *Yes* and *No* and 2.37% between *Yes* and *Probably No*. If **context contains a *yes* token**, the model almost always predicts *Yes*. This accounts for 19.46% of errors. As exemplified in the bottom left example in Table 4, this error occurs even if the *yes* token is before the yes-no question. We also observed **negation** in many errors. 11.08% of errors occur when there is a negation in either the question or the turn after (i.e., the turn after the question), and 7.28% when the negation is elsewhere in the context. In this case, the model predicts either *No* or *Probably No* although the gold label is *Yes*—likely because it (wrongly) learned that negation always indicates a *No* direct answer. The top and bottom right examples in Table 4 show errors that contain a negation in context.

7 Conclusions

Yes-no questions with indirect answers are common in real conversations—77.4% of 5-minute phone conversations include at least one. In this paper, we investigate the underlying direct answers to such questions in real conversations.

We have presented SwDA-IA, the first corpus with annotations for this task on top of real conversations. We show that determining the underlying direct answer requires taking into account context, as the ground truth changes depending on whether we show annotators context around the question. Our analysis also shows that context after the question is more useful, including turns by both the author of the question and the other speaker. Experimental results demonstrate that solving the task with real conversations is challenging (F1: 0.46). More importantly, (a) doing so barely benefits from fine-tuning with related tasks (MNLI and BoolQ) and (b) Circa, a corpus of synthetic questions and synthetic indirect answers, barely outperforms the majority baseline with real conversations.

Our future plans include designing models that better encode conversational context in order to obtain better results. We are also interested in exploring applications in dialogue systems to avoid inconsistencies due to misunderstandings of indirect answers to yes-no questions.

References

Penelope Brown and Stephen C Levinson. 1978. Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction*, pages 56–311. Cambridge University Press.

Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Z Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. *arXiv preprint arXiv:2005.04246*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Scott Grimm, and Christopher Potts. 2009. Not a simple yes or no: Uncertainty in indirect answers. In *Proceedings of the SIGDIAL 2009 Conference*, pages 136–143, London, UK. Association for Computational Linguistics.

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 48th Annual Meeting of the Association for Computational Linguistics*, pages 167–176, Uppsala, Sweden. Association for Computational Linguistics.

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*.

Sarah E. Finch, James D. Finch, Daniil Huryn, William M. Hutsell, Xiaoyuan S. Huang, Han He, and Jinho D. Choi. 2021. An Approach to Inference-Driven Dialogue Management within a Social Chatbot. In *Proceedings of the 4th Alexa Prize Socialbot Grand Challenge*, AlexaPrize'21.

Nancy Green and Sandra Carberry. 1999. Interpreting and generating indirect answers. *Computational Linguistics*, 25(3):389–435.

Julia Bell Hirschberg. 1985. *A theory of scalar implicature (natural languages, pragmatics, inference)*. Ph.D. thesis, University of Pennsylvania.

Beth Ann Hockey, Deborah Rossen-Knill, Beverly Spejewski, Matthew Stone, and Stephen Isard. 1997. Can you predict responses to yes/no questions? yes, no, and stuff. In *Fifth european conference on speech communication and technology*. Citeseer.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2020. Efficient dialogue state tracking by selectively overwriting memory. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 567–582, Online. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. “I'd rather just go to bed”: Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425, Online. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Ian Plewis and Thomas Coram Research Unit. 1982. Statistical methods for rates and proportions.

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.

Deborah Rossen-Knill, Beverly Spejewski, Beth Ann Hockey, Stephen Isard, and Matthew Stone. 1997. Yes/no questions and answers in the map task corpus.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Anna-Brita Stenström. 1984. *Questions and responses in English conversation*. Krieger Pub Co.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634, Florence, Italy. Association for Computational Linguistics.

Marta Tatu. 2005. Automatic discovery of intentions in text and its application to question answering. In *Proceedings of the ACL Student Research Workshop*, pages 31–36, Ann Arbor, Michigan. Association for Computational Linguistics.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

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 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Supplementary Materials

A.1 Collecting Yes-No Questions with Indirect Answers

Dialogue acts to select yes-no questions The process to collect yes-no questions from SwDA uses dialogue act annotations (Section 3.1). Here is the full list of dialogue acts we consider and their descriptions:

- `qh`: Rhetorical Questions
- `qy`: Yes-No question
- `qy^d`: Declarative Yes-No question
- `^g`: Tag-Question
- `qy^t`: Yes-No question about task
- `qy^r`: Yes-No question repeat self
- `qy^m`: Yes-No question mimic other
- `qy^h`: Question in response to a question
- `qy^c`: Yes-No question about communication
- `qy^2`: Yes-No question collaborative completion
- `qy(^q)`: Yes-No question quoted material
- `qy^g`: Yes-No question tag-question

