*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 220–228
Marseille, 11–16 May 2020


# Developing a Corpus of Indirect Speech Act Schemas

**Antonio Roque, Alexander Tsuetaki, Vasanth Sarathy, Matthias Scheutz**
Tufts University
Medford, MA, USA
{Firstname.Lastname}@tufts.edu

**Abstract**

Resolving Indirect Speech Acts (ISAs), in which the intended meaning of an utterance is not identical to its literal meaning, is essential to enabling the participation of intelligent systems in peoples' everyday lives. Especially challenging are those cases in which the interpretation of such ISAs depends on context. To test a system's ability to perform ISA resolution we need a corpus, but developing such a corpus is difficult, especialy given the contex-dependent requirement. This paper addresses the difficult problems of constructing a corpus of ISAs, taking inspiration from relevant work in using corpora for reasoning tasks. We present a formal representation of ISA Schemas required for such testing, including a measure of the difficulty of a particular schema. We develop an approach to authoring these schemas using corpus analysis and crowdsourcing, to maximize realism and minimize the amount of expert authoring needed. Finally, we describe several characteristics of collected data, and potential future work.

**Keywords:** corpus methodologies, corpus construction, crowdsourcing

## 1. Introduction

**Indirect speech acts** (ISAs) involve utterances whose literal meanings are not identical to their intended meanings (Searle, 1975). For example, the utterance "can you open the door?" has the surface form of a question and a **literal meaning** of an elicitation of information from the hearer regarding the hearer's ability to open the door. In some cases, the speaker's intent may in fact be to have the hearer reply with the answer to that question. However, if the **intended meaning** is a request that the hearer perform the action of opening the door, then an ISA has been performed.

Over the years, researchers have developed inferential (Perrault and Allen, 1980), idiomatic (Wilske and Kruijff, 2006), and hybrid (Briggs and Scheutz, 2013) implementations of ISA analysis. Studies have shown that people commonly use ISAs in everyday conversations as well as in task-based dialogues with robots (Williams et al., 2018), and that context is important for understanding ISAs (Gibbs Jr, 1979). In fact, we argue that the detection and interpretation of ISAs is modulated by numerous contextual factors, such as location, task, interactant roles, co-present objects, and recent dialogue history. Even slight changes in this context can change whether an utterance's intended meaning is different from its literal meaning. Intelligent systems that interact with people in real-world environments need to be able to use context to determine whether an utterance should be interpreted literally, or as part of an ISA. An important part of developing such a capability is testing it. But doing so requires answering several difficult questions regarding how to collect and represent such content.

In this paper we make the following contributions. We describe an approach to testing ISAs that is derived from relevant work in using collections of test problems to track progress in systems that perform reasoning tasks. We present a formal representation of ISA Schemas required for such testing, including a measure of the difficulty of a particular schema[1]. We develop an approach to authoring these schemas using a combination of corpus analysis and crowdsourcing, to maximize realism and minimize the amount of expert authoring needed. Finally, we describe several characteristics of collected data, and potential future work.

## 2. Related Work

AI researchers have developed a variety of language-processing tests that require reasoning about knowledge (Storks et al., 2019). Examples include COPA (Roemmele et al., 2011) and RTE (Rus et al., 2008). Trichelair et al (2019) describe some problems associated with the datasets used for such tests, including limited size and the predictable structure of their examples. The terms **corpus** and **corpora** are used for these datasets by e.g. Levesque et al. (2012) and Morgenstern et al. (2016), even though this usage differs somewhat from the traditional usage in linguistics, i.e. referring to a collection of texts and conversations.

One approach we find particularly inspiring is that of *Winograd Schemas* (WS), which are used to test a system's ability to perform anaphora resolution. The following is an example, reformatted from Levesque et al. (2012), of a WS problem:

> Statement: The trophy doesn't fit in the suitcase because it's too [big / small].
>
> Question: what is too [big / small]?
>
> Answer: [the trophy / the suitcase].

This example shows how schema are actually made up of two halves, called *options*. In Option 1, "big" is selected for the statement and question and the correct answer is "the trophy;" in Option 2, "small" is selected for the statement

[1]The plural of *schema* is *schemas* (OED Online, 2019); we adapt this term from *Winograd Schemas* as described in Section 2.

and question and the correct answer is "the suitcase." When a schema is presented, typically only one option is shown, and the system or person being tested needs to select the right answer.

We extract several important lessons from the experience of Winograd Schemas. First, regarding **levels of difficulty**: WS researchers distinguish between *easy* and *hard* WS problems. Easy WS problems are those that can be solved by (1) *statistical correlations* e.g. simply observing whether the query words co-occur more frequently with one of the possible solutions, (2) *selectional restrictions* in which the answer can be determined just by using definitions of the options, and (3) other simple *syntactic cues* (Bender, 2015). Hard WS problems require reasoning about knowledge. In the example above, determining what the pronoun "it" refers to requires reasoning that in general if object A does not fit in object B, then object A is bigger than object B. The words "big" and "small" are equally applicable to "trophy" and "suitcase," so a system that only uses statistical correlations will do no better than chance.

