

# “I’d rather just go to bed”: Understanding Indirect Answers

Annie Louis  
Google Research, UK  
annielouis@google.com

Dan Roth\*  
University of Pennsylvania  
danroth@seas.upenn.edu

Filip Radlinski  
Google Research, UK  
filiprad@google.com

## Abstract

We revisit a pragmatic inference problem in dialog: Understanding indirect responses to questions. Humans can interpret ‘*I’m starving.*’ in response to ‘*Hungry?*’, even without direct cue words such as ‘yes’ and ‘no’. In dialog systems, allowing natural responses rather than closed vocabularies would be similarly beneficial. However, today’s systems are only as sensitive to these pragmatic moves as their language model allows. We create and release<sup>1</sup> the first large-scale English language corpus ‘Circa’ with 34,268 (polar question, indirect answer) pairs to enable progress on this task. The data was collected via elaborate crowdsourcing, and contains utterances with yes/no meaning, as well as uncertain, middle-ground, and conditional responses. We also present BERT-based neural models to predict such categories for a question-answer pair. We find that while transfer learning from entailment works reasonably, performance is not yet sufficient for robust dialog. Our models reach 82-88% accuracy for a 4-class distinction, and 74-85% for 6 classes.

## 1 Introduction

Humans produce and interpret complex utterances even in simple scenarios. For example, for the polar (yes/no) question ‘*Want to get dinner?*’, there are many perfectly natural responses in addition to ‘yes’ and ‘no’, as in Table 1. How should a dialog system interpret these INDIRECT answers? Many can be understood based on the answer text alone, e.g. ‘*I would like that*’ or ‘*I’d rather just go to bed*’. For others, the question is crucial, e.g. ‘*Dinner would be lovely.*’ is a positive reply here, but a negative answer to ‘*Want to get lunch?*’. In this

### “Want to get some dinner together?”

“I know a restaurant we could get a reservation at.”  
“I have already eaten recently.”  
“I hope to make it home by supper but I’m not sure I can.”  
“Dinner would be lovely.”  
“I’d rather just go to bed.”  
“There’s a few new restaurants we could go to.”  
“I would like that.”  
“We could do dinner this weekend.”  
“I would like to go somewhere casual.”  
“I’d like to try the new Italian place.”

Table 1: A polar question with 10 indirect responses, taken from our corpus.

paper, we present the first large scale corpus and models for interpreting such indirect answers.

Previous attempts to interpret indirect yes/no answers have been small scale and without data-driven techniques (Green and Carberry, 1999; de Marneffe et al., 2010). However, recent success on many language understanding problems (Wang et al., 2019), the impressive generation capabilities of modern dialog systems (Zhang et al., 2019; Adiwardana et al., 2020), as well as the huge interest in yes/no question-answering (Choi et al., 2018; Clark et al., 2019) have created a conducive environment for revisiting this hard task.

We introduce *Circa*, a new dataset with 34K pairs of polar questions and indirect answers in the English language. This high quality corpus consists of natural responses collected via crowd workers, and goes beyond binary yes/no meaning to include conditionals, uncertain, and middle-ground answers. *Circa* contains many phenomena of interest, although the first step, which we address here, is how to robustly classify a question-answer pair into one of the above meaning categories. We find that BERT (Devlin et al., 2019) fine-tuned on entailment data is an effective initial approach, mirroring the success in question-answering work involving yes/no questions (Clark et al., 2019). It reaches

\* Work done at Google

<sup>1</sup>The corpus can be downloaded from <http://google/circa>

an accuracy of 85-88% for responses in the same situational context, and 6-10% lower accuracy on held-out scenarios. The answer text itself (as in ‘*I would like that*’) contains strong cues leading to 78-82% accuracy, however the best results come from jointly analyzing the question and the answer.

## 2 Related work

Indirect answers to polar questions are reasonably frequent and warrant deep study. Early work put the proportion at 13% in face-to-face and telephone conversations (Stenström, 1984), and at 27% in an instruction giving/following map task (Rossen-Knill et al., 1997; Hockey et al., 1997). For a more recent and larger analysis, consider the Cornell Movie Dialog Corpus (Danescu-Niculescu-Mizil and Lee, 2011). We heuristically mined yes/no questions and their following utterances, finding 6,327 pairs. Direct answers (i.e., answers with ‘yes’, ‘no’, ‘maybe’ and related terms such as ‘okay’, ‘yup’, etc.) only cover 53% of the pairs. This suggests that indirect responses could be even more frequent in natural open-domain dialogue.

Even when a direct answer is possible, speakers use indirect answers to be cooperative and address anticipated follow-up questions (Stenström, 1984), to provide explanations in the case of a negative answer (Stenström, 1984), to block misleading interpretations that may arise from a curt ‘yes’ or ‘no’ reply (Hirschberg, 1985), and since indirect answers may appear more polite (Brown and Levinson, 1978). But we lack a large corpus of such answers to study these multiple pragmatic functions. Our work aims to fill this gap.

On the computational side, there are impressive efforts towards planning, generation, and detection of indirect answers, albeit, on a small scale, and without data-driven approaches. Green and Carberry (1999)’s early work leverages discourse relations for generating indirect answers. For example, an ‘elaboration’ may be relevant for a ‘yes’ response, and a ‘contrast’ might help convey a ‘no’ answer. de Marneffe et al. (2009) reason about such answers using Markov Logic Networks. In subsequent work, de Marneffe et al. (2010) present one of the first data-driven studies into indirect answers containing scalar adjectives. They mine a set of 224 question-answer pairs from interview transcripts and dialog corpora. Using polarity information from review data, and manual coding of test samples, they achieve an accuracy of 71%

on three classes ‘yes’, ‘no’ and ‘uncertain’. Our work aims to collect a much larger and more diverse natural corpus, and demonstrates the first automatic approach using recent advances in natural language inference (NLI). de Marneffe et al. (2010) also demonstrated the first crowd annotation of indirect answers, and we draw on many aspects of their formulation for the creation of our corpus.

An unexpected limelight on yes/no questions has also arisen in recent question-answering (QA) work. Researchers have noticed that yes/no questions are complex and naturally arise (as high as 20%) when questions are posed one after the other in a conversation (Reddy et al., 2019; Choi et al., 2018). Their goal is to produce direct ‘yes’ or ‘no’ answers, but obtaining them requires inference against a paragraph or excerpt, an analogous task to our yes/no inference from indirect answers. Very recent work (Clark et al., 2019) has specifically sought to improve this ability in QA systems, by building a corpus of 16K yes/no factual questions paired with Wikipedia passages from which the yes/no answer can be inferred. Departing from factual texts, our focus is on single-turn *conversational* responses in everyday situations. The latter are faithful, cooperative, and grounded in world knowledge. Still, transfer learning from factual corpora could prove useful and we explore this too.

## 3 The Circa Corpus

Circa (meaning approximately) is our crowd-sourced corpus for research on indirect answers. It contains 34,268 question-answer pairs, comprising 3,431 unique questions with up to 10 answers to each. Each question-answer pair is also annotated with meaning based on a fixed set of categories.<sup>2</sup> We explain all aspects of the collection process below, but in the interest of space, further details (complete annotator instructions, interfaces, and examples) are in the Appendix.