- qy∧g∧t: Yes-No question tag-question about task
- qy∧g∧r: Yes-No question tag-question repeat self
- qy∧g∧c: Yes-No question tag-question about communication
- qy∧d∧t: Declarative Yes-No question about task
- qy∧d∧r: Declarative Yes-No question repeat self
- qy∧d∧m: Declarative Yes-No question mimic other
- qy∧d∧h: Declarative Yes-No question in response to a question
- qy∧d∧c: Declarative Yes-No question about communication
- qy∧d(∧q): Declarative Yes-No question quoted material
- qy∧c∧r: Yes-No question about-communication repeat self

Dialogue acts to discard yes-no questions with indirect answers We also use dialogue acts to discard yes-no questions with indirect answers (Section 3.1). Here is the full list of dialogue acts we consider and their descriptions:

- ny: Yes answers
- nn: No answers
- ny∧r: Yes answers repeat self
- nn∧r: No answers repeat self
- b: Acknowledge (Backchannel)
- br: Signal-non-understanding
- x: Non-verbal
- aa: Agree/Accept

Analyzing Yes-No Questions with Indirect Answers The yes-no questions with indirect answers in SwDA-IA have an average length of 39 tokens. 41.4% of the questions are indirect, i.e., do not contain a question mark despite they are yes-no questions. For example, *You must be the supervisor in this office.* is an indirect yes-no question. Figure 5 displays the most salient tokens in the yes-no questions and indirect answers. Finally, the most common first tokens in the questions are as follows:

- *Do*: 8%
- *Well*: 5%
- *Is*: 4%
- *You*: 3.9%
- *So*: 3.8%
- *And*: 3%

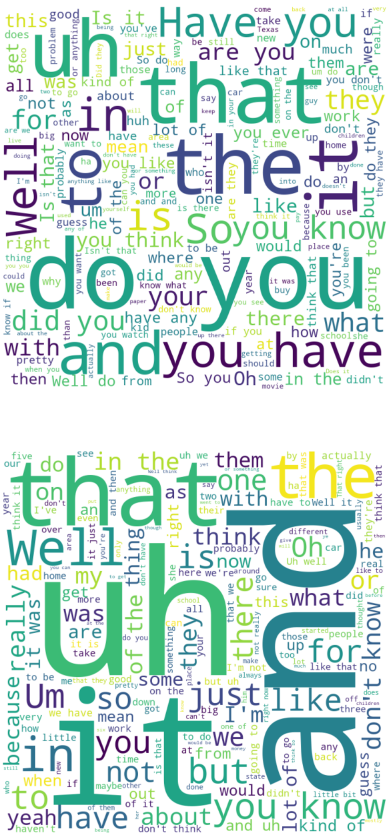

Figure 5: Word cloud displaying the most salient tokens of the yes-no questions (top) and indirect answers (bottom) in SwDA-IA.

- *Did*: 2.7%
- *Are*: 2.3%
- *Do*: 2.1%
- *Have*: 1.4%

A.2 Corpus Analysis: Context and Underlying Direct Answers

A.3 Experiments and Discussion

Tables 7–10 presents additional results complementing the results in Table 3 of the main paper. Table 11 presents the detailed out-of-domain evaluation (Section 5.1.1 in the main paper).

For all the models, we tuned the learning rate (values 5e-5, 3e-5, 2e-5), number of epochs (2, 3, 4), and batch size (16, 32) in an exhaustive combination of these parameters. Table 6 presents the hyperparameter settings with best performance on the development set fine-tuned with RoBERTa and no context.

A.4 Error Analysis

Table 12 presents the most common error types by our best model. This table complements Section 6 in the main paper.

t_{-3}, A: really
t_{-2}, B: But she started it
t_{-1}, A: That's good
t_0, B: *Well, she was depressed to begin with right?*
t_1, A: *That's one way to see it*
t_2, B: Yeah.
t_3, A: But, she also might be really calm.