This leads to the second important lesson we extract from the experience of Winograd Schemas, regarding **the importance of alternate options.** Having each schema be made up of two options enables the hard problems that cannot be solved by statistical correlations, because the statements in each option only vary slightly. (WS are often authored such that they only vary by a single word, though in some cases this may be several words.) This also ensures that the corpus of schemas is testing for those slight variations that create large changes in meaning.

This contributes to the third important lesson we extract from the experience of Winograd Schemas, regarding **the difficulty of developing a corpus of such schemas.**

The ideal solution would be to find several such paired halves in naturally-occurring data, but to the best of the community's knowledge there is no such source of naturally-occurring data. One obvious solution is to have experts construct them, but this is time-consuming and potentially leads to unrealistic data. These challenges are spelled out in more detail in Section 4. Our general approach to solving these problems is described in Section 5. An example corpus development is given in Section 6. But first, the next section provides our formal definition of ISA Schemas.

## 3. Defining ISA Schemas

Imagine a person injures their leg and goes to a doctor. In the doctor's office, the doctor says that they will begin by asking about the extent of the injury. The doctor says: "Can you run?" In this context, the utterance is clearly a question asking about the patient's capability.

After a few more questions, the doctor determines that the patient can in fact run, and that it is safe to test the extent of the patient's injury. The patient is taken to a treadmill in the exercise room, and the doctor says: "Can you run?" In this context, the utterance is clearly a request that the patient begin running.

So the utterance is identical in both cases; only the context has changed. This contextual change can be informally represented as follows.

| | |
|---|---|
| Utterance: | *Can you run?* |
| Context 1: | *a doctor talking to a patient,* |
| | *in a doctor's office,* |
| | *to collect information for diagnosis.* |
| Context 2: | *a doctor talking to a patient,* |
| | *in the exercise room,* |
| | *to test physical capability.* |

There are several cues for interpreting the utterance. First, *role* of the speaker and of the hearer. Second, *location*: in an office it is unusual to suddenly begin running, whereas on a treadmill it is perfectly natural. Third, *task*: in Context 1 the task is to collect information about the injury, whereas in Context 2 the task is to test physical capability; it is therefore more plausible that the utterance in Context 1 is asking about an ability and in Context 2 is requesting an action. Fourth, *copresent items*, such as a treadmill. Finally, *interaction history*: in Context 2 the patient has already answered the question and the doctor has already diagnosed that it was safe for the patient to run. All of this suggests that the utterance in Context 1 is asking about an ability, and the utterance in Context 2 is requesting an action. The literal meaning of the utterance is asking about an ability, so in Context 2 the utterance is an ISA.

This example shows how, for a given utterance, variations in context produce variations in interpretation. We therefore define an **ISA Schema** as

$$S_{ISA} = (u, l, c_{1...n}, i_{1...n})$$

where $u$ is an utterance, $l$ is the literal meaning of $u$, $c_{1...n}$ is a set of contexts in which $u$ is made (where a context is defined by a set of feature pairs), and $i_{1...n}$ is the intended meaning of $u$ for contexts 1 to n. So for a given context $c_x$, $u$ is an ISA iff $l \neq i_x$.

Figure 1 shows an example ISA Schema. This schema includes context features including the task at hand and the feature role, with a different intended meaning $i$ for each context. The utterance is not an ISA in context 1 (because its literal meaning is the same as its intended meaning) and the utterance is an ISA in context 2.

We are interested the difficult cases in which context affects intended meaning. So we additionally require that (1) at least one of the contexts must be an ISA (i.e. $l \neq i$ for that context) and at least one of the contexts must *not* be an ISA (i.e. $l = i$ for that context), (2) a system to interpret ISA should look at each context of the ISA separately, e.g. first ($u$, $l$, $c_1$) and then ($u$, $l$, $c_2$). This follows from the experience of WS Schemas as described in Section 2.

## 4. Techniques for Developing a Corpus of ISA Schemas

Having defined $S_{ISA}$, we next need to define how these $S_{ISA}$ can be obtained.

From the relevant work in developing corpora for reasoning tasks, we can identify several approaches. This section will show how any technique for developing a corpus of ISA Schemas will involve actions on a continuum involving trade-offs between naturalness, accuracy, and scalability. After this section describes each of these techniques

