Unravelling Indirect Answers to Wh-Questions: Corpus Construction, Analysis, and Generation

Zulipiye Yusupujiang and **Jonathan Ginzburg**
Université Paris Cité, CNRS, Laboratoire de Linguistique Formelle
zulipiye.yusupujiang@linguist.univ-paris-diderot.fr
yonatan.ginzburg@u-paris.fr

Abstract

Indirect answers, crucial in human communication, serve to maintain politeness, avoid conflicts, and align with social customs. Although there has been a substantial number of studies on recognizing and understanding indirect answers to polar questions (often known as yes/no questions), there is a dearth of such work regarding *wh*-questions. This study takes up the challenge by constructing what is, to our knowledge, the first corpus of indirect answers to *wh*-questions. We analyze and interpret indirect answers to different *wh*-questions based on our carefully compiled corpus. In addition, we conducted a pilot study on generating indirect answers to *wh*-questions by fine-tuning the pre-trained generative language model DialoGPT (Zhang et al., 2020). Our results suggest this is a task that GPT finds difficult.

1 Introduction

Indirect answers (INDs) to questions hold a distinctive position in the realm of human communication, as they provide related or implied information instead of offering the speaker's intentions or knowledge directly through an utterance's *grammatically governed content.* (i.e., *literal content*) (Ginzburg et al., 2022). Grasping the intrinsic nuances of indirect answers and accurately deducing the expected direct answer from them is essential to facilitate effective communication and information sharing between dialogue participants.

It is a natural part of human communication to produce and understand indirect answers. People use indirect speech to maintain politeness, avoid confrontations, adhere to social norms, or convey information without explicitly stating it (Searle, 1975; Brown et al., 1987). However, understanding and generating indirect answers to questions can be quite challenging for dialogue systems. To

engage in human-like conversation, these systems must be able to grasp the conversational context, background information, and relationships between participants. By accurately interpreting the meaning behind an indirect answer, the system can then provide a more appropriate response, contributing to a more natural interaction.

In the field of dialogue studies, considerable attention has been given to the interpretation and generation of indirect answers to polar questions (Green and Carberry, 1994a,b, 1999; de Marneffe et al., 2009, 2010; de Marneffe and Tonhauser, 2016; Louis et al., 2020; Damgaard et al., 2021). However, there still exists a gap when it comes to the identification and interpretation of indirect answers to *wh*-questions. Studying indirect answers to *wh*-questions is a challenging task for several reasons: a). Unlike polar questions that have only *yes* or *no* (or rather the propositions they convey in context) as direct, resolving answers, *wh*-questions can have a wide range of possible direct answers. This makes it harder to interpret indirect answers to *wh*-questions; b). Compiling a corpus of indirect answers to *wh*-questions is a challenging task, since indirect answers to *wh*-questions are significantly less frequent than those of polar questions. It requires annotating a huge number of *wh*-questions within conversational context to collect a reasonable amount of *WhQ-IND* pairs for analysis and training machine learning algorithms; c). The implied meaning of indirect answers to *wh*-questions often depends heavily on the context of the conversation. It usually also involves nuanced linguistic features like sarcasm, irony, and figurative expressions which can be a challenge for humans (overhearers) to interpret, let alone for dialogue systems.

The aim of this paper is, therefore, to conduct a preliminary study by constructing what is, to our knowledge, the first corpus of indirect answers to *wh*-questions, and to investigate how direct answers are deduced from indirect answers.

This paper is structured as follows: Section 2 consists of a literature review, whereas Section 3 provides the requisite theoretical background. In Section 4, we present the data collection and annotation process. In Section 5 we propose possible information resources needed for interpreting indirect answers to *wh*-questions. Following this, in Section 6 we briefly describe a pilot study on generating indirect answers by using a pre-trained language model. Our results suggest this is a task that GPT finds difficult. The final section offers conclusions and some potential future work.

2 Related Work

Several studies exist concerning the interpretation and generation of indirect answers to polar questions: Green and Carberry (1994a,b, 1999) proposed both pragmatic and computational methods for understanding and generating indirect answers to polar questions. Specifically, they introduced a discourse-plan-based strategy for implicatures and a combined reasoning model to simulate a speaker's incentive for offering pertinent, unsolicited information. Furthermore, they designed a computational model that is capable of interpreting and generating indirect answers to polar questions in English. Their model relies on shared knowledge of discourse strategies and coherence relations to recognize and formulate a responder's discourse plan for a complete response.

Takayama et al. (2021) released the corpus *DIRECT*, which provides 71,498 indirect-direct pairs together with multi-turn dialogue history extracted from the MultiWoZ dataset, and conducted three experiments to examine the model's ability to recognize and generate indirect and direct utterances. The *DIRECT* corpus provides triples of paraphrases for each user's utterance: *original utterance, indirect utterance*, and *direct utterance*. This is the first study that offers a large-scale corpus of pragmatic annotations, which is very useful for understanding users' intentions in dialogue systems.