Underlying direct answer ...
without context: Yes; *with context*: Middle

t_{-3}, A: I like to see you drive through Burger King now and the bags are recycled paper
t_{-2}, B: Uh huh.
t_{-1}, A: I feel like people are more aware of it or becoming more aware of it
t_0, B: *do you like to buy more recycled item*
t_1, A: *I'm not as good about searching something out like that*
t_2, B: oh really.
t_3, A: but I have been reading a lot about it

Underlying direct answer ...
without context: No; *with context*: Middle

Table 5: Examples of yes-no questions (t_0) with indirect answers (t_1). Determining the underlying direct answers requires taking into account more conversation context than the question and indirect answer: the ground truth annotated by humans changes depending on whether they also have access to context (t_{-3}, t_{-2}, t_{-1}, t_2, and t_3). This examples complement Table 2 in the main paper with additional examples of less frequent label changes.

Model	learning rate	epochs	batch size
BoolQ+Circa+SwDA-IA_QA	2e-5	3	16
SwDA-IA_QA	2e-5	4	16
MNLI+SwDA-IA_QA	2e-5	4	16
BoolQ+SwDA-IA_QA	3e-5	3	16
Circa+SwDA-IA_QA	2e-5	4	16
MNLI+BoolQ+SwDA-IA_QA	3e-5	3	16
MNLI+Circa+SwDA-IA_QA	2e-5	3	16
BoolQ+Circa+SwDA-IA_QA	2e-5	4	16
MNLI+BoolQ+Circa+SwDA-IA_QA	3e-5	3	16

Table 6: Models with hyperparameter settings of best performance on the development set fine-tuned with RoBERTa and no context.