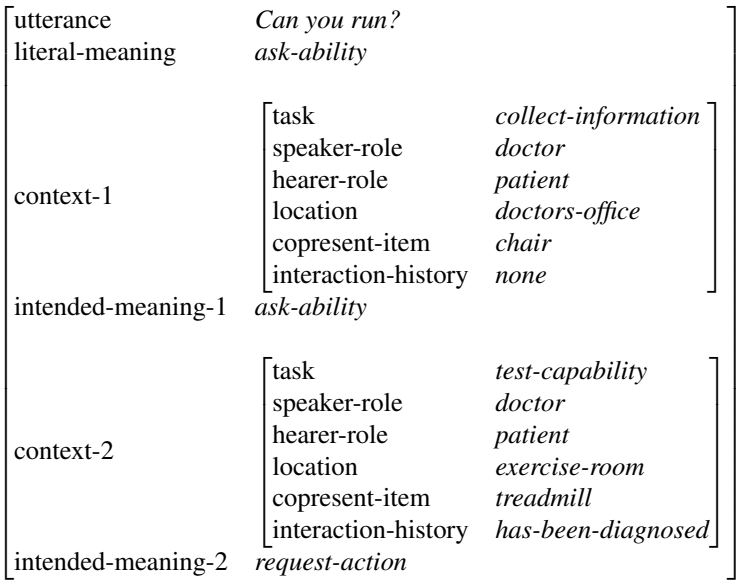


Figure 1: Example ISA Schema, where the intended meaning of an utterance varies based on context, and an Indirect Speech Act is performed in context 2 when the intended meaning is not equal to the utterance's literal meaning.

and the trade-offs they involve, the next section describes how we developed an approach that appropriately balances these trade-offs.

### 4.1. Expert Authoring

First, we ourselves could author $S_{ISA}s$; this is the approach that was initially taken by WS researchers.

The challenge is that this approach is time-consuming, although regarding WS, Levesque et al (2012) argue that the amount of effort required to author problems, while not negligible, is not insurmountable either. Indeed, Levesque et al. manually authored 273 schemas, although their approach does not scale to create thousands of schemas. However, once these expert-authored schemas are available, they have been shown to be translatable into other languages (Amsili and Seminck, 2017).

Another challenge is that this approach produces schemas that are potentially not representative of material that would be found "in the wild." However, Levesque et al (2012) argue that WS problems can be derived from natural data, and that in practice, WS problems can be combined with more-easily-obtained *Pronoun Disambiguation* problems, which are like one half of a WS, thereby being "easy" problems by definition (Levesque et al., 2012).

One of the disadvantages of having experts perform the authoring is that they may subconsciously tailor the resources authored based on the needs of their research. One of the advantages is that the resources are more likely to be high-quality, as the experts will have a strong idea of what the schema represents, and have spent time considering what they should look like. No matter what type of authoring approach is used, the corpus benefits from having the experts examine at least a subset of the corpus to ensure the appropriate level of quality is being maintained.

### 4.2. Non-Expert Authoring

A second approach we could use is to have non-experts author $S_{ISA}$.

This has been done with WS in several different ways. Rahman and Ng (2012) had 30 undergraduate students create WS, resulting in 941 schemas. A more scalable approach was taken by Sakaguchi et al (2020), who used the Amazon Mechanical Turk crowdsourcing platform to author a set of over 40,000 schemas.

The advantage of having non-experts author schemas is that it removes the authoring burden from researchers, thereby providing a more scalable approach that can generate larger corpora. It also lessens any potential subconscious bias. The disadvantage is that the non-expert authors are still performing a somewhat abstract language task, and therefore may produce results that are not necessarily tied closely to the reality of their daily language use. Another disadvantage is that non-experts are more likely to produce noisy output.

### 4.3. Non-Expert Validation

Regardless of the approach used to develop a schema, the corpus benefits from validation by examining the corpus. This can be done by experts as described above, but using experts does not scale to larger corpora.

In developing their WS, Bender et al (2015) conducted a study of how correctly humans intepreted a corpus of WS, and found that participants received an average score of 92%. That study also helped to identify ill-formed problems that had been human-authored, but which were not evident at the authoring stage because the author had access to both options of the WS (whereas the validator only had access to one of the options.)

In the same way, Zellers et al (2018) used Amazon Mechanical Turk to have human non-experts validate a multiple-choice corpus of grounded commonsense inference state-

ments.
The advantage of this approach is that it provides a "reality check" on the quality of the data. The disadvantage is that the non-expert validators may incorrectly process the corpus elements, either allowing substandard schemas or removing acceptable schemas.

### 4.4. Defining a Task to Collect Data

One intriguing possibility is to have subjects work on a task that "naturally" produces $S_{ISA}s$. This would require defining a task that, in the course of execution, would require someone to use a given utterance (1) in a context in which the utterance's literal meaning was intended and then (2) in a context in which the utterance's literal meaning was **not** intended.
Collecting $S_{ISA}s$ in this manner would have the advantage of being more natural than corpus-based non-expert collections. However, the disadvantage is that defining such a task is prohibitively difficult without a well-defined notion of what $S_{ISA}$ are, and how to characterize their quality. We therefore reserve this approach for future work.