In another recent work, Louis et al. (2020) created and released the first large-scale English corpus of more than 34K *polar question–indirect answer* pairs, named *Circa*. That is a collection of natural responses obtained by crowd-sourcing and contains responses with yes-no meaning, as well as uncertain, middle-ground, and conditional responses. The authors also conducted experiments by fine-tuning a multiclass classifier over the BERT

model (Devlin et al., 2019), and then further fine-tuned those models with polar question-answer pairs from the *Circa* corpus. They examined the performance of different models for the classification of polar question-indirect answer pairs into the following meaning categories: 1. STRICT labels: *Yes; No; Probably yes / sometimes Yes; Yes, subject to some conditions; Probably no; In the middle; neither yes nor no; I am not sure; Other; N/A.*, and 2. RELAXED labels: *Yes; No; Yes, subject to some conditions; In the middle, neither yes nor no; Other; N/A.* [1] The study evaluated various baseline models and compared the performance of the models using only questions, only answers, and both questions and answers. The results indicated that joint models (that is, models trained both with questions and answers) outperformed answer-only models. The study also highlighted the challenges of classifying uncertain or ambiguous responses and suggested that incorporating the right information for the task remains a challenge.

Taking inspiration from the research of Louis et al. (2020), Damgaard et al. (2021) studied how to understand indirect answers to polar questions. Instead of crowdsourcing, they collected polar questions and indirect answers from the transcripts of the *Friends* TV series. After manual annotations, they released the FRIENDS-QIA dataset with 5,930 *polar question–indirect answer* pairs in English, both with the majority label and with the raw annotations. They further experimented with Convolutional Neural Networks (CNNs) with different word embeddings: CNN with GloVe embeddings and CNN with BERT embeddings. Furthermore, an additional crowd layer was added to enable the model to learn from the disagreement of human annotators. As a result, CNNs trained with BERT embeddings outperformed CNNs trained with GloVe word embeddings when the model was trained both with questions and answers. Furthermore, using Convolutional Neural Networks (CNNs) to evaluate the task, the authors showed that there was still room for improvement in the interpretation of indirect answers. However, they also found encouraging improvements when explicitly modeling human disagreement in the annotations.

---

[1] The RELAXED labels were achieved by collapsing the more uncertain and confusing classes from the STIRICT labels: "Probably yes / sometimes Yes" → "Yes", "Probable No" → "No", and "I am not sure" → "In the middle, neither yes nor no".

3 Background

The taxonomy of the response space to questions we use is formally characterized using the KoS framework (Ginzburg, 2012) which provides a theory of dialogue context and dialogue management. The *Question-Specific* responses are the most important subgroup of the taxonomy of the response space to questions. This includes responses providing *answers* (*Direct Answers* and *Indirect Answers*), and *Dependent Questions* where the response to the original question depends on the response to the question-response to that original question. Other subgroups of the taxonomy are the *Metacommunicative* responses (*Clarification Response* and *Acknowledgement*), and the *Evasion* responses (*Motivation*, *Ignore*, *Change the topic*, and *Difficult to Provide an Answer*). Detailed descriptions of each class are presented in Appendix B.

Direct Answers are defined as those that, given a proposition: p, a question: q, p is a direct answer to q, if and only if p is *about* q, and is entailed by either the meet of q's atomic or negative atomic answer set.[2] Indirect Answers are distinguished from direct answers under two basic conditions: a). the indirect answer p is not a direct answer to the question q, and b). the indirect answer p, together with a *bridging proposition bridgeprop* (some shared knowledge), entails r, which is a direct answer to the question q. The formal definition of indirect answers is stated as follows:

> Given $p : Prop, q : Question, dgb : DGBType$ InDirectAns(p,q,dgb) iff ¬DirectAns(p,q) and there exist $bridgeprop, r : Prop$ such that DirectAns(r,q) and In(dgb.FACTS, bridgeprop) and $\rightarrow (p \wedge bridgeprop, r)$. (Ginzburg et al., 2022)

As reflected in the definition, the implied direct answer from the indirect answer can be inferred with the help of shared knowledge during the conversation and some domain-independent information. However, in some cases, the interpretation of indirect answers might involve reasoning about the speaker's intentions. Thus, the process of inference will be influenced by the specific perspective, knowledge, goal, or interests of the individual making the inference.

In the following section, we present our methods and processes for collecting a corpus of indirect answers to *wh*-questions.

4 Corpus Collection

We aim to collect the first publicly available corpus of indirect answers to various content questions in English dialogue. To start with, we follow the annotation guidelines for the entire response space of the questions presented in previous works by Ginzburg et al. (2019, 2022), and also updated their annotation guidelines by adding extra instructions specific to indirect answers to *wh*-questions. We annotated various *wh*-questions and their corresponding responses from four different English corpora. Namely, BNC (Burnard, 2007), Cornell-Moive corpus (Danescu-Niculescu-Mizil and Lee, 2011), COCA (The Corpus of Contemporary American English, Davies, 2010), and LLC (The London–Lund corpus of spoken English, Svartvik, Jan, 1990).