Our data consists of very short dialogues, each containing a question and its indirect answer. The texts are varied in semantic and syntactic forms, and grounded in 10 different situational contexts. Table 2 presents some examples, showing that the answers go beyond binary yes/no distinctions. We created this data via an elaborate crowd-annotation exercise which comprised of 4 *steps* described next.

<sup>2</sup>The data can be downloaded from <http://goo.gle/circa>

<p><b>S1: Talking to a friend about food preferences.</b>  Q: "Do you like pizza?"  A: "I like it when the toppings are meat, not vegetable."</p>
<p><b>S2: Talking to a friend about music preferences.</b>  Q: "Do you like guitars?"  A: "I practice playing each weekend."</p>
<p><b>S3: Talking to a friend about weekend activities.</b>  Q: "Are you available this Sunday evening?"  A: "What did you have in mind?"</p>
<p><b>S4: Talking to a friend about book preferences.</b>  Q: "Are you a fan of comic books?"  A: "I read an Archie every time I have lunch."</p>
<p><b>S5. Your friend is visiting from out of town.</b>  Q: "Would you like to go out for dinner?"  A: "I could go for some Mexican."</p>
<p><b>S6. Two colleagues leaving work on a Friday.</b>  Q: "Long week?"  A: "I've had worse weeks."</p>
<p><b>S7. You friend is planning to buy a flat in New York.</b>  Q: "Does the flat's price fit your long-term budget?"  A: "Well, if it doesn't I will definitely refinance my mortgage."</p>
<p><b>S8. Your friend is thinking of switching jobs.</b>  Q: "Do you have to travel far?"  A: "My commute is about 10 minutes."</p>
<p><b>S9. Two childhood neighbours unexpectedly run into each other at a cafe.</b>  Q: "Are you going to the high school reunion in June?"  A: "I forgot all about that."</p>
<p><b>S10. Meeting your new neighbour for the first time.</b>  Q: "Did you move from near-by?"  A: "I am from Canada."</p>

Table 2: Examples of questions and answers in our 10 dialogue scenarios.

### 3.1 Step 1: Dialog scenarios

We designed 10 diverse prompts to serve as situational contexts. These everyday situations can lead to productive dialog, and simple yes/no questions (i.e., they do not depend on elaborate prior context for initiating a conversation). We designed them manually to intentionally cover a number of situations that encourage variety in questions and responses. As only a small number were needed, and the desiderata are not trivial, a crowd task was not suitable for prompt development. In Table 2, S1–S10 are the titles of the 10 dialog scenarios (each also consisting of a longer description).

The rest of the data were collected using crowd workers. We ran pilots for each step of data collection, and perused their results manually to ensure clarity in guidelines, and quality of the data. We recruited native English speakers, mostly from the USA, and a few from the UK and Canada. We did not collect any further information about the crowd workers.

### Annotator instructions

In this task, we will ask you to provide yes/no questions in a social dialogue situation.

**Example:** Suppose that you are trying to learn about a friend's movie preferences, but can only ask yes/no questions. Provide 5 useful questions that can be answered "Yes" or "No".

**Note:** (1) We are looking for variety in the questions. For instance:

'Have you watched Star Wars?'  
'Do you like movies with a complicated plot?'  
'Did you enjoy the last movie we saw together?'  
'Want to go watch a thriller this weekend?'  
'Are you into the Avengers series?'

Note that the questions have different forms as well as different content.

(2) Remember that the setting is a conversation with a friend (or neighbour or colleague). Please keep the questions casual, so they would be natural during an informal conversation.

Table 3: Annotator instructions for Step 2.

### 3.2 Step 2: Question collection

In this phase, we ask annotators to write yes/no questions for a given dialog scenario.

Table 3 shows the instructions displayed to annotators.<sup>3</sup> (Also see Figure 1 and Table 11.) 100 annotators each provided 5 questions per scenario, resulting in 5,000 questions.

Questions where annotators did not adhere to topic provided were removed. Of the remaining 4,710 questions, 84% were unique. Understandably, some scenarios had more repetitions than others: For food preferences, 76% of the questions were unique, as opposed to 95% when talking about a friend's job change. Below we show the most and least common questions in the food context.

#### Most common food questions

25 times	Do you like spicy food?
11 times	Are you vegetarian?
10 times	Do you eat meat?

#### Sample of least common food questions

1 time	Have you ever tried vegan cuisine?
1 time	Do you have a gluten allergy?
1 time	Are you familiar with Thai food?

The most common one '*Do you like spicy food?*' was suggested by 25 out of 100 annotators.

One important aspect of our design is that each annotator was asked to provide five questions at the same time. Obvious questions often showed up as the *first* question, with later questions becoming more complex and more interesting. Question properties such as length, inverse document frequency, and type-token ratio confirm this difference.

<sup>3</sup>In annotator instructions, we always use examples different from the actual scenarios in the exercise.

### 3.3 Step 3: Answer elicitation

We sampled 350 non-redundant questions per scenario, with an equal number from the 5 question positions (see previous section)<sup>4</sup>. Using a different set of annotators than Step 2, we then elicited 10 indirect answers for each question. The annotator instructions are provided in Table 4. For faster annotation, and to encourage diverse answers, we displayed five questions (in same situational context) simultaneously. The five questions were chosen to be diverse (based on cosine similarity between nouns and adjectives in the questions), and annotators were instructed to treat them independently.<sup>5</sup> See Figure 2 for an example display.

Importantly, a key design consideration was that we do *not* instruct annotators to produce an answer with a specific meaning. Rather the annotator composes a natural response without reflecting upon a required meaning. We believe this flexibility is important to ensure that varied meanings (even ambiguous ones) are present in our data. This is verified in our analysis in Section 3.6.

Table 1 provides example answers from this step. Note that the 10 answers have varied meanings departing from definite ‘yes’ and ‘no’. The answers were high quality and diverse in form. 83% of answers appear only once in the corpus. At the same time, there are a few prototypical answers for the different meanings. Below we show the most repeated responses (and frequency) for ‘yes’ and ‘no’ meanings, and also for cases where the answer is conditional upon some situation, and middle ground responses.

‘yes’ answers		‘no’ answers	
59	I would love to	21	I don’t drink
40	Let’s do it	18	I prefer pop
40	That would be good	18	I wish!
‘conditional yes’		‘middle-ground’	
9	If the weather is nice	14	I’m not sure yet
6	If I can afford it	10	Which one?
6	Depends where you want to go	9	It’s OK

These responses indicate that strong lexical signals for meaning are present in the answer.

### 3.4 Step 4: Marking interpretations

Finally, we ask a third set of annotators to mark interpretations for all the QA (question, indirect

<sup>4</sup>For some scenarios, we obtained slightly fewer questions due to high repetitiveness at earlier question positions.

<sup>5</sup>Note that the 10 answers for each question in the Circa corpus are always from 10 different annotators.

#### Annotator Instructions

You will be given a social situation, for example, talking to your friend or neighbour.

**Task:** You will be asked to respond to a question from your friend/neighbour but without using the words ‘yes’ or ‘no’ (or similar words like ‘yeah’, etc). Please provide a possible answer, it does not have to be your real opinion. Rather you should provide a possible answer from which your friend will be able to infer whether you mean ‘yes’, ‘no’, ‘maybe’ or ‘sometimes’.