### 4.5. Extracting from Existing Corpora

This approach is a variation of the previous technique: rather than create a task to collect data, we could extract schemas from an existing corpus. However, as described in the previous technique, there is no single corpus that naturally contains the schemas that we need. We therefore looked through sets of corpora to see if an utterance in one corpus might have an identical or near-identical utterance in another corpus, in different contexts, that we could combine together to make an $S_{ISA}$.
Indeed, this enabled us to perform expert authoring of several schemas that were thus tied to existing corpora. However, because this approach used expert authoring, it did not scale. We considered using automated approaches to extract the $S_{ISA}s$, but we identified several challenges: (1) automatically identifying when utterances were "similar enough" (i.e. whether the utterances had to be exactly similar word-for-word), (2) automatically extracting the context information in a way that was comparable across corpora, and (3) automatically determining the intended meaning of the utterances. Indeed, accomplishing (3) was one of the motivations for developing an ISA corpus to begin with.
Using existing corpora promises to result in highly-realistic schemas. Therefore, although it is currently infeasible to extract schemas entirely from corpora, the next section shows how we use data (such as utterances and contexts) from existing corpora whenever possible.

## 5. A General Approach to Developing a Corpus of ISA Schemas

As described above, any step in developing a corpus of ISA Schemas will involve actions on a continuum between explicit authoring and natural occurrence, with explicit authoring providing scalability problems. This section therefore defines a general approach which maximizes both scalability and ties to realistic data. The following section describes an execution of this general approach.
Given: the definition of $S_{ISA}s$ presented in Section 3.
**Step 1**: we identify prospective *utterances* for the $S_{ISA}s$. Rather than author these ourselves, we automatically extract them from a corpus. We review the utterances only long enough to remove any that, for example, have unusual characters. In other words, to help ensure scalability, we do not individually examine them.
**Step 2**: for each context feature, we author several *suggested context values* by extracting them from corpora. This involves expert authoring, but in terms of scalability the number of these suggested context values is constant per corpus (i.e. a corpus with 100 schemas may have the same number of suggested context values as a corpus with 10,000 schemas.)
**Step 3**: we use non-experts to *author* values for the context features, guided by the the suggested contexts, enabled by crowdsourcing as described in Section 4.2.
**Step 4**: we use non-experts to *validate* the schemas that have been authored, enabled by crowdsourcing as described in Section 4.3.
**Step 5**: we use a limited amount of expert authoring to produce the best-validated schemas. To maintain scalability, we minimize the amount of expert authoring, such as only requiring experts to act as "tie-breakers" as described in the example below.

## 6. Example Development of a Corpus of ISA Schemas

We now describe a corpus development whose goals were the following. We wanted to ensure that we had defined an approach to collecting a corpus of $S_{ISA}s$ in a way that minimized authoring effort while maximizing realism and maintaining scalability. For our first effort, we also wanted to produce a corpus that was small enough to be closely examined, while ensuring that we used scalable approaches. From our familiarity with several corpora (Li et al., 2017; Eric and Manning, 2017; Sun et al., 2019) we decided to focus on utterances of the form "Can you...?"
For **Step 1**, we used a script to extract all utterances from the DailyDialog corpus (Li et al., 2017) that took the form "Can you...?", and we organized these utterances in a hash table with a key of the first three words to nominally cluster like sentences. We then randomly extracted 250 sentences and manually discarded all sentences that on a quick reading seemed to have unclear grammar or were not very understandable. This left a set of 215 utterances.
For **Step 2**, we authored suggested context features as follows. First, we took a sample of 20 utterances from the 215 utterances identified in Step 1. Next, we identified 2 possible contexts for each of these 20 utterances, and collected all unique context features. The suggested context features we identified are shown in Figure 2.
For **Step 3**, we constructed a web-based GUI, shown in Figure 3, that used the elements of Steps 1 and 2. One of the utterances identified in Step 1 is shown at the top of the GUI. The context features from Step 2 are provided as suggestions in a drop-down menu, which when selected, populates the editable text input next to each context feature name. As shown in Figure 3, the authors are encouraged to modify the editable text or to enter text that was not suggested. They also may enter "irrelevant" for any given fea-

| Context Feature | Suggested Context Values |
|---|---|
| Location | School, Store, Street, Restaurant, Offices, Car on the road, Pool, University, Hotel, Home, Phone call, Nondescript |
| Task | getting contact information, looking for a hotel, asking about a book, mailing a letter, talking about a show, general dialogue, planing a meeting, planning errands, looking for a rest stop, grocery shopping |
| Speaker Role | Volunteer, Administrator, Tourist, Tour Guide, Local, Customer, Employee, Driver, Assistant, Waiter, Teacher, Student, Interviewer, Interviewee, Translator, Public Servant |
| Hearer Role | Volunteer, Administrator, Tourist, Tour Guide, Local, Customer, Employee, Driver, Assistant, Waiter, Teacher, Student, Interviewer, Interviewee, Translator, Public Servant |
| Copresent Item | car, buildings, hotel, theater, letter, package, phone, computer, toy, stamps, stone, food, utensil, notepad, pictures, clothes, fabric, money |
| Interaction History | this is the beginning of the interaction, they have talked about a plan to accomplish the task, the person is beginning an order, they have already talked about the copresent item, they have talked about events related to the task, they have determined that the person wants a specific item, the person has just arrived at the location, they are having difficulty communicating |

Figure 2: Suggested Context Features and Values, manually extracted from corpora by the authors