4.1 Annotations

There are several steps involved in collecting the corpus of indirect answers to *wh*-questions:

- Step 1: we started by investigating the collections of question-answer pairs from the BNC with the response space annotations, shared by the authors of Ginzburg et al. (2022) on the OSF platform.[3] We re-annotated those collections following our updated guidelines and then extracted the *WhQ-IND* pairs.
- Step 2: we searched for various *wh*-questions (involving the *wh*- words *what, why, how, which, when, where and who*) and their responses using the SCoRE [4] search engine for the BNC. Table 1 presents the search patterns used for each *wh*-question, the number of examples obtained from them, and also the number of examples we annotated for this study. During this annotation process, we only focused on adjacent pairs of *wh*-questions and their responses, uttered by two distinct interlocutors. In addition, we also eliminated utterances in which the content is unclear (for instance, cases where the main parts of the utter-

---

[2]For the detailed description of the definition and formalization, see Ginzburg et al. (2022); for a detailed discussion of *Aboutness*, see (Ginzburg and Sag, 2000, pp. 129–149).

[3]https://osf.io/mq6r7/
[4]http://www.dcs.qmul.ac.uk/imc/ds/score/saved.html

Search Pattern	Search Result	Annotated
^when <V??>, ?	420	98
^where <V??>, ?	1877	94
^why <V??>, ?	1328	656
^how <V??>, ?	1640	359
^what <V??>, ?	7965	318
^who <V??>, ?	1696	366
^which <?N?> <V??>,	225	149
Total	**15151**	**2040**

Table 1: Search patterns from BNC, their results, and the number of annotated examples in this study.

ance are not available and marked with *<unclear>* tag, thereby reducing understanding of the utterance's meaning). As a result, we collected 35 *wh*-question and indirect answer pairs from 2040 examples of annotated *wh*-questions.

- Step 3: Ginzburg et al. (2022) reported that the CornellMovie corpus has the highest percentage of indirect answers in their data set. Therefore, we also annotated dialogues from the CornellMovie corpus and collected 12 pairs of *wh*-question and indirect answer pairs.

- Step 4: We searched for *wh*-questions and their responses in the conversational part of the London-Lund Corpus of Spoken English (LLC) corpus. This resulted in a total of 21 *wh*-question and indirect answer pairs.

- Step 5: we utilized the Corpus of Contemporary American English (COCA) [5], and searched for different types of *wh*-questions using various search patterns. The details of the search patterns are provided in Appendix C. Most of the examples taken from this corpus are from the sub-corpora: Movie, TV, and Spoken. An intern who is studying for a master's degree in English linguistics, specially trained in dialogue semantics, participated in this process. He went through at least 400 examples (around 1200 examples for some *wh*-question types) for each type of question and selected examples that are potential *WhQ-ID* pairs. These examples were then checked by the first author of this paper. In the end, we obtained 390 *wh*-question and indirect answer pairs from around 5000 *wh*-questions from the COCA corpus.

4.2 Corpus Description

The annotation and re-checking processes resulted in a collection of 458 *wh*-question and indirect

answer pairs. Among these, 390 examples were selected from the COCA corpus, 35 from BNC, 12 from CornellMovie, and 21 from the LLC corpus. The collected *WhQ-IND* pairs, their annotations, and the updated annotation guidelines are shared with the public on the OSF platform: https://osf.io/zuhvp/.

The number of indirect answers collected for various *wh*-questions also varies. As presented in Table 2, almost half (214 out of 458) of the collected examples are *how*-questions. Other frequent questions are *what*-questions and *why*-questions, 75 and 63 examples, respectively. In addition, we found 32, 31 and 29 examples, respectively, from *where*-questions, *when*-questions and *who*-questions. However, we only found 14 examples from *which*-questions.

wh-question	No. Indirect answers
What	75
Why	63
How	214
Which	14
When	31
Where	32
Who	29
Total	**458**

Table 2: Distribution of indirect answers across different *wh*-questions.

Inter Annotator Agreement To evaluate the reliability of the corpus annotation, we performed an experiment to determine whether the response in each dialogue instance within our corpus qualifies as an indirect answer.

In this annotation experiment, four annotators participated: the first author (referred to as First Annotator), an English L2 speaker enrolled in a Ph.D. program in linguistics and an expert in response space annotation tasks; an intern (referred to as Second Annotator), an English L2 speaker pursuing a master's degree in English linguistics; a volunteer native English speaker (referred to as Third Annotator) who is pursuing a master's degree in English linguistics, and another volunteer (referred to as Fourth Annotator), an English L2 speaker enrolled in a Ph.D. program in English linguistics. Before starting the annotation process, all annotators familiarized themselves with the updated annotation guidelines. Additionally, they underwent several training sessions and discussed any disagreements

together to ensure a shared understanding of the annotation criteria. In the end, they co-annotated 65 *WhQ-IND* pairs from the collected examples. Each of the four annotators, when marking an indirect answer, was also required to infer and supply the implied direct answer from the indirect answer.