**Example:** Here are three such answers to your friend’s question about movies.

*Do you like movies with sad endings?*

(a) I often watch them. (Meaning=Yes)

(b) I prefer movies which make me laugh. (Meaning=No)

(c) When the plot is also good. (Meaning=Sometimes)

Table 4: Annotator instructions for Step 3.

#### Annotator instructions

You will be shown short dialogues between two friends/colleagues (X and Y) in a certain context. For example:

Context: X wants to know about Y’s movie preferences.

X: “Do you like movies with sad endings?”

Y: “I often watch them.”

In all the dialogues, X asks a simple ‘Yes/No’ question, and Y answers it with a short sentence or a phrase.

**Task:** We need your help to interpret Y’s answer. Read the dialog and tell us how you think X will interpret Y’s answer. Your options are: *X will think that Y means*

(1) ‘yes’

(2) ‘probably yes’ / ‘sometimes yes’

(3) ‘yes, subject to some conditions’

(4) ‘no’

(5) ‘probably no’

(6) ‘in the middle, neither yes nor no’

(7) ‘I am not sure how X will interpret Y’s answer’

If Y’s response does not fit any of the above, please choose the option (8) ‘Other’, and leave a short comment.

For our example above, the likely interpretation is ‘yes’.

Table 5: Annotator instructions for Step 4.

answer) pairs from Step 2. In particular, they are asked to judge how the question-seeker would interpret the answer provided. As in most NLP tasks, interpretations will vary, and so we obtain *five* annotations per pair.

The correct label categories are not readily clear, but the variety of examples in our corpus made it certain that just ‘yes’ and ‘no’ will not suffice. Building on prior work by [de Marneffe et al. \(2010\)](#), and a pilot experiment, we identified categories that can be annotated reliably. These are shown in Table 5. The annotators were asked to assume the dialogs were co-operative and casual. They were advised to use ‘probably yes/no’ when they cannot infer a definite meaning. If X will not infer a ‘yes’ without some condition being met, then the class ‘yes, subject to some conditions’ was to be chosen. Detailed instructions with exact phrasing,

<b>Yes</b> Q: Do you have any pets? A: My cat just turned one year old.	<b>Probably yes / sometimes yes</b> Q: Do you like mysteries? A: I have a few that I like.	<b>Yes, subject to some conditions</b> Q: Do you enjoy drum solos? A: When someone’s a master.
<b>No</b> Q: Do you have a house? A: We are in a 9th floor apartment.	<b>Probably no</b> Q: Are you interested in fishing this weekend? A: It’s supposed to rain.	<b>In the middle</b> Q: Did you find this week good? A: It was the same as always.

Table 6: Example question and answer pairs where all 5 annotators agreed on the label.

and practice questions are in the appendix (Figure 3, Tables 13 and 14). Annotators took an average of 23 seconds per question-answer pair.

Annotators were also given an option to flag improper content. We remove those QA pairs which were flagged by even one of the five annotators. The authors also read every question, and used a blacklist of words for additional filtering. The remaining 34,268 pairs comprise the final corpus.

### 3.5 Gold standard labels

Each (question, indirect answer) pair was marked by five annotators, so we use majority judgement as the gold standard, subject to *at least three* annotators making that judgement.

We use two aggregation schemes. The STRICT scheme keeps all eight class distinctions from Table 5. A more RELAXED label is computed by collapsing the uncertain classes with the definite ones: ‘probably yes / sometimes yes’ → ‘yes’, ‘probably no’ → ‘no’, and ‘I am not sure how X will interpret Y’s answer’ → ‘In the middle, neither yes nor no’. These classes were commonly confused by the raters. The ‘Other’ class was used mostly when the question was not polar (e.g. disjunctive ones such as ‘Do you like to dine-in or take-out?’). To illustrate the richness of the Circa corpus, Table 6 shows one QA pair from each class.

### 3.6 Label distributions

We now analyze the distribution of the gold standard labels. For STRICT labels (Table 7), only 8% of the examples (marked ‘N/A’) do not receive a majority vote. The most frequent class is ‘yes’ (42% of the data). ‘No’ is less frequent (32%). The third most frequent is ‘conditional yes’, indicating that conditional preferences may be common. The ‘probably’ classes are around 3-4%, each with over a thousand examples in the corpus. There is also a notable number of ‘in the middle’ examples. With the collapsed RELAXED labels (Table 8), the proportion of ‘yes’ and ‘no’ increase slightly, and ‘N/A’ examples drop to only 2%. These distributions reflect the rich patterns in our data.

Label	STRICT	
Yes	14,504	(42.3%)
No	10,829	(31.6%)
Probably yes / sometimes yes	1,244	(3.6%)
Yes, subject to some conditions	2,583	(7.5%)
Probably no	1,160	(3.4%)
In the middle, neither yes nor no	638	(1.9%)
I am not sure	63	(0.2%)
Other	504	(1.5%)
N/A	2,743	(8.0%)

Table 7: Distribution of STRICT gold standard labels. ‘N/A’ indicates lack of majority agreement.

Label	RELAXED	
Yes	16,628	(48.5%)
No	12,833	(37.5%)
Yes, subject to some conditions	2,583	(7.5%)
In the middle, neither yes nor no	949	(2.8%)
Other	504	(1.5%)
N/A	771	(2.2%)

Table 8: Distribution of RELAXED gold standard labels. ‘N/A’ indicates lack of majority agreement.

### 3.7 Annotator agreement

The Fleiss kappa scores are 0.61 for STRICT and 0.76 for RELAXED labels (p-values < 0.0001) indicating substantial agreement. In fact, full agreement, as in Table 6 where all five annotators agree on the STRICT class, occurs for 49% of pairs, a high proportion given the complexity of the task. When the labels are RELAXED, this reaches 71%. The full agreement distributions are:

Agreement	STRICT	RELAXED
5 annotators	49.1%	71.0%
4 annotators	23.8%	17.2%
3 annotators	19.1%	9.6%

### 3.8 Dialog scenarios and question type

As one would expect, different scenarios prompt different types of questions and answers, and hence different label distributions. For the ‘friend switching jobs’ scenario, 54.5% of elicited answers are have ‘yes’ meaning (RELAXED labels). For book preferences, this is only 42%.

Perhaps more unexpected is that a few questions

have labels predominantly of the same kind. For example, all 10 answers to ‘*Ready for the weekend?*’ receive a ‘yes’ label, and they are all ‘no’ for ‘*Are you offering the asking price?*’. While the first question is largely rhetorical, the second involves common sense: Most people negotiate real-estate prices. We found that 3% of questions (95) have all answers with the same label, and for another 20%, 8-9 answers have the same label. A model would still need to identify the label as either ‘yes’ or ‘no’, but these skews indicate that there may be some (weak) signals for the label in the question itself. After all, *who isn’t ready for the weekend?*

### 3.9 Annotation Protocol

While our annotation method is comprehensive, one might wonder how alternative approaches would have fared. We have not performed controlled tests of different approaches but we briefly document our choices to aid future work.

We performed pilot annotations for each step of our process. Our goal was to start with less restrictive settings, where annotators are given minimal and simple instructions. If we did not receive quality responses, we intended to give more specific directions. For example, in an initial pilot, sometimes annotators gave long answers which may be unrealistic in conversations eg. ‘*I can tell you, without a shadow of a doubt, that there are very few things that I enjoy more than sitting in front of my computer.*’ So in later annotations for all tasks, we added instructions ‘to keep the conversation casual’ which at least reminds annotators about this need. We performed spot checks on the results, but did not perform controlled tests.