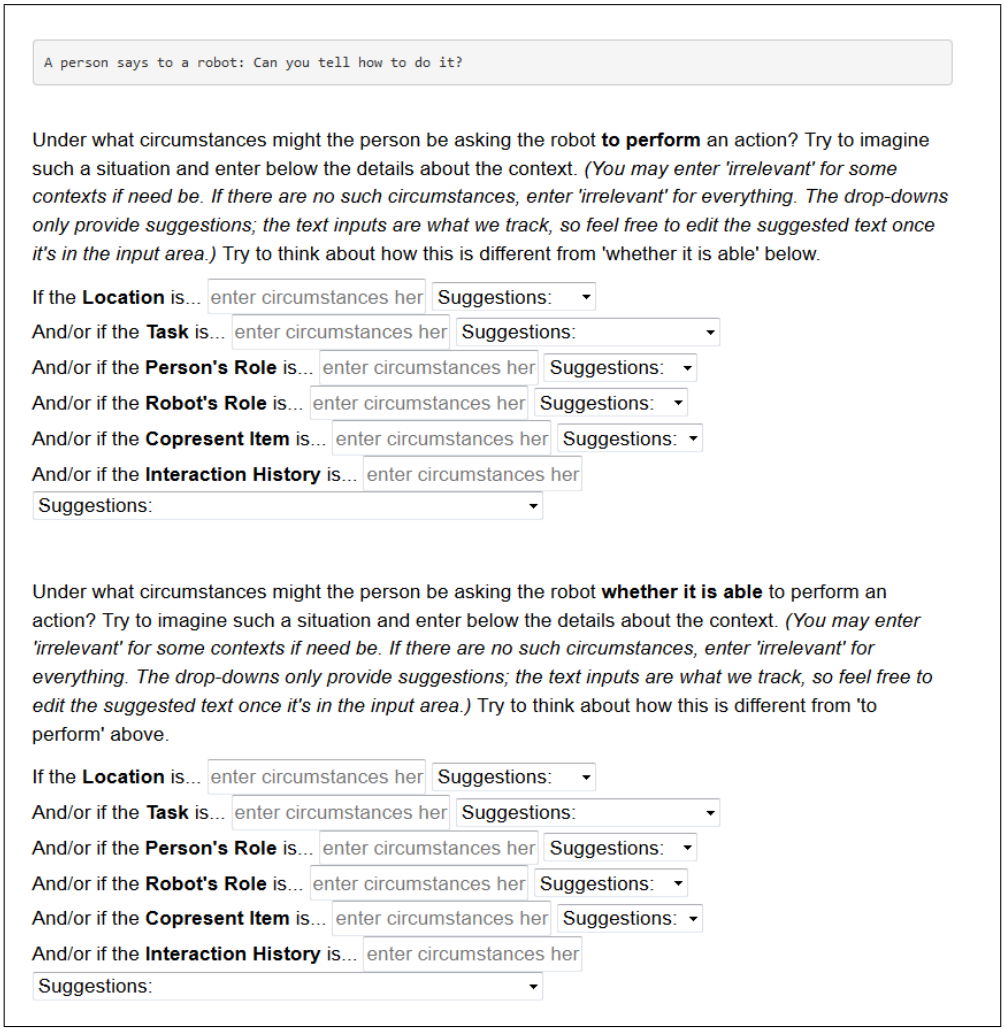


Figure 3: Example of the web-based GUI used for non-expert authoring of ISA Schemas.

ture, or to enter "irrelevant" for all features if they cannot imagine an appropriate scenario.

We used this GUI to collect data using the Amazon Mechanical Turk and psiTurk (McDonnell et al., 2012) crowdsourcing tools, adhering to the oversight of an Institutional Review Board. Before being shown the GUI in Figure 3, the participants were given the following overview:

> On each of the following pages, you will be shown a description of a human-robot interaction. On the same page you will see several questions for you to answer related to that description.
>
> We are interested in how context can change interpretation. Imagine a doctor says to you: "Can you move your arm?" In some circumstances this might be a request for you to perform an action, such as moving your arm out of the way. In other circumstances it might be a question about whether you are able to perform an action, such as after a serious injury. That is why we are asking questions about circumstances.

We used the word "circumstances" to help make the idea of "context" intuitive to the participants. We framed the "Can you..." utterance in the context of a robotic interaction in part because that is the scenario that we are interested in, and in part to provide a concrete basis for the annotators to consider.

The GUI shown in Figure 3 was used to generate 92 $S_{ISA}s$ (80 schemas authored by 20 participants authoring 4 schemas each, and 12 additional schemas being authored by participants who did not complete the process of authoring a set of 4.)

As a first characterization of the collected data, we began by considering the difficulty of the schemas collected. As described earlier, WS are characterized as either "easy" or "hard" depending on how amenable they are to statistical analysis, which is affected by the limited number of words that are changed between the two options. We therefore define the **contextual difficulty** of an $S_{ISA}$ as the number of nonzero context features that are non-identical between its two contexts. So if two contexts, with different intended meanings, are identical except for 1 feature, then it has a contextual difficulty of 1, which is the maximum difficulty. Contextual difficulty is defined for a nonzero number of contexts because when two contexts sets are identical but have different intended meanings, this indicates that the meaning is ambiguous. For our current purposes, we are taking note of when such ambiguities occur but we leave a full exploration of ambiguities in $S_{ISA}$ for future work. Additionally, the existence of schemas whose two halves are identical could also be the result of authors who did not understand the task.