We calculated the inter-annotator agreement score among four annotators using Fleiss's Kappa (Fleiss, 1971; Fleiss et al., 2003) and Krippendorff's Alpha (Krippendorff, 2011) methods in Python. As a result, the agreement scores among the four annotators are rather low: Fleiss's κ is -0.51, and Krippendorff's α is 0.025. This indicates substantial disagreement among the four annotators. In addition, we also calculated the inter-annotator agreement level between annotators with the average pairwise Cohen's Kappa scores (Carletta, 1996) using the *Scikit-learn* (Pedregosa et al., 2011) data mining and data analysis tool in Python with its *sklearn.metrics* package. The pairwise Cohen's κ obtained are presented in Table 3. These pairwise agreement scores (0.22 - 0.44) indicate that the agreement between the annotators ranges from fair to moderate agreement.

Annotators	Cohen's κ
First vs. Second	0.44
First vs. Third	0.28
First vs. Fourth	0.38
Second vs. Third	0.33
Second vs. Fourth	0.22
Third vs. Fourth	0.36

Table 3: The average pairwise Cohen's Kappa scores between annotators.

The low inter-annotator agreement scores can be attributed to the fact that annotating and interpreting indirect answers is a highly inference-based task with inherent subjectivity and pragmatic complexity. To further address this issue, 60 *wh*-question indirect answer pairs from the collected corpus were randomly selected and then annotated by both authors of the paper (both are experts in the response space classification task). In this way, our aim was to evaluate inter-annotator agreement among expert annotators. Cohen's Kappa score between the two experts is 0.60, which indicates a moderated to substantial agreement between the experts. This agreement score also corroborates the difficulty in annotating *WhQ-IND* pairs.

We hypothesize that the low levels of agreement

among annotators arise because identifying indirect answers to *wh*-questions involves a high level of pragmatic complexity. In addition to relying on the annotation guidelines, annotators need to use their semantic and pragmatic knowledge and experience, as well as their subjective judgments for identifying and inferring indirect answers. These low inter-annotator agreement results are also in line with the inter-annotator results reported in Ginzburg et al. (2022), who note a sharp decline when including annotations of indirect answers to calculate annotator agreements on different sets of response types. Yusupujiang et al. (2022) also reported that automatic classification results obtained for indirect answers are pretty low: F1-scores are 0.25 and 0.07 on their full taxonomy and coarser taxonomy respectively. Therefore, the authors suggest that a targeted set of features is necessary to automatically classify indirect answers.

5 Interpreting Indirect Answers to *Wh*-questions

Wh-questions are one of the most commonly observed question types in English conversation. Stivers (2010) reported that among the 328 questions that occurred in a videotaped American English conversation 27% ($n = 90$) of the questions were *wh*-questions. She indicated that the two commonest *wh*-questions types were *what*-questions (38%) and *how*-questions (23%). Other frequent types were *why*-questions (16%) and *when*-questions (12%). *Where*- and *who*-questions only accounted for 8% and 3% of their corpus, respectively. However, the distribution of *wh*-question types can vary depending on many other factors, such as conversational context, cultural and individual communication styles, as well as the specific nature of conversations.

Fox and Thompson (2010) presented the grammatical and interactional characteristics of different responses to *wh*-questions by studying a collection of 73 examples from American English conversations. The authors identified two broader types of responses to the *wh*-questions: *phrasal* and *clausal* responses. Their study suggested that phrasal responses provided simple answers to *wh*-questions, while clausal responses, specifically, clausal Phrase-in-Clause (*PiC*) responses, often signaled trouble with the question or sequences even though they also provided answers. Furthermore, the main types of clausal responses (that

is, full-clause responses) usually did not provide answers to the question, instead, they treated an assumption in the question as problematic or provided "no-access" responses, such as *I don't know*, or *he/she/they don't know*. It is worth mentioning that, the *"treating an assumption as problematic"* function of the full-clause responses corresponds to the "Clarification Response", precisely, the "Correction" response type, while the *"no-access" responses* correspond to the "Difficult to provide an answer" response type in the response space taxonomy provided by Ginzburg et al. (2019, 2022).

5.1 Information Sources

Ginzburg et al. (2022) proposed to categorize indirect answers into two main types: *shallow* and *deep* indirect answers. Shallow indirect answers are those where the implied direct answers are inferred only based on some shallow shared knowledge and domain-independent erotetic reasoning (also known as interrogative or questioning reasoning); whereas deep indirect answers require reasoning about the speaker's intentions, beliefs, and some domain-specific knowledge. Therefore, based on their suggestions, we further divide the information that one might need to interpret indirect answers into 9 categories as follows:

Basic linguistic knowledge: this is based on significant competence in the language used (grammar, vocabulary, etc.). As in Dialogue (1), the word (*daily*) used in the indirect answer helps questioner A to infer the implied direct answer from B's indirect answer, which is *"The last time it was inspected was yesterday/today."* Thus, A is required to have a good understanding of basic English grammar and vocabulary for interpretation.

(1) A: When was the last time that line was inspected, commander?
     B: It's inspected daily. [COCA Corpus]

Shared knowledge: this involves shared or communally established knowledge during conversations.

(2) *previous utterances*:I also had extraordinary hearing. During dinner, I could tune out the cacophony of chewing, slurping, chewing, cutlery scraping against plates, chewing, …
     A: Why aren't you eating, Sheldon?