	All labels			Yes	ProbYes	Middle	ProbNo	No
	P	R	F1	F1	F1	F1	F1	F1
Majority Baseline	0.18	0.42	0.25	0.59	0.00	0.00	0.00	0.00
RoBERTa without context and tuned with . . .								
MNLI	0.24	0.20	0.12∗	0.32	0.00	0.14	0.00	0.00
BoolQ	0.20	0.35	0.25	0.50	0.00	0.00	0.00	0.00
Circa	0.31	0.34	0.32	0.58	0.04	0.05	0.03	0.35
MNLI+BoolQ	0.08	0.20	0.10∗	0.32	0.00	0.00	0.00	0.00
MNLI+Circa	0.20	0.23	0.21	0.42	0.09	0.00	0.04	0.14
BoolQ+Circa	0.19	0.18	0.11∗	0.22	0.00	0.00	0.00	0.24
MNLI+BoolQ+Circa	0.26	0.29	0.24	0.50	0.00	0.20	0.04	0.03
SwDA-IA_Q	0.18	0.43	0.26	0.45	0.18	0.22	0.22	0.18
SwDA-IA_A	0.42	0.45	0.43∗	0.60	0.29	0.40	0.09	0.31
SwDA-IA_QA	0.44	0.45	0.44∗	0.62	0.31	0.33	0.08	**0.44**
MNLI+SwDA-IA_QA	0.46	0.43	0.45∗†	0.58	0.25	0.44	0.16	0.36
BoolQ+SwDA-IA_QA	0.39	0.42	0.39∗†	0.58	0.22	0.32	0.06	0.36
Circa+SwDA-IA_QA	0.42	0.44	0.42∗†	0.60	0.27	0.24	0.21	0.36
MNLI+BoolQ+SwDA-IA_QA	0.47	0.44	0.45∗†	0.62	**0.32**	0.30	0.18	0.32
MNLI+Circa+SwDA-IA_QA	0.41	0.43	0.42∗†	0.62	0.29	0.28	0.22	0.31
BoolQ+Circa+SwDA-IA_QA	0.46	0.48	**0.46∗†**	**0.64**	0.22	**0.45**	**0.24**	0.42
MNLI+BoolQ+Circa+SwDA-IA_QA	0.41	0.40	0.40∗†	0.54	0.30	0.34	0.20	0.34
RoBERTa with full context and tuned with . . .								
MNLI	0.19	0.16	0.13∗	0.12	0.00	0.20	0.00	0.00
BoolQ	0.20	0.18	0.10∗	0.32	0.00	0.00	0.00	0.00
Circa	0.24	0.28	0.27	0.55	0.03	0.00	0.00	0.20
MNLI+BoolQ	0.23	0.16	0.14∗	0.30	0.00	0.00	0.00	0.14
MNLI+Circa	0.30	0.15	0.13∗	0.20	0.02	0.24	0.05	0.16
BoolQ+Circa	0.24	0.33	0.28	0.50	0.24	0.00	0.16	0.07
MNLI+BoolQ+Circa	0.21	0.26	0.23	0.34	0.28	0.08	0.00	0.10
SwDA-IA_Q	0.32	0.39	0.33	0.57	0.08	0.26	0.02	0.24
SwDA-IA_A	0.38	0.40	0.38	0.53	0.22	0.26	0.12	0.38
SwDA-IA_QA	0.41	0.43	**0.43∗**	**0.59**	0.24	**0.32**	0.14	**0.39**
MNLI+SwDA-IA_QA	0.44	0.42	0.42∗†	0.54	**0.38**	0.16	0.14	0.38
BoolQ+SwDA-IA_QA	0.38	0.44	0.39∗†	0.59	0.20	0.23	0.00	0.36
Circa+SwDA-IA_QA	0.36	0.40	0.39∗†	0.58	0.11	0.30	0.00	0.38
MNLI+BoolQ+SwDA-IA_QA	0.38	0.39	0.37∗†	0.56	0.18	0.16	0.10	0.32
MNLI+Circa+SwDA-IA_QA	0.36	0.38	0.37∗†	0.58	0.20	0.18	0.11	0.31
BoolQ+Circa+SwDA-IA_QA	0.39	0.42	0.40∗†	0.58	0.24	0.26	0.14	0.28
MNLI+BoolQ+Circa+SwDA-IA_QA	0.34	0.40	0.36∗†	0.56	0.19	0.20	**0.16**	0.18
RoBERTa with annotator-selected context and tuned with . . .								
MNLI	0.18	0.15	0.08∗	0.30	0.00	0.02	0.00	0.14
BoolQ	0.02	0.12	0.04∗	0.20	0.00	0.00	0.00	0.00
Circa	0.30	0.34	0.32	0.58	0.04	0.04	0.00	0.38
MNLI+BoolQ	0.22	0.18	0.18∗	0.30	0.00	0.00	0.00	0.21
MNLI+Circa	0.30	0.41	0.34∗	0.58	0.14	0.04	0.00	0.33
BoolQ+Circa	0.33	0.41	0.28	0.58	0.06	0.06	0.04	0.04
MNLI+BoolQ+Circa	0.24	0.34	0.30	0.56	0.24	0.00	0.00	0.00
SwDA-IA_Q	0.34	0.37	0.35∗‡	0.54	0.22	0.12	0.11	0.23
SwDA-IA_A	0.46	0.42	0.44∗‡	0.58	0.24	0.40	0.17	0.36
SwDA-IA_QA	0.43	0.47	0.45∗	0.62	0.24	0.40	0.12	0.42
MNLI+SwDA-IA_QA	0.45	0.43	0.44∗†	0.62	0.31	0.30	0.12	0.38
BoolQ+SwDA-IA_QA	0.42	0.44	0.43∗†‡	0.58	0.30	0.35	0.14	0.41
Circa+SwDA-IA_QA	0.44	0.40	0.42∗†	0.58	0.16	0.50	0.12	**0.51**
MNLI+BoolQ+SwDA-IA_QA	0.44	0.42	0.42∗†	0.59	0.24	0.30	**0.19**	0.37
MNLI+Circa+SwDA-IA_QA	0.52	0.48	**0.49∗†‡**	**0.64**	**0.40**	**0.42**	0.16	0.42
BoolQ+Circa+SwDA-IA_QA	0.43	0.46	0.45∗†	0.62	0.27	0.40	0.17	0.40
MNLI+BoolQ+Circa+SwDA-IA_QA	0.44	0.46	0.44∗†‡	0.64	0.30	0.37	0.15	0.41

Table 7: Results obtained with the test set. These results complement Table 3 in the main paper (Table 3 is subsumed by this one). We indicate statistical significance (McNemar's test (McNemar, 1947) with $p < 0.05$) as follows: $*$ indicates better results with respect to the baseline, † with respect to the same fine-tuning except SwDA-IA, and ‡ with respect to the model with full context. Training with real conversation (SwDA-IA) is crucial, and annotator-selected context yields the best results.