The number of examples of schemas for each contextual difficulty level, for the first set of collected data, is shown in Table 1. In addition, 14 0-level (i.e. ambiguous) schemas were identified. Note that the measure of contextual difficulty is affected by the existence of the "irrelevant" keyword which the schema creators were instructed to use; when measuring contextual difficulty "irrelevant" is considered a variable which matches any other word.

| Contextual Difficulty | Examples Found |
|---|---|
| 1 | 14 |
| 2 | 14 |
| 3 | 10 |
| 4 | 13 |
| 5 | 12 |
| 6 | 15 |

Table 1: Difficulty Level of Schemas, based on differences in context features, in a collection of 78 $S_{ISA}s$, where 1 is most difficult and 6 is least difficult.

For **Step 4**, we developed and used a GUI as shown in Figure 4 to perform non-expert validation through crowdsourcing. The purpose of this was to determine the extent to which human perception of intended meaning corresponded with the authored intended meaning.

Each schema authored in the previous step was split into two halves based on their different context. Then an annotator was presented with the half-schema in the GUI to identify whether: (1) the intended meaning is to perform an action, (2) the intended meaning is a question about whether the hearer is capable of performing the action, (3) whether the half-schema is ambiguous or not obvious, or (4) whether the text is incoherent (which may be due to the nature of the automated extraction of "Can you..." questions, or due to schema creators who completed the task poorly.)

171 judgments were completed by 12 annotators perfoming 13 judgments each, plus 15 judgments by annotators who did not complete the process of authoring a full set of 13. (15 of these judgments were therefore "duplicate" judgments of schemas that had already been judged.)

The results are shown in Table 2. We note that non-expert validators had achieved a score of 92% on WS, as described in Section 4.3. That was on expert-authored WS, and authoring WS is arguably easier than authoring $S_{ISA}s$, so we expected a fairly low accuracy rate on this, and indeed the result of 37% (63 out of 171) bears this out. We believe this low accuracy was due to both the difficulty of ISA authoring and the use of non-experts, though further work will need to be done to determine this.

| *Validator Response* | *Total* | *Req. Action* | *Ask Ability* |
|---|---|---|---|
| Agreement | **63** | 37 | 26 |
| Disagreement | **76** | 45 | 31 |
| Ambiguous | **26** | 13 | 13 |
| Incoherent | **6** | 1 | 5 |

Table 2: Results of Non-Expert Validation Study 1. Validator responses (agreement/disagreement with intended meaning, ambiguous, incoherent), total counts, and breakdowns by intended meaning (requested action and asking about ability.)

We performed a second validation study to confirm the results of the first one. 155 judgments were completed by 12 annotators performing 13 judgments each, where 1 judg-

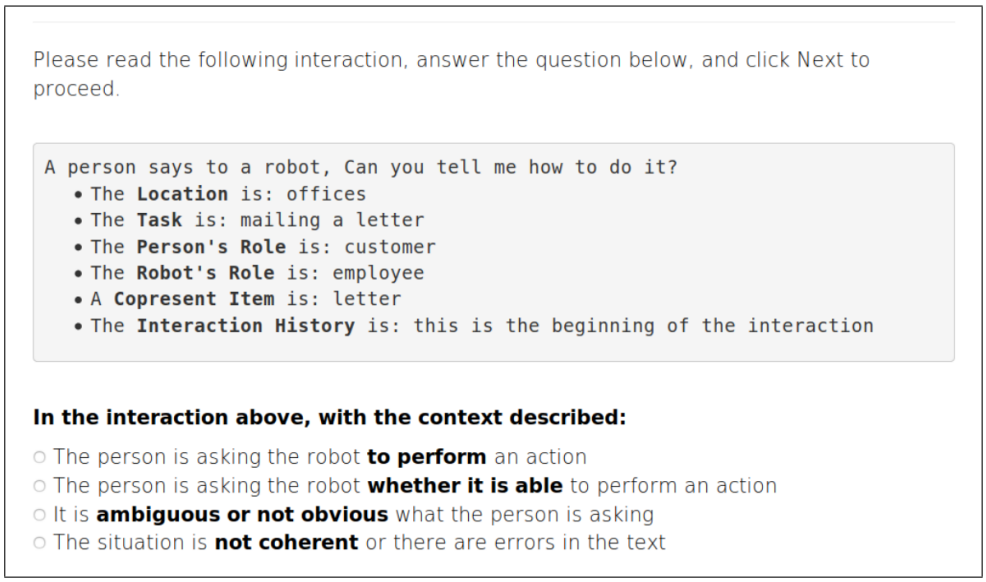


Figure 4: Example of the web-based GUI used for non-expert validation of ISA Schemas.