     B: How can I with that horrible noise? [COCA Corpus]

From the previous utterances in Dialogue (2), one learns that Sheldon has very sensitive hearing. Therefore, the noise around Sheldon is the reason he is not eating. In contrast, in Dialogue (3), by providing the indirect answer *"Look what happened in 2018."*, Speaker B invites Speaker A to recall events that happened in 2018 to infer the direct answer to his question. Here, Speaker B believes that Speaker A shares the same communal memory as he does, and is capable of finding the requested information in this way.

(3) *previous utterances*:AXELROD: Yes. So, that lack of enthusiasm if it's Joe Biden, right, on the one side, Donald Trump on the other, I can tell you whose voters are going to be more enthusiastic.
     A: Well, how do you know that? How do you know that?
     B: Look what happened in 2018. [COCA Corpus]

Speaker's intentions/goals: the speaker conveys the messages indirectly by mentioning her/his goals or intentions. As shown in Dialogue (4), we can learn of Speaker B's intentions of *"[getting] married to that woman"*, so can infer the direct answer that the person that Speaker B is talking to is his girlfriend.

(4) A: Who are you talking to? Your girlfriend? I didn't know you had a girlfriend.
     B: I'm probably gonna marry this one. [COCA Corpus]

Speaker's belief/interest: some indirect answers convey speakers' beliefs or interest in a subject/topic, so correctly identifying these is the way to interpret the direct answer to the original *wh*-questions.

(5) A: Man, how do you know this shit's safe?
     B: These guys know what they're doing. Don't worry. They've tested it on dogs and everything. [COCA Corpus]

In Dialogue (5), Speaker B indicates her/his trust in the ability of those group of people who

invented the (*medical items or drugs*). Therefore, Speaker B's full trust in those people is the basis for her/him to (believe he) know(s) that the item invented by those people is safe.

Relationships between speakers: Indirect answers can be used between strangers to be polite and to exude more professionalism, or to avoid conflict in an employer-employee relationship. On the other hand, among close friends or family members, indirect answers might be used to make the conversation more casual based on their vast amount of shared knowledge. Thus, in Dialogue (6), Speaker B's response, *"Like you don't know."* indicates that Speaker A already knows the reason based on their relationship and shared history. However, a third party might not be able to infer Speaker B's implied direct answer because of not being in that relationship.

(6) *previous utterances*:Carl: Okay, here she is. She'll clear up this whole thing. What are you doing here?! Uh, Carl... What's goin' on? It's not what it looks like.

A: Why are you wearing that?

B: Like you don't know. [COCA Corpus]

Nuanced linguistic features: these include idioms, slang, figurative expressions etc. As in Dialogue (7), the figurative expression *"I'm right inside your head."* usually implies that she/he understands the other person's thoughts, feelings, and motivations.

(7) A: How do you know that?

B: I'm right inside your head. [COCA Corpus]

Common sense: this involves common knowledge about the world, certain social norms, customs, etc. In order to infer the implied direct answer *"I'm not very hungry now"* to the question about Speaker B's hunger level in Dialogue (8), one is required to understand what *"being flexible about eating time"* means.

(8) PREVIOUS UTTERANCES:Would you like to suggest a time for eating? Would I? Either of you

A: <laughs> how hungry are you Ken? <laughs>

B: I can I could eat now, or I could manage to wait. I'm quite flexible. [LLC Corpus]

Visual context can provide important cues for interpreting indirect answers, especially when analyzing multimodal dialogue settings. The Dialogue (9) is taken from the CornellMovie corpus, so is a dialogue in a movie scenario. Both speakers are in the same physical space and, hence, share visual context. Thus, Speaker A can identify the person requested by looking in the direction provided by Speaker B, *"At the end of the bar."*.

(9) A: Who said that?

B: At the end of the bar. [CornellMovie Corpus]

Non-verbal cues: we can utilize tone of voice, facial expressions, body language, etc. to better understand speakers' motivations and intentions. This is very useful when considering multimodal dialogues. For instance, in the constructed example of Dialogue (10), the parent can infer from the child's guilty facial expression and body behaviors that the child broke the window.

(10) *scenario*:A parent enters a room and notices a broken window. So the parent initiates the following dialogue:

A: Who broke the window?

B: (The child looks guilty and tries to avoid eye contact with the parent.) [Constructed example]

5.2 Statistical Analysis of Information Sources

To study which information sources are more frequently needed for the interpretation of indirect answers to *wh*-questions, we conducted a pilot study using the examples in our collected corpus of *WhQ-IND* pairs. The first author of this paper selected 141 examples (examples whose indirectness has been annotated with high confidence) for annotation with the 9 possible information sources presented above in Section 5.1 as a pilot study.