	Yes			ProbYes		
	P	R	F1	P	R	F1
Majority Baseline	0.42	1.00	0.59	0.00	0.00	0.00
RoBERTa without context and tuned with …						
MNLI	0.34	0.38	0.32	0.00	0.00	0.00
BoolQ	0.49	0.52	0.50	0.00	0.00	0.00
Circa	0.51	0.69	0.58	0.33	0.02	0.04
MNLI+BoolQ	0.28	0.34	0.32	0.00	0.00	0.00
MNLI+Circa	0.40	0.43	0.42	0.17	0.06	0.09
BoolQ+Circa	0.15	0.49	0.22	0.00	0.00	0.00
MNLI+BoolQ+Circa	0.48	0.52	0.50	0.00	0.00	0.00
SwDA-IA_Q	0.38	0.48	0.45	0.17	0.19	0.18
SwDA-IA_A	0.58	0.62	0.60	0.29	0.30	0.29
SwDA-IA_QA	0.60	0.64	0.62	0.30	0.34	0.31
MNLI+SwDA-IA_QA	0.56	0.60	0.58	0.19	0.3	0.25
BoolQ+SwDA-IA_QA	0.55	0.59	0.58	0.15	0.49	0.22
Circa+SwDA-IA_QA	0.59	0.63	0.60	0.22	0.30	0.27
MNLI+BoolQ+SwDA-IA_QA	0.60	0.64	0.62	0.30	0.35	**0.32**
MNLI+Circa+SwDA-IA_QA	0.60	0.64	0.62	0.24	0.30	0.29
BoolQ+CircaQ+SwDA-IA_QA	0.61	0.65	**0.64**	0.15	0.49	0.22
MNLI+BoolQ+Circa+SwDA-IA_QA	0.52	0.56	0.54	0.28	0.32	0.30
RoBERTa with full context and tuned with …						
MNLI	0.22	0.11	0.12	0.00	0.00	0.00
BoolQ	0.28	0.35	0.32	0.00	0.00	0.00
Circa	0.54	0.56	0.55	0.28	0.02	0.03
MNLI+BoolQ	0.28	0.32	0.30	0.00	0.00	0.00
MNLI+Circa	0.18	0.22	0.20	0.29	0.02	0.02
BoolQ+Circa	0.48	0.53	0.50	0.22	0.28	0.24
MNLI+BoolQ+Circa	0.38	0.29	0.34	0.25	0.30	0.28
SwDA-IA_Q	0.59	0.54	0.57	0.30	0.05	0.08
SwDA-IA_A	0.50	0.56	0.53	0.21	0.23	0.22
SwDA-IA_QA	0.58	0.59	**0.59**	0.22	0.26	0.24
MNLI+SwDA-IA_QA	0.50	0.58	0.54	0.33	0.41	**0.38**
BoolQ+SwDA-IA_QA	0.56	0.62	0.59	0.18	0.22	0.20
Circa+SwDA-IA_QA	0.62	0.56	0.58	0.29	0.07	0.11
MNLI+BoolQ+SwDA-IA_QA	0.57	0.54	0.56	0.12	0.23	0.18
MNLI+Circa+SwDA-IA_QA	0.54	0.60	0.58	0.24	0.18	0.20
BoolQ+Circa+SwDA-IA_QA	0.55	0.61	0.58	0.22	0.26	0.24
MNLI+BoolQ+Circa+SwDA-IA_QA	0.59	0.54	0.56	0.24	0.16	0.19
RoBERTa with annotator-selected context and tuned with …						
MNLI	0.28	0.33	0.30	0.00	0.00	0.00
BoolQ	0.18	0.23	0.20	0.00	0.00	0.00
Circa	0.59	0.58	0.58	0.30	0.02	0.04
MNLI+BoolQ	0.28	0.32	0.30	0.00	0.00	0.00
MNLI+Circa	0.57	0.59	0.58	0.10	0.18	0.14
BoolQ+Circa	0.59	0.58	0.58	0.28	0.03	0.06
MNLI+BoolQ+Circa	0.54	0.60	0.56	0.22	0.26	0.24
SwDA-IA_Q	0.52	0.56	0.54	0.20	0.24	0.22
SwDA-IA_A	0.56	0.60	0.58	0.26	0.22	0.24
SwDA-IA_QA	0.58	0.64	0.62	0.20	0.28	0.24
MNLI+SwDA-IA_QA	0.59	0.65	0.62	0.30	0.32	0.31
BoolQ+SwDA-IA_QA	0.54	0.62	0.58	0.26	0.34	0.30
Circa+SwDA-IA_QA	0.56	0.60	0.58	0.18	0.14	0.16
MNLI+BoolQ+SwDA-IA_QA	0.58	0.62	0.59	0.20	0.28	0.24
MNLI+Circa+SwDA-IA_QA	0.62	0.66	**0.64**	0.38	0.43	**0.40**
BoolQ+Circa+SwDA-IA_QA	0.60	0.65	0.62	0.26	0.29	0.27
MNLI+BoolQ+Circa+SwDA-IA_QA	0.63	0.66	0.64	0.38	0.32	0.30

Table 8: Results obtained with the test set. We present Precision (P), Recall (R) and F1 scores for the *Yes* and *Probably Yes* labels. These results complement Table 3 in the main paper.