| *Validator Response* | *Total* | *Req. Action* | *Ask Ability* |
|---|---|---|---|
| Agreements | **68** | 38 | 30 |
| Disagreements | **68** | 36 | 32 |
| Ambiguous | **13** | 6 | 7 |
| Incoherent | **6** | 3 | 3 |

Table 3: Results of Non-Expert Validation Study 2. Validator responses (agreement/disagreement with intended meaning, ambiguous, incoherent), total counts, and breakdowns by intended meaning (requested action and asking about ability.)

ment was left blank. The judgments for this validation study were made on the same schema as the first validation study. The results are shown in Table 3; the human accuracy rating of 44% (68 out of 155) is comparable to the first study.

Another reason to perform a second validation study is to enable inter-rater reliability: in other words, to determine the extent to which crowdsourcing annotators agree on the intention of the authored utterance.

There were 107 comparisons in which an annotator $a_1$ from the first validation study and an annotator $a_2$ from the second validation study looked at the same half-schema and determined that it was neither ambiguous nor incoherent. Of those 107 comparisons, in 30 cases $a_1$ and $a_2$ agreed with the schema author. In 12 of those cases they agreed that it was an ask-ability, and in 18 cases they agreed that it was a request-action. However, recall that these annotations are actually on half-schemas. In fact, of those 30 cases only 1 schema is formed from an appropriate set of half-schemas. This indicates the limits of attempting to rely completely on non-expert authoring of $S_{ISA}s$: it appears to be possible, but the yield is extremely low.

That is why for **Step 5** the expert author serves as tiebreaker for the 107 half-schemas which the validators agreed were coherent and unambiguous (but for which the validators disagreed about the intended meaning). This resulted in a total of 36 full $S_{ISA}s$.

One of the goals of this particular corpus development was to determine the strengths and shortcomings of our approach, so at this point we closely examined and manually edited the entire corpus (whose size had been limited to enable this.) In general, this step would be avoided to minimize expert authoring for scalability purposes. Alternately, a randomly-selected subset of the data could be examined to get a sense of the quality of data being produced.

Generally, the schemas produced after tiebreaking only required a few edits (2-3 of the context features per schema) to reach the level of quality that satisfied an expert author. However, some of the more utterances were ambiguous and therefore especially difficult for the participants to develop a schema from. For example "can you tell me about it" was difficult as most of the time the implied meaning is for the hearer to both respond about their capability and also to perform the action. This was found in most of the "can you $X$ me $Y$" utterances. This suggests that utterances of this particular sub-form are best not used for schema development by non-experts. In total, 6 of the 36 utterances produced by tie-breaking were identified as ambiguous in this way, resulting in a final total of 30 $S_{ISA}s$, or 60 data points for testing an ISA interpretation system.

## 7. Discussion and Future Work

To ensure the quality of the corpora produced, crowdsourcing with non-experts and the extraction of schema elements from corpora should not be seen as techniques that replace expert authors altogether, but instead as techniques that greatly reduce the authoring burden of experts, and also provide a basis for schemas that are grounded in reality. Consider the schema shown in Figure 5, which was produced by non-expert agreement alone (i.e. the two non-expert validators agreed with the non-expert author.) Even this schema would benefit from expert authoring modifications: in context 2, for example, the task could be re-stated as "trying to find the right person to talk to."

| | | |
|---|---|---|
| utterance | *Can you tell me something about the patent law in our country from the book you are reading?* | |
| literal-meaning | *ask-ability* | |
| context-1 | task | *asking about a book* |
| | speaker-role | *student* |
| | hearer-role | *teacher* |
| | location | *school* |
| | copresent-item | *notepad* |
| | interaction-history | *this is the beginning of the interaction* |
| intended-meaning-1 | *request-action* | |
| context-2 | task | *getting contact information* |
| | speaker-role | *administrator* |
| | hearer-role | *student* |
| | location | *phone call* |
| | copresent-item | *phone* |
| | interaction-history | *they have talked about events related to the task* |
| intended-meaning-2 | *ask-ability* | |

Figure 5: ISA Schema authored from non-expert crowdsourcing.

Future work includes further exploring authoring techniques to improve the effectiveness of the non-experts. There are at least two types of improvements we could consider. First, we could explore incremental changes such as refining the GUIs that the non-experts use. For example, we could revise the suggested context feature options in the drop-down menus. In Step 3 of our example corpus development, the non-expert authors often identified one or more contexts which changed the utterance's meaning, but the non-expert authors often failed to identify those contexts which were irrelevant, and often left several "confound" context features such as co-present objects. Also, the non-expert authors often seemed to be confused about the difference between tasks and roles. Although they were generally able to identify a relevant location, the authors often did not take advantage of writing in their own feature values. Finally, the feature values could be modified. For example, we could describe roles in terms of the dynamic between people, such as "transactional" for customer-employee and "leader-follower" for manager-employee.