As indicated in Table 4, Basic linguistic knowledge (30.50%) and Common Sense (24.11%) are the two most frequent information sources used for inferring direct answers from indirect answers. The third frequently used information source is the Nuanced linguistic features in the indirect answers, which accounts for 14.18% of all information sources in our anno-

Information Source	How	Why	What	When	Where	Which	Who	Freq. %
Basic linguistic knowledge	11	7	13	10	0	0	2	30.50% (43)
Common sense	23	2	3	1	1	1	3	24.11% (34)
Nuanced linguistic features	14	2	2	0	0	1	1	14.18% (20)
Shared knowledge	7	4	3	0	0	0	0	9.93% (14)
Speaker's intentions/goals	5	2	1	0	0	1	4	9.22% (13)
Speaker's beliefs/interests	7	4	1	0	0	0	0	8.51% (12)
Relationships between speakers	0	1	2	0	0	0	0	2.13% (3)
Visual context	1	0	0	0	0	0	1	1.42% (2)
Non-verbal cues	0	0	0	0	0	0	0	0 %
Total	**68**	**22**	**25**	**11**	**1**	**3**	**11**	**141**

Table 4: Distribution of information sources.

tations. Furthermore, the Shared knowledge, Speaker's intentions/goals, and Speaker's beliefs/interests have similar distributions, which are 9.93%, 9.22%, and 8.51% respectively. Other types of information sources seem to have quite lower frequency: Relationships between speakers (2.13%), Visual context (1.42%), and Non-verbal cues (0%).

In addition, we can learn from Table 4 that, most of the indirect answers to *how*-questions can be interpreted based on Common sense and Nuanced linguistic features. For *what*- and *when*-questions, Basic linguistic knowledge seems to be used more in interpreting their indirect answers. However, due to the imbalanced number of examples for each type of *wh*-question in our current data set, our results concerning the distribution of information sources must be viewed as quite provisional.

6 Generation of INDs to *wh*-questions

As a pilot study, we fine-tuned the pre-trained response generation model Di-aloGPT (medium) (Zhang et al., 2020) with our collected corpus of indirect answers to *wh*-questions (458 examples), and tested the fine-tuned model's ability to generate indirect answers to *wh*-questions in a new test set.

Experimental Setup We fine-tuned our model by using Hugging Face's *"Transformer"* library. During the training, we randomly split the corpus into training and evaluation sets with a ratio of $4:1$. We set the number of training epochs to $num\_train\_epochs = 10$, with a per device training batch size of 4. The model also saves its result every $10,000$ steps, while also applying a weight decay of 0.01 to avoid overfitting. In addition, we adopted a step-wise evaluation strategy *evaluation_strategy="steps"*, to evaluate the model

every 500 steps during the training phase. Furthermore, we set *load_best_model_at_end=True*, to load the model that had the best performance during the evaluation steps. Finally, the input format of the data for fine-tuning is *"[PH] Previous dialogue history + [Q] Wh-Questions + [R] indirect answers + <|endoftext|>"*.

Evaluation We tested the performance of the fine-tuned model on 20 new *wh*-questions selected from the annotated 2040 examples of BNC *wh*-questions, where the original responses to these 20 examples were direct answers. We then deleted their original direct answers and created a test set with a format, *"[PH] Previous dialogue history + [Q] Wh-Questions + [R]"*. The fine-tuned model generated responses to those new *wh*-questions, and we evaluated the performance of the model by manually determining if the model-generated responses were indirect answers. However, only one example in 20 was an indirect answer. Details of the generated responses are presented in Appendix A for reference.

7 Conclusions and Future Work

In this paper, we have addressed the challenge of interpreting indirect answers to *wh*-questions. We started by collecting indirect answers to *wh*-questions from four different English corpora (BNC, CornellMovie, COCA, and LLC), and constructed a small corpus of 458 *WhQ-IND* pairs along with pre-question utterances and post-response utterances. Building such a corpus is highly labour intensive, given the difficulty of the task of classifying responses as indirect, as we demonstrated in several inter-annotator studies, including ones involving expert annotators.

In addition, we developed a scheme of 9 possible information sources used to infer direct answers from indirect answers and found—provisionally,

given problems with the imbalanced nature of our data set— that *Basic linguistic knowledge, Common sense*, and *Nuanced linguistic knowledge* are the three most frequently used information sources for the interpretation of indirect answers to *wh*-questions. Finally, we also conducted a preliminary experiment for generating indirect answers to *wh*-questions by fine-tuning a large-scale response generation language model, DialoGPT. The results of this latter experiment are hampered by the small amount of our current data set, but also suggest that this is a tricky task for GPT.

There are several clear limitations of the current study, which future work should improve on: (1). Since the size of the collected corpus is small, there is a need to continue collecting a more balanced and larger corpus of indirect answers to *wh*-questions; (2). The proposed 9 possible information sources need to be further evaluated, related to established components of context, and tested across annotators; (3). We hope to improve the performance of our generation model by fine-tuning it with a larger corpus. Other methods, such as few-shot learning, data augmentation, and transfer learning techniques may help improve the model performance on generating indirect answers to *wh*-questions.

Acknowledgements

References

Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Lou Burnard, editor. 2007. *Reference guide for the British National Corpus (XML Edition)*. Oxford University Computing Services on behalf of the BNC Consortium. Acess 20.03.2017.

Jean Carletta. 1996. Assessing agreement on classification task: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Cathrine Damgaard, Paulina Toborek, Trine Eriksen, and Barbara Plank. 2021. "I'll be there for you": The One with Understanding Indirect Answers. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 1–11, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

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 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Mark Davies. 2010. The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing*, 25(4):447–464.