	Middle		
	P	R	F1
Majority Baseline	0.00	0.00	0.00
RoBERTa without context and tuned with . . .			
MNLI	0.22	0.18	0.14
BoolQ	0.00	0.00	0.00
Circa	0.35	0.02	0.05
MNLI+BoolQ	0.00	0.00	0.00
MNLI+Circa	0.00	0.00	0.00
BoolQ+Circa	0.00	0.00	0.00
MNLI+BoolQ+Circa	0.13	0.48	0.20
SwDA-IA_Q	0.14	0.50	0.22
SwDA-IA_A	0.50	0.29	0.40
SwDA-IA_QA	0.29	0.35	0.33
MNLI+SwDA-IA_QA	0.39	0.46	0.44
BoolQ+SwDA-IA_QA	0.30	0.35	0.32
Circa+SwDA-IA_QA	0.19	0.31	0.24
MNLI+BoolQ+SwDA-IA_QA	0.24	0.35	0.30
MNLI+Circa+SwDA-IA_QA	0.24	0.29	0.28
BoolQ+Circa+SwDA-IA_QA	0.38	0.50	**0.45**
MNLI+BoolQ+Circa+SwDA-IA_QA	0.28	0.38	0.34
RoBERTa with full context and tuned with . . .			
MNLI	0.23	0.18	0.20
BoolQ	0.00	0.00	0.00
Circa	0.00	0.00	0.00
MNLI+BoolQ	0.00	0.00	0.00
MNLI+Circa	0.22	0.26	0.24
BoolQ+Circa	0.00	0.00	0.00
MNLI+BoolQ+Circa	0.20	0.04	0.08
SwDA-IA_Q	0.22	0.28	0.26
SwDA-IA_A	0.25	0.27	0.26
SwDA-IA_QA	0.31	0.32	**0.32**
MNLI+SwDA-IA_QA	0.14	0.18	0.16
BoolQ+SwDA-IA_QA	0.27	0.21	0.23
Circa+SwDA-IA_QA	0.33	0.27	0.3
MNLI+BoolQ+SwDA-IA_QA	0.14	0.18	0.16
MNLI+Circa+SwDA-IA_QA	0.19	0.15	0.18
BoolQ+Circa+SwDA-IA_QA	0.24	0.28	0.26
MNLI+BoolQ+Circa+SwDA-IA_QA	0.18	0.22	0.20
RoBERTa with annotator-selected context and tuned with . . .			
MNLI	0.19	0.01	0.02
BoolQ	0.00	0.00	0.00
Circa	0.30	0.02	0.04
MNLI+BoolQ	0.00	0.00	0.00
MNLI+Circa	0.29	0.02	0.04
BoolQ+Circa	0.32	0.03	0.06
MNLI+BoolQ+Circa	0.00	0.00	0.00
SwDA-IA_Q	0.10	0.16	0.12
SwDA-IA_A	0.38	0.44	0.40
SwDA-IA_QA	0.39	0.41	0.40
MNLI+SwDA-IA_QA	0.35	0.28	0.30
BoolQ+SwDA-IA_QA	0.36	0.33	0.35
Circa+SwDA-IA_QA	0.48	0.52	0.50
MNLI+BoolQ+SwDA-IA_QA	0.26	0.34	0.30
MNLI+Circa+SwDA-IA_QA	0.40	0.44	**0.42**
BoolQ+Circa+SwDA-IA_QA	0.38	0.44	0.40
MNLI+BoolQ+Circa+SwDA-IA_QA	0.35	0.40	0.37

Table 9: Results obtained with the test set. We present Precision (P), Recall (R) and F1 scores for the *Middle* label. These results complement Table 3 in the main paper .