For answer elicitation, our simple instruction pilot produced answers with diverse meanings. The distribution of meanings also varied according to the question (some questions had a mix of meanings, others were skewed towards a few meanings). So we decided against explicitly asking annotators to provide a certain meaning as that would create uniform meaning distributions, which may turn out unrealistic and miss dominant tendencies. We also considered that the explicit approach may reduce answer quality (for example, if a person who eats meat were to have to answer indirectly that they are vegetarian, they may be more likely to get their facts wrong, or try to rush through the task). In fixing our choice, we took care to ensure that quality was not affected, and answers were varied.

## 4 Learning Task

Obviously answers contain numerous cues for subsequent dialog flow, but in this first work we focus on meaning prediction. Specifically, given a question-answer pair, we classify it into one of the meaning categories in Tables 7 and 8.

We consider two experimental settings: In the *matched* setup, we assume that the response scenario is seen during training (randomly dividing our corpus examples into 60% training, 20% each for development/test). The *unmatched* setting is aimed at understanding the performance on unseen scenarios, i.e., whether models can learn the semantics of yes/no answers beyond the patterns specific to individual scenarios. As our data contains 10 scenarios, we carry out 10 leave-one-out tasks, each time holding out one scenario (for example, ‘buying a flat in New York’) as the test data, and use the remaining nine for training and development.

For both the *matched* and *unmatched* setting, we consider two variants of the classification problem: STRICT (with 6 different labels, namely all except ‘other’ and ‘N/A’ in Table 7) and RELAXED with 4 labels (Table 8). We ignore the examples without a majority label and also those marked ‘unsure’ or ‘other’. Thus our experiment data sizes are:

Experimental Setting	Train	Dev.	Test
STRICT-matched	18,574	6,192	6,192
RELAXED-matched	19,795	6,599	6,599
STRICT-unmatched	24,746	3,115	3,095
RELAXED-unmatched	26,404	3,289	3,299

For the *unmatched* setting, these sizes are the average across the 10 leave-one-out sets.

## 5 Models

Building upon recent NLI systems, our approach leverages representations from unsupervised pre-training, and finetunes a multiclass classifier over the BERT model (Devlin et al., 2019). However, we first consider other models for related tasks.

### 5.1 Related baselines and corpora

**BOOLQ** is a question-answering dataset focused on factual yes/no questions (Clark et al., 2019). Here yes/no questions from web search queries are paired with Wikipedia paragraphs that help answer the question. There are 9.4k train, 3.2k development, and 3.2k test set examples, with two target classes, namely ‘yes’ and ‘no’. We train our own BOOLQ model with BERT pre-training. It reaches

a development accuracy of 74.1%<sup>6</sup>. This model only predicts two classes ‘yes’ and ‘no’.

**MNLI.** The MultiNLI corpus (Williams et al., 2018) is a large corpus for textual entailment. It consists of premise-hypothesis sentence pairs which are marked with three target classes ‘entailment’, ‘contradiction’ and ‘neutral’. There are 392K train, 9K dev., and 9K test examples.

Although not applicable to all indirect answers, semantic consequence can be leveraged for interpreting some of them. For the question (Q) ‘*Do you like Italian food?*’, consider two possible answers (A) ‘*I love Tuscan food.*’ and (B) ‘*I prefer Mexican cuisine.*’. Let  $Q'$  be the declarative (positive) sentence derived from Q i.e.  $Q' = 'I like Italian food'$ . ( $Q'$  can be obtained by inverting the subject and auxiliary, and changing pronoun to first person). The meaning of A and B above, can then be obtained from an entailment system:  $A \implies Q'$ , hence ‘yes’, while B contradicts  $Q'$ , hence ‘no’.

We thus obtain predictions from an MNLI system, and map the predicted three NLI classes in a post-processing step: ‘contradiction’  $\rightarrow$  ‘no’, ‘entailment’  $\rightarrow$  ‘yes’, and ‘neutral’  $\rightarrow$  ‘in the middle’. Note that this approach cannot predict all the classes in the corpus. Before prediction, we rewrite our questions into declarative form using syntactic rules on a constituency parse. Performance is much worse without this rewriting. Our models for the MNLI task start from a BERT checkpoint and reach a development accuracy of 84%.

This model improves with the syntactic rewriting of questions which we are able to do fairly accurately. We based the rewrite rules on 50 questions. On a different set of 50 questions, 38 were rewritten fully accurately (manual inspection). Some errors arose from incorrect parsing and some are due to deficient rules. For example, we do not handle verb re-inflection. So ‘*Did you enjoy the movie?*’ gets rewritten into ‘*I did enjoy the movie.*’ rather than ‘*I enjoyed the movie.*’ This rewriting technique helped only the MNLI model. For other finetuning based models, which involve training, the models are able to learn from the question form itself.

**Majority baseline.** This method predicts the most frequent label, ‘yes’, for all examples.

## 5.2 Training with Question or Answer only

**Answer only.** In many cases, the answer to a question suffices for predicting the label (see Table

1 and Section 3.3); for example “*I would like that.*” or “*I wish!*”. This *Answer only* model fine-tunes BERT to predict the class based only on the answer. Similar experiments are done on NLI datasets to test if the hypothesis alone is at times sufficient for entailment prediction (Poliak et al., 2018; Gururangan et al., 2018). Such results are problematic for entailment (since it is *defined* to depend on the truth of the premise). In contrast, our problem is primarily about the meaning of answers. This experiment will provide insight into the cues within indirect responses, an aspect not understood so far.

**Question only.** Some questions commonly elicit certain answers (see Section 3.8). These models test how well the question predicts the label.

## 5.3 Question-Answer Pair models

These models take both the question and the answer. They all finetune BERT checkpoints, and the question-answer pair is passed with a separator.

**BERT-YN** is BERT finetuned only on our Circa corpus (YN).

We also explore how to transfer the strength of parameters learned for three related inference tasks.

**BERT-BOOLQ-YN** finetunes a BOOLQ model checkpoint (see Section 5.1) on our corpus, with a new output layer. Since BOOLQ is a Yes/No question answering system, even if developed for a different domain, we expect to learn many semantics of yes/no answers from this data.

**BERT-MNLI-YN** is first fine-tuned on the MNLI corpus, followed by our YN data. This configuration tests if the signals we hoped to capture with the out-of-the-box MNLI model (Section 5.1) can be strengthened by training on our target task.

**BERT-DIS-YN.** As discussed, indirect answers also have discourse relations with the speaker’s intent (Green and Carberry, 1999). We implement this idea via a discourse connective prediction task. Consider again the texts from Section 5.1: The likely connective between  $Q'$  and A is ‘because’ as in ‘*I like Italian food [because] I love Tuscan food.*’. For  $Q'$  and B, ‘but’ would be more reasonable: ‘*I like Italian food [but] I prefer Mexican cuisine.*’. We hypothesize that these discourse relations will help learn the yes/no meaning via transfer learning.