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.

Marie-Catherine de Marneffe and Judith Tonhauser. 2016. Inferring Meaning from Indirect Answers to Polar Questions: The Contribution of the Rise-Fall-Rise Contour. In *Questions in discourse*, pages 132–163. Brill, Leiden, The Netherlands.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2003. *Statistical Methods for Rates and Proportions*. John Wiley & Sons.

Barbara A Fox and Sandra A Thompson. 2010. Responses to Wh-Questions in English Conversation. *Research on Language and Social Interaction*, 43(2):133–156.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford.

Jonathan Ginzburg and Ivan A. Sag. 2000. *Interrogative Investigations: the form, meaning and use of English Interrogatives*. Number 123 in CSLI Lecture Notes. CSLI Publications, Stanford: California.

Jonathan Ginzburg, Zulipiye Yusupujiang, Chuyuan Li, Kexin Ren, Aleksandra Kucharska, and Pawel Lupkowski. 2022. Characterizing the response space of questions: data and theory. *Dialogue & Discourse*, 13(2):79–132.

Jonathan Ginzburg, Zulipiye Yusupujiang, Chuyuan Li, Kexin Ren, and Paweł Łupkowski. 2019. Characterizing the Response Space of Questions: a Corpus Study for English and Polish. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 320–330, Stockholm, Sweden. Association for Computational Linguistics.

Nancy Green and Sandra Carberry. 1994a. Generating indirect answers to yes-no questions. In *Proceedings of the Seventh International Workshop on Natural Language Generation*.

Nancy Green and Sandra Carberry. 1994b. A Hybrid Reasoning Model for Indirect Answers. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 58–65, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

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

Klaus Krippendorff. 2011. Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5(2):93–112.

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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

John R Searle. 1975. Indirect speech acts. In *Speech acts*, pages 59–82. Brill.

Tanya Stivers. 2010. An overview of the question–response system in American English conversation. *Journal of Pragmatics*, 42(10):2772–2781.

Svartvik, Jan, editor. 1990. *The London–Lund corpus of spoken English : Description and research*, volume 82 of *Lund Studies in English*. Lund University Press. Book Editor.

Junya Takayama, Tomoyuki Kajiwara, and Yuki Arase. 2021. DIRECT: Direct and Indirect Responses in Conversational Text Corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1980–1989, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zulipiye Yusupujiang, Alafate Abulimiti, and Jonathan Ginzburg. 2022. Classifying the response space of questions: A machine learning approach. In *SemDial 2022-26th Workshop on the Semantics and Pragmatics of Dialogue*, pages 59–69.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Appendix A: Details of the responses generated as indirect answers

Generated Responses:

(1) *[PH] And she's a bit of a slag. She don't really, you know, I mean she's got a bit of a reputation. I suppose everyone has but I hear about her a lot. In school and everything. Which one was that? Pardon?*
Question: Which one was this?
Response: Pardon. I think I know who you are.

(2) *[PH] Is it a big fire? Is it? You've seen it. I can't remember.*
Question: Which room is it in?
Response: I don't know. I left the building when I was nine.

(3) *[PH] Let me put that question to the hundred women here, are you looking forward to Christmas? Button one for yes, and button two for no. And, this hundred don't share my apparent cynicism, seventy three of them are, twenty seven said no. Of those who said yes, why?*
Question: What are you looking forward to?
Response: I'm looking forward to a good time.

(4) *[PH] Any problems with these? Erm no. They they they were they were straightforward, just. Okay.*
Question: How did you do this one?
Response: I just, uh... I used a program called CamScanner.

(5) *[PH] We want a difference of nineteen somewhere don't we? That's easy. We drop the calculator everywhere. Well you can drop the calculator if you want it would be better if you pick 'em up and stick the numbers in.*
Question: How're you gonna do?
Response: Well, I'm going to use my powers for good. (Indirect Answer)

(6) *[PH] So nine hundred, it works out at nine pound odd What would two thirds of three quarters come to roughly. Just sort of thinking about it having a guess. A half.*
Question: How did you get that?
Response: <pause> I think I know. I think I know. I think I know.

(7) *[PH] What I, where I gonna go? You're not going anywhere, you're staying here Can I help you plant the? Bulbs, er probably not cos I think I'll have to do while you're at school cos daddy tells me it's gonna rain this afternoon, I'll have to do that this morning, pity that.*
Question: Why can't we do them tomorrow?
Response: I don't know, I think it's best if we wait till tomorrow.

(8) *[PH] he save, he keeps swords. He saves those. Well, oh do they save them? Mm.*
Question: Why do they save them?
Response: Well, they're not exactly cheap.

(9) *[PH] Do you want mum to come to Argos with me tomorrow morning? Why? Well she isn't, I don't think she's doing anything tomorrow morning.*
Question: Are you asking me? Why are you asking me?
Response: Why are you asking me? Why are you asking me?