The second type of change we could make involves breaking up the authoring task performed by non-experts. In Step 4 of our example corpus development, we noted that we achieved a low accuracy rate on non-expert validation, and we suggested that the difficulty of ISA authoring might be a cause of this. Specifically, the non-expert authoring of $S_{ISA}s$ in Step 3 is a challenging mental task involving linguistic phenomena that most people have not extensively reflected upon. It might therefore be better to break up that authoring task into several sub-tasks that people find more intuitive. For example, Step 3a might involve presenting an utterance and asking the non-expert author to imagine a context in which that utterance had the intended meaning "request-action." If so, they would be prompted to write a brief paragraph in plain language explaining *why* the intended meaning followed from the context, specifically listing the relevant context details. Step 3b of the process would involve presenting these brief paragraphs to a second set of non-expert authors, and asking them to author the context feature/value pairs (such as "task: mailing a letter,") i.e. the $c_{1...n}$ of the $S_{ISA}s$. The advantages of this approach are: (1) the cognitive tasks of authoring the scenario and formalizing the context are separated, allowing non-expert authors to focus on one at a time, (2) Step 3b serves as an initial validation check, (3) the $S_{ISA}s$ could then also include a *meta-why-n* field explaining in plain language the scenario author's reasoning for the given interpretation in context $n$. The potential disadvantage of this approach is that it may involve more non-expert authors, although the total amount of time those authors take may be lower; this can be investigated and quantified.

To summarize, we have defined an approach that balances multiple techniques to develop a corpus of $S_{ISA}s$. The utterances in Step 1 are extracted from corpora; the context features in Step 2 are authored by experts from corpora; the initial schemas in Step 3 are authored by non-experts using crowdsourcing; the validations in Step 4 are similarly performed by non-experts using crowdsourcing; in Step 5 an expert author acts as tie-breaker to produce validated schemas. The example corpus development shows that non-experts are capable of developing schema halves which can then be assembled by expert authors.

There is much possible follow-up work involving improving authoring techniques by incremental GUI performance, as well as by splitting up the authoring tasks. Finally, additional issues for further study will doubtlessly be identified when the corpus of $S_{ISA}s$ is actually used to investigate a system's Indirect Speech Act processing.

## Acknowledgments

We are grateful to the anonymous reviewers for their helpful comments.

## 8. Bibliographical References


Amsili, P. and Seminck, O. (2017). A Google-proof collection of French Winograd Schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 24–29.

Bender, D. (2015). Establishing a human baseline for the Winograd Schema Challenge. In *MAICS*, pages 39–45.

Briggs, G. M. and Scheutz, M. (2013). A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Eric, M. and Manning, C. D. (2017). Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the SIGDIAL 2017 Conference*.

Gibbs Jr, R. W. (1979). Contextual effects in understanding indirect requests. *Discourse Processes*, 2(1):1–10.

Levesque, H., Davis, E., and Morgenstern, L. (2012). The Winograd Schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

McDonnell, J., Martin, J., Markant, D., Coenen, A., Rich, A., and Gureckis, T. (2012). psiTurk (version 1.02)[software]. New York, NY: New York University.

Morgenstern, L., Davis, E., and Ortiz, C. L. (2016). Planning, executing, and evaluating the Winograd Schema Challenge. *AI Magazine*, 37(1):50–54.

OED Online. (2019). schema, n. https://www.oed.com/view/Entry/172307 (accessed December 02, 2019).

Perrault, C. R. and Allen, J. F. (1980). A plan-based analysis of indirect speech acts. *Computational Linguistics*, 6(3-4):167–182.

Rahman, A. and Ng, V. (2012). Resolving complex cases of definite pronouns: the Winograd Schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789. Association for Computational Linguistics.

Roemmele, M., Bejan, C. A., and Gordon, A. S. (2011). Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Rus, V., McCarthy, P. M., McNamara, D. S., and Graesser, A. C. (2008). A study of textual entailment. *International Journal on Artificial Intelligence Tools*, 17(04):659–685.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. (2020). WINOGRANDE: An adversarial Winograd Schema Challenge at scale. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*.

Searle, J. R. (1975). Indirect speech acts. *Syntax & Semantics, 3: Speech Act*, pages 59–82.

Storks, S., Gao, Q., and Chai, J. Y. (2019). Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., and Cardie, C. (2019). DREAM: A challenge dataset and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*.

Trichelair, P., Emami, A., Trischler, A., Suleman, K., and Cheung, J. C. K. (2019). How reasonable are common-sense reasoning tasks: A case-study on the Winograd schema challenge and SWAG. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3382–3387, Hong Kong, China, November. Association for Computational Linguistics.

Williams, T., Thames, D., Novakoff, J., and Scheutz, M. (2018). Thank you for sharing that interesting fact!: Effects of capability and context on indirect speech act use in task-based human-robot dialogue. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 298–306. ACM.

Wilske, S. and Kruijff, G.-J. (2006). Service robots dealing with indirect speech acts. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4698–4703. IEEE.

Zellers, R., Bisk, Y., Schwartz, R., and Choi, Y. (2018). SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.