We use 400K examples (matching the MNLI data size) of explicit connectives and their arguments (a subset of Nie et al. (2019)). We aim to predict the 5 connectives (because, but, if, when,

<sup>6</sup>Clark et al. (2019) report 78% (dev.) with BERT-Large.

and) based on their arguments. This task itself can be done with a development accuracy of 82%. The best checkpoint is then finetuned on YN data.

## 6 Experiments

We use pre-trained BERT models (with 12 layers, 768 hidden units, and 12 attention heads, 110M parameters) for all our experiments. The experiments were done on a single Cloud TPU, and finetuning on our corpus takes under 30 minutes.

### 6.1 Setup and Hyperparameter Tuning

For the base models (MNLI, BOOLQ, and DIS), we tuned the learning rate (values  $5e-5$ ,  $3e-5$ ,  $2e-5$ ), the number of epochs (2, 3, 4), and train batch size (16, 32) in an exhaustive combination. For finetuning on yes/no data, we tune the learning rate while setting the epochs to 3 and training batch size to 32. We also perform three random restarts in each configuration. Performance was stable across the restarts (accuracy variation  $\leq 1\%$ ). So we take the best model on the development set as our final model. The best hyperparameters are in the Appendix. For the unmatched setting, we do 10 leave-one-out tasks. Here we take the best parameters from the matched setting, and use that configuration to train all 10 experiments.

### 6.2 Results

We report the development accuracy and detailed test results in Table 9 (RELAXED setting) and Table 10 (STRICT setting). For the unmatched setting, we report the mean accuracy and standard deviation across the 10 folds, and the min and max values. We first discuss results for the RELAXED labels. The findings for STRICT are similar.

The majority baseline (‘yes’ class) leads to an accuracy of 49%. The MNLI to yes/no label mapping (no finetuning) is reasonable in terms of F-score for the ‘no’ class, but is poor for ‘yes’. BOOLQ is the best baseline with 63% accuracy. However, there is no recall for labels other than ‘yes’ and ‘no’.

The question-only and answer-only results are noteworthy. The question-only model outperforms the majority baseline. On the other hand, the answer text contains strong signals, with 82% accuracy, or about 20% better than the best baseline.

But models using both question and answer reach 5-6% greater accuracy and significantly outperform the answer-only model (McNemar’s test, p-values  $< 1e-6$ ). As expected, these joint models

are necessary when a string is a possible answer to multiple questions. An answer-only model is easily misled in these cases, as the examples below show:

(1) “Is there something you absolutely won’t eat?” “I really dislike bananas.”  
Answer-only prediction: ‘No’  
Question+Answer prediction: ‘Yes’

(2) “Do you need a nap?”  
“I have plenty of energy.”  
Answer-only prediction: ‘Yes’  
Question+Answer prediction: ‘No’

The best F-scores are obtained by an MNLI transfer task, reaching 88.2% accuracy in the matched setting. But it is not significantly better than a no-transfer BERT-YN model (McNemar’s test).

The unmatched setting shows that the models are worse when a scenario has not been seen in the training data. While it may not be possible for every conversational system to generalize across scenarios, a semantic classification such as yes/no should ideally be robust to such changes. Instead we see a 6-10% accuracy gap between the in-scenario test accuracies, and minimum out-of-scenario accuracy. The best accuracies are reached by a MNLI model. The highest accuracy on a scenario is 90% (‘music preferences’, ‘weekend activities’), and lowest is 82% (‘buying a flat in New York’ and ‘switching jobs’). The latter scenarios are quite different than the rest, indicating scope for improving the models.

The ‘in the middle’ class has much worse results compared to the rest. The class has low frequency but also comprises responses of different types. Uncertain responses such as ‘*I am not sure.*’ appear easy to classify. But responses which do not take a stance: ‘*Do you know if it’s raining outside? I’m prepared regardless.*’ are harder. Sometimes, the interpretation is left to the judgement of the listener. Eg. ‘*travelling an hour away*’ could be interpreted as ‘far away’ or ‘close by’ depending on the context and perceptions of the listener. These cases need models to deeply connect the question and answer, and are missed by our technique.

The general trends for STRICT labels is similar: The best accuracy is again reached with MNLI pretraining. It is 85% for the matched case, a small but significant gain over BERT-YN (McNemar’s test, p-value  $< 0.02$ ). The accuracy is 10% lower (74%) for hardest experiment in leave-one-out.

For STRICT labels, the ‘probably no’ class is hardest to predict even though it is as frequent as ‘probably yes’ and close to double the size of the ‘in the middle’ class. We found that ‘probably no’



Model	Accuracy		Matched setting Test F-Score				Unmatched setting Test Accuracy			
	Dev.	Test	Yes	No	C.Yes	Mid	Mean	Std.	Min.	Max.
<b>Baselines (no finetuning)</b>										
Majority class	50.2	49.3	66.0	0.0	0.0	0.0	50.4	4.3	43.6	56.8
MNLI	28.4	28.9	34.4	52.8	0.0	6.9	28.1	2.8	24.2	34.1
BOOLQ	64.2	<b>62.7</b>	<b>71.1</b>	<b>59.6</b>	0.0	0.0	<b>63.3</b>	2.7	<b>58.3</b>	<b>66.5</b>
<b>BERT finetuned on Question or on Answer</b>										
BERT-YN (Question only)	56.4	56.0	63.1	54.1	9.1	1.0	53.3	2.9	48.0	58.4
BERT-YN (Answer only)	83.0	<b>81.7</b>	<b>83.9</b>	<b>80.3</b>	<b>88.9</b>	<b>18.6</b>	<b>80.1</b>	5.8	<b>71.4</b>	<b>87.8</b>
<b>BERT finetuned on Question + Answer</b>										
BERT-YN	88.4	<u>87.8</u>	89.8	87.9	89.9	28.2	85.5	3.9	79.0	90.2
BERT-MNLI-YN	89.6	<b>88.2</b>	<b>90.4</b>	<b>88.5</b>	89.3	29.4	<b>87.1</b>	3.0	<b>81.9</b>	<b>90.3</b>
BERT-DIS-YN	88.0	87.4	89.4	87.4	<b>90.0</b>	<b>35.2</b>	85.5	3.5	78.9	89.4
BERT-BOOLQ-YN	87.7	87.1	89.0	86.9	89.6	30.9	85.3	3.7	78.8	89.4

Table 9: Performance on the **relaxed** labels. The highest value in each column is in bold. For matched setting, we show dev. and test accuracies, and also F-scores for the 4 labels (‘yes’, ‘no’, ‘conditional yes’ and ‘in the middle’). In unmatched setting, we report summaries of 10 leave-one-out experiments. *BERT-YN is significantly better than ‘Answer only’ (McNemar’s test,  $p$ -value  $< 1e-6$ ); BERT-MNLI-YN is not significantly better than BERT-YN.*