(10) *[PH] Yes, Christmas eve, well. Did you have many guests at sunrise? Well we, we had about sixty erm, people that trudged up to the top of the hill including erm, one who didn't know there was a wedding, he was a runner, and then he and he had tagged the top of Arthur's Seat and was astonished to see a crowd there.*
Question: What do you wear to a mountain top wedding in the middle of winter?
Response: Well, I'm a guy so I don't really have to worry about that sort of thing.

(11) *[PH] Erm forty nine. Alright and five squared? Twenty five. Right.*
Question: What's forty nine and twenty five come to?
Response: Erm five centimeters per second.

(12) *[PH] yes It means you can pack in a lot more things in your day, you can work, you can look after your children, you can take them out, you can do a lot of things that you couldn't do otherwise. mm, mm, yeah Er from the necessity in an area with no public transport*
Question: where's that?
Response: Er, I don't know, I moved to London. And I'm not even a very good driver.

(13) *[PH] Yeah? Negative one. Now the tricky one, let's do plus four make it plus five, plus five a a takeaway a negative three. Okay?*
Question: Where do we start?
Response: Well, I'm sure we can come up with a few ideas.

(14) *[PH] Right. Twelve over three and Okay. Now the notation, do you understand the notation?*
Question: Where's the notation?
Response: I don't know, I'm just?

(15) *[PH] I would only been six or seven when the war started. It started in nineteen fourteen. First I mind about it was the first zeppelin that came over. Great big thing like a huge balloon and lived out at Ness and I mind me mother coming and getting us early and this great big thing come over.*
Question: Who was in it?
Response: Zebediah Kerman, a Russian spy that was captured during the Russian Revolution.

(16) *[PH] Yes I'm getting you out treasure. This this er thing's got stuck on the rails. he's got Yes but one of them's going to get in.*
Question: Who's going to get in first and who's going to get in later?
Response: I'm gon na get in there, but I can't get in the car.

(17) *[PH] What are you going to do? I'm, making another lot of marmalade. Again?*
Question: Who's going to eat it all?
Response: I'll make it up to you.

(18) *[PH] No, but I just want you to tell me. Science. With whom? I don't know.*
Question: When are you supposed to come up here again?
Response: Well, I'm supposed to be in Las Vegas by the time I'm eighteen.

(19) *[PH] Oh yeah. do the one that you can do. Okay, well. What what we need to do, is make sure these scripts are done before half term really, don't we.*
Question: When's half term?
Response: It starts in a few minutes.

(20) *[PH] Have you had any erm problems in the past? No. No problems at all? No.*
Question: When was the last time you saw your doctor?
Response: I haven't.

Appendix B: Taxonomy for Responses to Queries

Full-Taxonomy	Description
Direct Answer (**DA**)	the response directly offers an answer to the question.
Indirect Answer (**IND**)	the answer to the question can be indirectly inferred from this utterance.
Dependent questions (**DP**)	the answer to the original question depends on the answer to this query response.
Clarification Response (**CR**)	the speaker asks for extra information to confirm (s)he understood the question correctly, requires additional information to understand it better, or provides some information to clarify/correct misinformation from the previous utterance.
Acknowledgement (**ACK**)	the speaker acknowledges that (s)he heard the question, such as mhm, aha,... etc.
Motivation (**MOTIV**)	a query response about the motivation of asking the initial question.
Ignore (**IGNORE**)	the utterance does not relate to the question, but to the situation.
Change the topic (**CHT**)	the utterance signals that the speaker does not want to answer the question, instead (s)he changes the topic, and gives an evasive response.
Difficult to provide an answer (**DPR**)	the speaker indicates that (s)he does not know the answer, or it is difficult for her/him to provide an answer, so points at a different information source,
OTHER	utterance that does not fit in any of the categories above.

Appendix C: Details of search patterns and annotated questions from the COCA corpus

Search Pattern	Annotated Questions	Number of INDs
what * * * PUNC	What do you think?	28
what are * * *	What are you * ?	14
	What are you going to * ?	17
	Why are you doing this?	27
	Why are you still here?	3
	Why are you following me?	6
	Why are you calling me?	1
	Why are you so nervous?	2
	Why are you so happy?	1
	Why are you wearing that?	3
why are * * * PUNC	Why are you protecting him?	1
	Why aren't you eating?	1
	Why are you so calm?	1
	Why are you ignoring me?	1
	Why are you helping us?	1
	Why are you here?	1
	Why are you so late?	1
	Why are you so surprised?	1
	How do you know that?	35
	How do you do that?	7
	How do you explain that?	25
how do * * * PUNC	How do you know this?	40
	How do you figure that?	13
	How do we do that?	26
	How do you feel?	51
	Which one do you want?	3
which one * * * PUNC	Which one do you like?	5
	Which one do you think?	2
who was * * * PUNC	Who was on the phone?	1
who is * * * PUNC	Who is responsible for this?	1
	Who are all these people?	2
	Who are you working for?	2
who are * * * PUNC	Who are you looking for?	1
	Who are you voting for?	1
	Who are you talking to?	14
	Who are you talking about?	1
when was * * * PUNC	When was the last time?	23
when is * * * PUNC	When is he coming back?	2
where did * * * PUNC	Where did you get that?	21
where * * * PUNC	Where is he now?	4
Total		**390**