	No			ProbNo		
	P	R	F1	P	R	F1
Majority Baseline	0.00	0.00	0.00	0.00	0.00	0.00
RoBERTa without context and tuned with …						
MNLI	0.00	0.00	0.00	0.00	0.00	0.00
BoolQ	0.00	0.00	0.00	0.00	0.00	0.00
Circa	0.37	0.34	0.35	0.25	0.01	0.03
MNLI+BoolQ	0.00	0.00	0.00	0.00	0.00	0.00
MNLI+Circa	0.23	0.16	0.14	0.25	0.02	0.04
BoolQ+Circa	0.18	0.50	0.24	0.00	0.00	0.00
MNLI+BoolQ+Circa	0.30	0.02	0.03	0.21	0.04	0.04
SwDA-IA_Q	0.20	0.16	0.18	0.15	0.48	0.22
SwDA-IA_A	0.50	0.38	0.31	0.18	0.05	0.09
SwDA-IA_QA	0.43	0.44	**0.44**	0.15	0.06	0.08
MNLI+SwDA-IA_QA	0.56	0.30	0.36	0.14	0.21	0.16
BoolQ+SwDA-IA_QA	0.34	0.38	0.36	0.28	0.04	0.06
Circa+SwDa-IA_QA	0.35	0.37	0.36	0.15	0.48	0.21
MNLI+BoolQ+SwDA-IA_QA	0.28	0.34	0.32	0.12	0.30	0.18
MNLI+Circa+SwDA-IA_QA	0.28	0.33	0.31	0.21	0.23	0.22
BoolQ+Circa+SwDA-IA_QA	0.38	0.44	0.42	0.22	0.26	**0.24**
MNLI+BoolQ+Circa+SwDA-IA_QA	0.33	0.34	0.34	0.19	0.21	0.20
RoBERTa with full context and tuned with …						
MNLI	0.00	0.00	0.00	0.00	0.00	0.00
BoolQ	0.00	0.00	0.00	0.00	0.00	0.00
Circa	0.15	0.25	0.20	0.00	0.00	0.00
MNLI+BoolQ	0.28	0.10	0.14	0.00	0.00	0.00
MNLI+Circa	0.24	0.12	0.16	0.28	0.03	0.05
BoolQ+Circa	0.28	0.05	0.07	0.28	0.13	0.16
MNLI+BoolQ+Circa	0.30	0.06	0.10	0.00	0.00	0.00
SwDA-IA_Q	0.22	0.26	0.24	0.20	0.01	0.02
SwDA-IA_A	0.34	0.40	0.38	0.20	0.10	0.12
SwDA-IA_QA	0.38	0.42	**0.39**	0.10	0.18	0.14
MNLI+SwDA-IA_QA	0.40	0.36	0.38	0.19	0.10	0.14
BoolQ+SwDA-IA_QA	0.40	0.34	0.36	0.00	0.00	0.00
Circa+SwDa-IA_QA	0.34	0.40	0.38	0.00	0.00	0.00
MNLI+BoolQ+SwDA-IA_QA	0.30	0.34	0.32	0.11	0.10	0.10
MNLI+Circa+SwDA-IA_QA	0.29	0.32	0.31	0.20	0.07	0.11
BoolQ+Circa+SwDA-IA_QA	0.24	0.32	0.28	0.28	0.10	0.14
MNLI+BoolQ+Circa+SwDA-IA_QA	0.16	0.20	0.18	0.20	0.14	**0.16**
RoBERTa with annotator-selected context and tuned with …						
MNLI	0.10	0.18	0.14	0.00	0.00	0.00
BoolQ	0.00	0.00	0.00	0.00	0.00	0.00
Circa	0.32	0.40	0.38	0.00	0.00	0.00
MNLI+BoolQ	0.20	0.25	0.21	0.00	0.00	0.00
MNLI+Circa	0.36	0.29	0.33	0.00	0.00	0.00
BoolQ+Circa	0.29	0.02	0.04	0.28	0.02	0.04
MNLI+BoolQ+Circa	0.00	0.00	0.00	0.00	0.00	0.00
SwDA-IA_Q	0.20	0.26	0.23	0.08	0.16	0.11
SwDA-IA_A	0.33	0.39	0.36	0.12	0.20	0.17
SwDA-IA_QA	0.40	0.44	0.42	0.11	0.14	0.12
MNLI+SwDA-IA_QA	0.36	0.40	0.38	0.10	0.16	0.12
BoolQ+SwDA-IA_QA	0.44	0.39	0.41	0.16	0.13	0.14
Circa+SwDa-IA_QA	0.45	0.55	**0.51**	0.14	0.11	0.12
MNLI+BoolQ+SwDA-IA_QA	0.34	0.40	0.37	0.18	0.20	**0.19**
MNLI+Circa+SwDA-IA_QA	0.40	0.43	0.42	0.16	0.17	0.16
BoolQ+Circa+SwDA-IA_QA	0.41	0.40	0.40	0.13	0.22	0.17
MNLI+BoolQ+Circa+SwDA-IA_QA	0.36	0.44	0.41	0.13	0.17	0.15

Table 10: Results obtained with the test set. We present Precision (P), Recall (R) and F1 scores for the *No* and *Probably No* labels.These results complement Table 3 in the main paper.