Model	Accuracy		Matched setting Test F-Score				Unmatched setting Test Accuracy					
	Dev.	Test	Yes	P.Yes	C.Yes	No	P.No	Mid	Mean	Std.	Min.	Max.
<b>Baselines (no finetuning)</b>												
Majority class	47.5	47.0	63.9	0.0	0.0	0.0	0.0	0.0	46.9	3.9	40.0	52.3
MNLI	26.3	27.4	36.6	0.0	0.0	53.0	0.0	4.9	26.4	3.2	21.7	32.7
BOOLQ	59.4	<b>59.2</b>	<b>70.4</b>	0.0	0.0	<b>57.0</b>	0.0	0.0	<b>58.9</b>	3.0	<b>53.8</b>	<b>63.7</b>
<b>BERT finetuned on Question or Answer</b>												
BERT-YN (Question)	53.7	52.8	62.3	3.2	19.7	51.1	0.0	4.7	49.4	4.0	41.9	56.7
BERT-YN (Answer)	77.3	<b>77.8</b>	<b>82.5</b>	<b>49.5</b>	<b>90.2</b>	<b>77.3</b>	<b>16.2</b>	<b>26.9</b>	<b>75.8</b>	5.8	<b>65.4</b>	<b>82.8</b>
<b>BERT finetuned on Question + Answer</b>												
BERT-YN	83.6	<u>84.0</u>	88.7	49.9	<b>90.2</b>	85.4	18.6	42.6	81.2	4.6	71.8	85.6
BERT-MNLI-YN	85.0	<b>84.8</b>	<b>89.8</b>	<b>51.8</b>	89.8	<b>86.6</b>	18.0	41.3	<b>82.8</b>	4.0	<b>74.4</b>	<b>86.7</b>
BERT-DIS-YN	83.8	83.3	87.9	50.2	90.5	84.1	21.2	<b>50.8</b>	81.5	4.5	73.1	86.3
BERT-BOOLQ-YN	83.1	83.4	88.2	51.2	89.1	84.5	<b>22.1</b>	43.7	81.1	4.3	73.3	85.8

Table 10: Performance on the **strict** labels. The highest value in each column is in bold. *BERT-YN is significantly better than ‘Answer only’ (McNemar’s test,  $p$ -value  $< 1e-6$ ), and BERT-MNLI-YN is better than BERT-YN ( $p$ -value  $< 0.02$ ).*

examples are heavily (69%) mis-predicted into the ‘no’ class. Utterances which explicitly convey uncertainty eg. (*‘I don’t believe so’*) or comparison (*‘Is everything good? Not the greatest.’*) are somewhat easier to predict. On the other hand, in:

“Have you ever bought a romance novel?” “I prefer to read horror books.”  
Best model: “No”  
Gold standard: “Probably no”

The speaker prefers horror genre, but it does not preclude ever buying a romance novel. These subtleties are understandably harder for systems.

Overall, while MNLI based transfer learning has led to small improvements, incorporating the right information for the task remains a challenge.

## 7 Conclusion

We have presented a new dataset containing natural indirect yes/no answers, as well as other significant pragmatic moves in the form of conditionals and uncertain utterances. Our first approach towards automatic interpretation is promising, but there is a significant gap especially for examples outside training scenarios. Our model does not yet classify additional information in responses (*‘Dinner? Let’s go for Italian.’* indicates not only a ‘yes’ answer, but also a preference for Italian food.). Moreover, we have explored the phenomena in English. There are exciting avenues for multilingual work to account for language and cultural differences.

## References

- Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a Human-Like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977*.
- Penelope Brown and Stephen C Levinson. 1978. Universals in Language Usage: Politeness Phenomena. In E.N. Goody, editor, *Questions and Politeness: Strategies in Social Interaction*, pages 56–311. Cambridge University Press.
- Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau 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.
- 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.
- Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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 4171–4186.
- Nancy Green and Sandra Carberry. 1999. Interpreting and Generating Indirect Answers. *Computational Linguistics*, 25(3):389–435.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.
- Julia B. Hirschberg. 1985. *A Theory of Scalar Implicature*. Ph.D. thesis, University of Pennsylvania.
- Beth Ann Hockey, Deborah Rossen-Knill, Beverly Spejowski, 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*.
- 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: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 136–143.
- 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.
- Allen Nie, Erin Bennett, and Noah D. Goodman. 2019. DisSent: Learning Sentence Representations from Explicit Discourse relations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 4497–4510.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis-Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.
- 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 Spejowski, Beth Ann Hockey, Stephen Isard, and Matthew Stone. 1997. Yes/No Questions and Answers in the Map Task Corpus. Technical report, University of Pennsylvania.
- Anna-Brita Stenström. 1984. Questions and Responses in English Conversation: Lund Studies in English. *Claes Schaar and Jan Svartvik, editors, Lund Studies in English 68, CWK Gleerup, Malmö Sweden*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *the Proceedings of ICLR*.
- 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.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. *arXiv preprint arXiv:1911.00536*.

## A Hyperparameter settings

For the base models (MNLI, BOOLQ, DIS), we tuned the learning rate (values  $5e-5$ ,  $3e-5$ ,  $2e-5$ ), the number of epochs (2, 3, 4), and train batch size (16, 32) in an exhaustive combination of these parameters. The best performance on development data was obtained with the following settings:

Model	learning rate	no. epochs	batch size
MNLI	$2e-5$	3	16
BOOLQ	$3e-5$	4	16
DIS	$2e-5$	2	32

For finetuning on yes/no data (matched setting), we tune the learning rate while setting the number of epochs to 3, and training batch size to 32. We also perform three random restarts in each configuration. Performance was stable across the restarts (accuracy variation  $\leq 1\%$ ). We take the best model on the development set as our final model. The chosen learning rates are in the table below:

Model	STRICT	RELAXED
BERT-YN (Question only)	$2e-5$	$3e-5$
BERT-YN (Answer only)	$2e-5$	$2e-5$
BERT-YN	$3e-5$	$3e-5$
BERT-MNLI-YN	$2e-5$	$5e-5$
BERT-DIS-YN	$3e-5$	$5e-5$
BERT-BOOLQ-YN	$3e-5$	$5e-5$

For the unmatched setting, we do 10 leave-one-out experiments. Here we use the best parameters from the matched setting, and use the same configuration for training in all 10 experiments.

## B Annotation Instructions

We now detail the complete instructions to annotators, along with interface examples, prompt texts, and practice items.

**Question collection.** In this step, an annotator is shown a scenario, and asked to provide five yes/no questions. The instructions, and interface for an example item are in Figure 1. The descriptions used for each of the 10 dialog scenarios are in Table 11.

**Answer elicitation.** Similarly, the instructions and interface for collecting answers are in Figure 2. Note that we show 5 questions on each screen, to reduce annotation time. The 5 questions are taken from the same scenario but such that they are not too similar. We enforce non-redundancy by keeping the pairwise similarity between any two questions on the same screen to less than 0.35 (measured by cosine similarity between the adjectives and nouns in the questions).

We have the same 10 scenarios but their descriptions are changed slightly to suit to the answer elicitation task. These prompts are in Table 12.

**Marking interpretations.** Finally, the question-answer pairs are annotated with meaning categories. Our complete instructions and annotator interface are in Figure 3. Again, the scenario descriptions are modified for the task, and are given in Table 13.

This step is fairly complex, so every annotator also worked through 8 practice questions before starting the annotation. After they answered them, the correct answers were shown along with an explanation. These examples are given in Table 14.

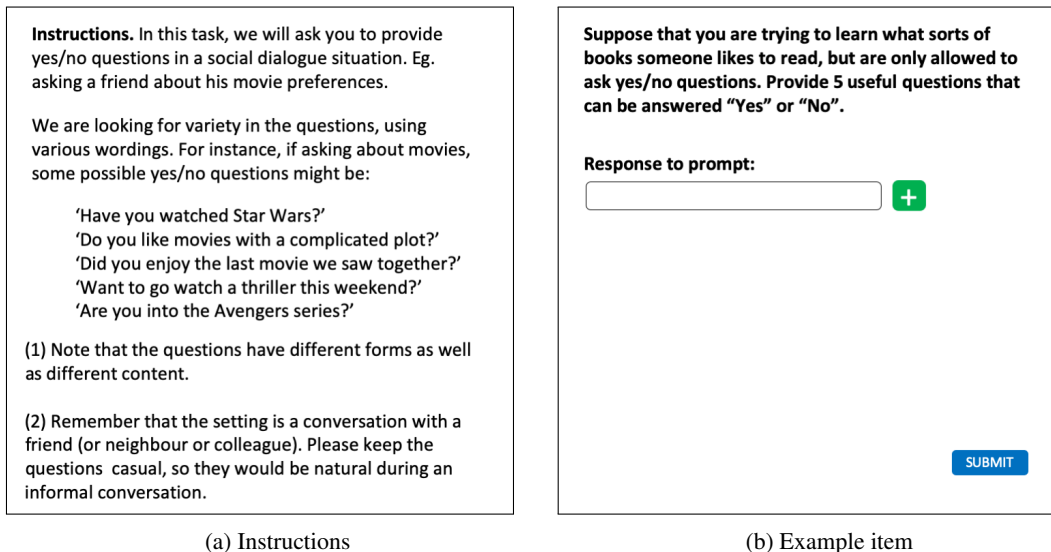


Figure 1: Annotator interface for question collection.

- |   |  |
|---|--|
| <p>(1) Suppose that you are trying to learn about a friend's food preferences, so that you can recommend a local restaurant, but can only ask yes/no questions. Provide 5 useful questions that can be answered "Yes" or "No".</p> <p>(2) Suppose that you are trying to learn what sorts of activities a friend likes to do during the weekend, so that you can recommend local activities that might interest them, but can only ask yes/no questions. Provide 5 useful questions that can be answered "Yes" or "No".</p> <p>(3) Suppose that you are trying to learn what sorts of books someone likes to read, but are only allowed to ask yes/no questions. Provide 5 questions you might ask about books that can be answered "Yes" or "No".</p> <p>(4) Suppose that you are meeting your new neighbour for the first time. You want to find out more about him/her, but you are only allowed to ask yes/no questions. Provide 5 questions you might ask him/her that can be answered "Yes" or "No".</p> <p>(9) Suppose that your friend tells you he is thinking of buying a flat in New York. In this context, provide 5 questions you might ask him/her. You are only allowed to ask yes/no questions.</p> | <p>(5) On a Friday evening, you are leaving work and see your friend (and colleague) also at the door ready to leave. Provide 5 questions you might ask him/her that can be answered "Yes" or "No".</p> <p>(6) Suppose that you are trying to learn about a friend's interests related to music. For instance, you could ask about music tastes, instruments played, events they go to, etc. You are only allowed to ask yes/no questions. Provide 5 questions you might ask.</p> <p>(7) Your friend has arrived from out of town to visit you. Provide 5 questions you might ask your friend when he/she arrives and during your time together. You are only allowed yes/no questions.</p> <p>(8) Suppose that you are at a cafe, and you run into your childhood neighbour. You haven't seen each other or had any contact for many years. Ask a few questions of your childhood neighbour. You are only allowed yes/no questions. Provide 5 questions you might ask.</p> <p>(10) Your friend (not a colleague) is considering switching his/her job. You do not know much about the aspects of his/her job. Ask a few yes/no questions in this context. Provide 5 yes/no questions.</p> |
|---|--|

Table 11: Descriptions of the 10 scenarios in the question collection step.

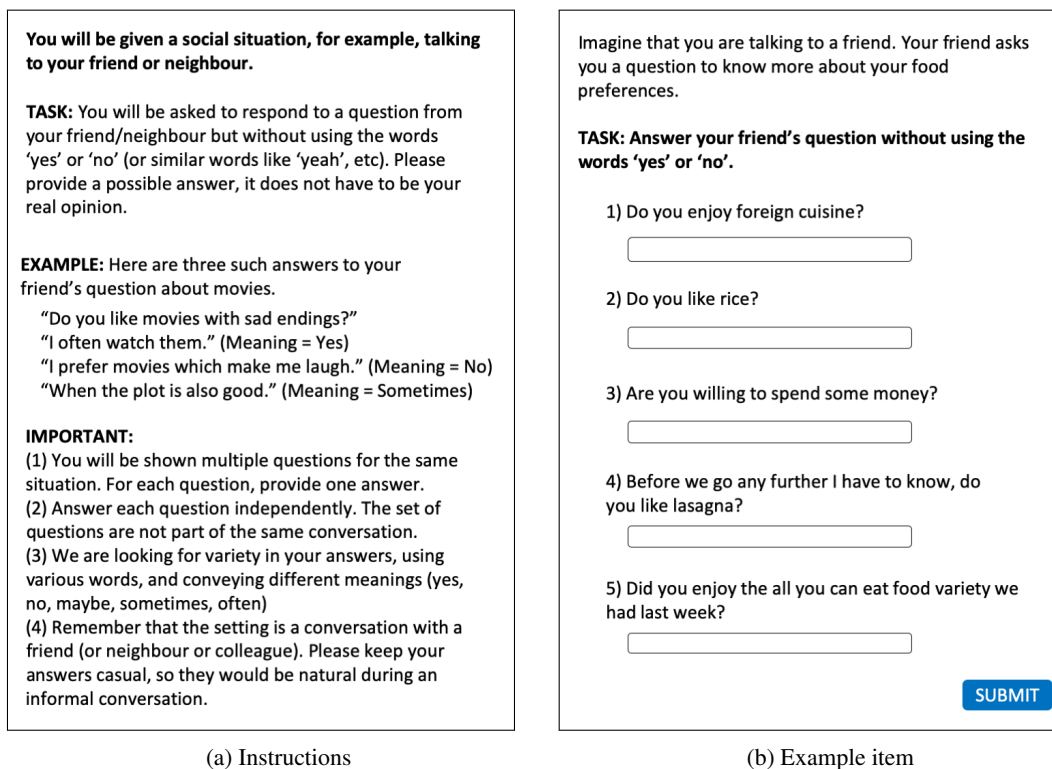


Figure 2: Annotator interface for answer elicitation.