	All labels			Yes	ProbYes	Middle	ProbNo	No
	P	R	F1	F1	F1	F1	F1	F1
Majority Baseline	0.22	0.40	0.28	0.61	0.00	0.00	0.00	0.00
RoBERTa without context and tuned with …								
MNLI	0.14	0.06	0.08	0.30	0.00	0.02	0.00	0.14
BoolQ	0.22	0.16	0.20	0.36	0.00	0.00	0.00	0.12
Circa	0.23	0.32	0.27	0.55	0.03	0.00	0.04	0.20
MNLI+Circa	0.24	0.14	0.18	0.30	0.05	0.08	0.00	0.14
MNLI+BoolQ+Circa	0.20	0.18	0.20	0.43	0.00	0.00	0.00	0.18
SwDA-IA_Q	0.20	0.25	0.18	0.30	0.24	0.06	0.00	0.10
SwDA-IA_A	0.31	0.37	0.32	0.56	0.13	0.04	0.00	**0.34**
SwDA-IA_QA	0.28	0.37	0.34	0.58	**0.19**	0.18	0.11	0.30
MNLI+SwDA-IA_QA	0.29	0.36	0.34	0.56	0.12	0.29	0.00	0.40
BoolQ+Circa+SwDA-IA_QA	0.26	0.39	**0.35**	**0.58**	0.16	**0.22**	**0.12**	0.32
MNLI+BoolQ+Circa+SwDA-IA_QA	0.31	0.32	0.30	0.50	0.19	0.21	0.10	0.20
RoBERTa with full context and tuned with …								
MNLI	0.17	0.13	0.10	0.12	0.00	0.15	0.00	0.00
BoolQ	0.18	0.16	0.09	0.25	0.00	0.00	0.00	0.00
Circa	0.24	0.18	0.20	0.42	0.09	0.14	0.00	0.04
BoolQ+Circa	0.20	0.30	0.22	0.44	0.22	0.00	0.10	0.07
MNLI+BoolQ+Circa	0.18	0.22	0.20	0.30	0.24	0.05	0.00	0.10
SwDA-IA_Q	0.25	0.15	0.16	0.36	0.09	0.00	0.03	0.12
SwDA-IA_A	0.28	0.32	0.30	0.43	0.17	0.23	0.21	0.18
SwDA-IA_QA	0.34	0.38	**0.32**	**0.54**	0.18	**0.24**	0.10	0.21
MNLI-SwDA-IA_QA	0.32	0.37	0.32	0.54	**0.22**	0.12	0.11	**0.23**
BoolQ+Circa+SwDA-IA_QA	0.29	0.33	0.30	0.44	0.18	0.23	**0.22**	0.18
MNLI+BoolQ+Circa+SwDA-IA_QA	0.20	0.27	0.25	0.40	0.10	0.18	0.10	0.15
RoBERTa with annotator-selected context and tuned with …								
SwDA-IA_QA	0.27	0.35	0.33	0.55	0.20	0.12	0.11	0.23
BoolQ+SwDA-IA_QA	0.36	0.34	0.34	0.52	0.24	0.18	0.08	0.30
MNLI+Circa+SwDA-IA_QA	0.39	0.37	**0.38**	**0.55**	**0.24**	**0.30**	0.08	**0.35**
MNLI+BoolQ+Circa+SwDA-IA_QA	0.30	0.35	0.32	0.48	0.17	0.23	**0.18**	0.24

Table 11: Out-of-domain evaluation, i.e., testing with 200 questions from MRDA (Section 5.1.1). These results complement Table 3 in the main paper, which present in-domain evaluation (i.e., testing with SwDA-IA).

Error Type	%	Gold	Pred.
Long sentences	21.20	ProbYes	Yes
	15.51	Yes	ProbYes
	9.49	Yes	Middle
	6.01	ProbNo	No
	3.80	ProbNo	Yes
	3.80	Yes	No
	2.37	Yes	ProbNo
Explicit *yes* in context	11.87	Middle	Yes
	7.59	ProbNo	Yes
Negation in . . .			
question or ind. answer	7.59	Yes	No
	3.48	Yes	ProbNo
in context	3.80	Yes	No
	3.80	Yes	ProbNo

Table 12: Qualitative analysis of the errors made by our best model (Table 3, second row from the bottom). We indicate the frequency of the most common combinations of gold and predicted labels for each error type. *Context* here does not include the yes-no question and indirect answer. This table complements Section 6 in the main paper.