- |  |   |
|--|---|
| <p>(1) Imagine that you are talking to a friend. Your friend asks you a question to know more about your food preferences. Answer your friend's question without using the words 'yes' or 'no'.</p> <p>(2) Imagine that you are talking to a friend. Your friend asks you a question to know more about what activities you like to do during weekends. Answer your friend's question without using the words 'yes' or 'no'.</p> <p>(3) Imagine that you are talking to a friend. Your friend asks you a question to know more about what sorts of books you like to read. Answer your friend's question without using the words 'yes' or 'no'.</p> <p>(4) Imagine that you have just moved into a neighbourhood. One of your new neighbours is a friendly person, and asks you a question. Answer your neighbour's question without using the words 'yes' or 'no'.</p> <p>(9) Imagine that you have just told your friend that you are thinking of buying a flat in New York. Your friend asks you a question to know more about your plans. Answer your friend's question without using the words 'yes' or 'no'.</p> | <p>(5) Imagine that at the end of the week you are leaving work and see your friend (and colleague) also at the door ready to leave. Your friend asks you a question. Answer your friend's question without using the words 'yes' or 'no'.</p> <p>(6) Imagine you are talking to a friend. Your friend asks you a question to know more about your interests related to music. Answer your friend's question without using the words 'yes' or 'no'.</p> <p>(7) Imagine that you have just travelled from a different city to visit your friend. Upon your arrival, your friend asks you a question. Answer your friend's question without using the words 'yes' or 'no'.</p> <p>(8) Imagine that you run into your childhood neighbour at a cafe. You haven't seen each other or had any contact for many years. Your childhood neighbour asks you a question. Answer your neighbour's question without using the words 'yes' or 'no'.</p> <p>(10) Imagine that you have just told your friend that you are considering switching your job. Your friend asks you a question to know more about your plans. Answer your friend's question without using the words 'yes' or 'no'.</p> |
|--|---|

Table 12: Descriptions of the 10 scenarios as used in the answer elicitation step.

**Please read these instructions carefully:**

You will be shown short dialogues between two friends/colleagues (X and Y) in a certain context. For example:

**Context** : X wants to know about Y's movie preferences.  
**X** : "Do you like movies with sad endings?"  
**Y** : "I often watch them."

In all the dialogues, X asks a simple 'Yes/No' question, and Y answers it with a short sentence or a phrase.

**Task:** We need your help to interpret Y's answer

Read the dialog and tell us how you think X will interpret Y's answer. Your options are:

X will think that Y means

1. Yes
2. Probably yes' / 'sometimes yes
3. Yes, subject to some conditions
4. No
5. Probably no
6. In the middle, neither yes nor no
7. I am not sure how X will interpret Y's answer

If Y's response does not fit any of the above, please choose the option 'Other' and leave a short comment.

**Example annotation:**

**Context** : X wants to know about Y's movie preferences.  
**X** : "Do you like movies with sad endings?"  
**Y** : "I often watch them."

X will think that Y means (1) Yes

Previous 1 2 Next

(a) Instructions

**Note (please read carefully):**

- X and Y are friends. All their conversations are in casual contexts.
- The 'Context' is provided to tell you about the situation where the dialogue happened.
- You should assume that Y has good intentions and always provided a relevant answer to the question
- Think only about how X will interpret Y's answer in the given context and to the given question. Do not make a choice based on what the answer means in general.
- Use 'probably yes/sometimes yes' when the answer is not a definite yes. Similarly for 'probably no'.
- Use 'yes, subject to some conditions' only if X will not interpret a 'yes' without that condition being satisfied.
- Use the options 'probably yes', 'sometimes yes' and 'yes, subject to some conditions' only when X will not interpret the answer as a definite 'yes'.

**Eg. X** : "Are you up for a movie?"  
**Y** : "Only to a comedy."  
 X will think that Y means  
 (3) yes, subject to some conditions.

In contrast, in the following, Y means 'yes' even if his/her answer also contains a condition.  
**X** : "Have you ever seen a Star Wars movie?"  
**Y** : "Only every time I'm persuaded by my Star Wars fan wife."  
 X will think that Y means  
 (1) yes (because Y has seen at least one Star Wars movie)

- If the dialog contains offensive, obscene or sensitive content, please mark the checkbox provided. Such examples will be rare if any.

Previous 1 2 Next

(b) Additional notes

Instructions

**Context** : Y has just travelled from a different city to meet X

**X** : "Did you stop anywhere on the way?"  
**Y** : "I try to incorporate my errands into every trip I make."

X will think that Y means

Yes  
 Probably yes / sometimes yes  
 Yes, subject to some conditions  
 No  
 Probably no  
 In the middle, neither yes nor no  
 I am not sure how X will interpret Y's answer  
 Other

Please enter any comments you wish to make (optional)

This dialog contains offensive/sensitive content

SUBMIT

(c) Example item

Figure 3: Annotator interface for marking interpretation.

(1) X wants to know about Y's food preferences.	(5) X and Y are colleagues who are leaving work on a Friday at the same time.
(2) X wants to know what activities Y likes to do during weekends.	(6) X wants to know about Y's music preferences.
(3) X wants to know what sorts of books Y likes to read.	(7) Y has just travelled from a different city to meet X.
(4) Y has just moved into a neighbourhood and meets his/her new neighbour X.	(8) X and Y are childhood neighbours who unexpectedly run into each other at a cafe.
(9) Y has just told X that he/she is thinking of buying a flat in New York.	(10) Y has just told X that he/she is considering switching his/her job.

Table 13: Descriptions of the 10 scenarios as used in the interpretation marking step.

<p>(1) Context: X wants to know about Y's food preferences.  X: "Do you eat red meat?"  Y: "I am a vegetarian."  Answer: No (Vegetarians do not eat meat. So X will interpret it as a 'no' answer.)</p>
<p>(2) Context: X wants to know about Y's food preferences.  X: "Have you had bulgogi?"  Y: "I am not sure."  Answer: In the middle, neither yes nor no. (Here Y's response is non-committal. So X's best option is to interpret it as neither a 'yes' nor a 'no')</p>
<p>(3) Context: Y has just travelled from a different city to meet X.  X: "Did you stop anywhere on the way?"  Y: "I had to get gas a few times."  Answer: Yes (X will infer a definite 'yes' because Y stopped for gas.)</p>
<p>(4) Context: Y has just travelled from a different city to meet X.  X: "Did you stop anywhere on the way?"  Y: "I try to incorporate my errands into every trip I make."  Answer: Probably yes / sometimes yes (In response to the question, Y mentions that his general tendency is to do errands on the way. So it is likely that he did the same on this trip. Hence 'probably yes')</p>
<p>(5) Context: X wants to know about Y's music preferences.  X: "Would you go to a punk rock show?"  Y: "It depends on who is playing."  Answer: Yes, subject to some conditions (Y might go depending on the artist who is performing. Hence X will interpret that Y is posing a condition for going to the punk show.)</p>
<p>(6) Context: X wants to know about Y's music preferences.  X: "Would you go to a punk rock show?"  Y: "I think I'd enjoy something less chaotic a little more."  Answer: Probably no (Y's indicates that he'd prefer some other activity. But he does not completely rule out the possibility of going to a punk show. Hence X will interpret it as 'probably no'.)</p>
<p>(7) Context: X and Y are colleagues who are leaving work on a Friday at the same time.  X: "Is your department busy this time of year?"  Y: "We usually end up working overtime."  Answer: Yes (Since Y is regularly working overtime, X will infer that Y's department is busy.)</p>
<p>(8) Context: X and Y are colleagues who are leaving work on a Friday at the same time.  X: "Is your department busy this time of year?"  Y: "Just as busy as the rest of the year."  Answer: In the middle, neither yes nor no (Since it is not known how busy Y's department is in general, it is unclear if Y is busy at this time. In such cases pick 'in the middle, neither yes nor no'. Note that if X had background knowledge that Y's department is usually never busy, he could interpret the answer as a 'no' or vice versa. You do not have to assume such information is available when providing your answer.)</p>

Table 14: The 8 practice questions used to train annotators for marking interpretations.