Resolving Indirect Referring Expressions for Entity Selection

Mohammad Javad Hosseini **Filip Radlinski** **Silvia Pareti** **Annie Louis**
Google Research
{javadh,filiprad,spareti,annielouis}@google.com

Abstract

Recent advances in language modeling have enabled new conversational systems. In particular, it is often desirable for people to make choices among specified options when using such systems. We address this problem of reference resolution, when people use natural expressions to choose between the entities. For example, given the choice 'Should we make a Simnel cake or a Pandan cake?' a natural response from a dialog participant may be *indirect*: 'let's make the green one'. Such natural expressions have been little studied for reference resolution. We argue that robustly understanding such language has large potential for improving naturalness in dialog, recommendation, and search systems. We create `AltEntities`[1] (Alternative Entities), a new public dataset of 42K entity pairs and expressions (referring to one entity in the pair), and develop models for the disambiguation problem. Consisting of indirect referring expressions across three domains, our corpus enables for the first time the study of how language models can be adapted to this task. We find they achieve 82%-87% accuracy in realistic settings, which while reasonable also invites further advances.

1 Introduction

Natural dialog often requires resolving referring expressions (REs), not only within and across texts, but also for grounding natural language expressions to specific entities or images. We focus on a specific conversational setting where a speaker's utterance intends to disambiguate between known named entities. While many aspects of RE resolution have been studied extensively, past work has focused on pragmatic reasoning (Dale and Reiter, 1995; Frank and Goodman, 2012), influence of discourse (Orita et al., 2015), and multimodal (e.g., image) context (Zhang et al., 2018).

[1]Our dataset can be found at https://github.com/google-research-datasets/AltEntities

Did you mean a Simnel or **Pandan** cake?
It looks surprisingly green in color
Without any frosting or fruit
It is made from some leaf
Comes from Indonesia
Isn't the Easter one

Table 1: Responses to the question which intend to choose Pandan cake over the alternative.

In the specific case of dialog, when people make choices, the natural REs are not always item names, spatial locations or attributes present in the question. For instance when the choice is among items with similar names (perhaps disambiguating automatic speech recognition errors), or items with difficult to pronounce names, or where the user does not even recall which name is correct but instead recalls some higher level attribute, the user may choose an *indirect* expression (Table 1). Most related to our work, Celikyilmaz et al. (2014) previously studied REs in response to a set of related items (e.g., Harry Potter movies) shown in a user interface. Their work both contains direct (using entity name), indirect, as well as locational (entity's position on the screen) expressions. Predating recent advances in language models (LMs), their best model is a decision tree classifier consuming knowledge graph metadata.

In this work, we created the `AltEntities` corpus by a multi-step process, soliciting crowdworkers to provide diverse yet *realistic* natural expressions for selecting entities in three domains: BOOKS, RECIPES, and MUSIC. To obtain natural and casual dialogic language, we introduce a novel cartoon-based annotation approach (Figure 1). `AltEntities` consists of 6,247 alternative questions (presenting two entities) along with 42,529 REs. In this context, REs are typically definite noun phrases with a pronominal head and a restrictive relative phrase or one of its reduced variants.

Our experiments are based on fine-tuned BERT

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 12313–12335
July 9-14, 2023

(Devlin et al., 2019) and T5 (Raffel et al., 2020) LMs. We assess the representation of entity names as well as other sources of entity information. We find that the results depend significantly on the *type* of entity information provided to the models alongside the REs: If a LM only has access to the entity names but no other information, a case that might happen especially for long tail entities, accuracy is around 60%. On the other hand, if a LM is (unrealistically) given entity information that is identical to that shown to annotators producing the REs, accuracy is very high (up to 95%). However, if the model (more realistically) only has access to generic information that may or may not overlap with annotators' knowledge (Section 5), accuracy of our models is only 82%-87%, leaving significant room for methodological improvements.

2 Related Work

Our work adds to recent efforts to allow users to speak more naturally to conversational systems. Here, we present the most related studies focusing on the properties of REs as well as their resolution.

Alternative Questions. Our questions belong to the class of *alternative* questions (e.g. *'Are you staying or leaving?'*). Several studies have focused on the form and semantics of such questions, and differences from yes/no questions particularly on the basis of prosody (Beck and Kim, 2006; Biezma and Rawlins, 2012; Pruitt and Roelofsen, 2013).

This paper focuses on the deep understanding of answers to such alternative questions when they are posed for selecting between two entities.

Speaker-Listener Cooperation. The research in this space follow the Rational Speech Act Theory (Frank and Goodman, 2012), where the way speakers and listeners reason about each others' intentions and beliefs explains which attributes speakers pick to describe an entity, and how listeners disambiguate the entity. Vogel et al. (2013); Monroe et al. (2017) focus on the pragmatic reasoning involved during the conversation which helps in reaching a common understanding of the topic. Wilkes-Gibbs and Clark (1992) study how REs change as the conversation proceeds. In an experiment, they show that participants start from long and indefinite descriptions of images, but end up with short and definite references. Jordan and Walker (2005) study the subproblem of content and attribute selection for generating object descriptions.

In our data collection, we assume a conversation between two humans in three dialog turns, where the first two turns prime the RE produced in the last turn (Section 3).

Common Ground. In addition to the interlocutors' intentions, their prior or shared knowledge also plays an important role in how they understand each other's utterances. Sometimes the common knowledge arises from a shared situation, e.g., in navigation dialog (Engonopoulos et al., 2013; Misu et al., 2014; Fang et al., 2014) or the presence of a visual space (Yu et al., 2018; Bernardi and Pezzelle, 2021). In the latter, the common ground is given, i.e., it is assumed the image is what all participants in the interaction see in the same way. In many other situations, e.g., in a dialog between two friends about a movie or a book, the common ground is hidden and we can only make assumptions of what information participants share.

In this work, during data collection, we assume that annotators have access to rich common ground involving multiple modalities such as text, image, and video (Section 3.3). During model training inference, we explore performance with varying levels of background information (Sectoin 5.2).

Implicature Understanding. This paper advances the broad area of understanding implicature in dialog. For example, a few recent papers developed datasets and models for indirect boolean responses (without saying 'yes' or 'no') (Pragst and Ultes, 2018; Louis et al., 2020; Takayama et al., 2021; Damgaard et al., 2021). Interestingly, Ruis et al. (2022) shows that LLMs cannot solve such implicatures in a zero-shot setting.

RE resolution. There are few prior studies around the data and models for resolution tasks such as ours. Stoyanchev et al. (2021) built a method where references to items from prior context in a dialog are resolved by detecting state updates. Unlike our work, their REs focus on attributes (e.g., *Italian* in *the Italian restaurant*) discussed in prior dialog. Celikyilmaz et al. (2014) collect REs to a target item among others shown on a screen (e.g., a set of Harry Potter movies). Their expressions contain both direct (reference to entity name) and indirect references, where the latter comprise about 25% of the data (≈ 6K REs). To aid the resolution of indirect ones, they include features which capture the overlap between an expression and knowledge graph attributes for each item.

Our work creates a large scale corpus (42K REs) exclusively for indirect REs, and explores how LMs

encode the knowledge for disambiguation.

3 Collecting Rich Referring Expressions

To maximize generalizability, we collect data in three domains: BOOKS, RECIPES, and MUSIC. These were selected to cover a diverse variety of entity types with different kinds of available information — e.g. plot summaries for books, images for recipes, and lyrics and videos for songs. We performed careful and detailed annotations, and explain the annotation steps in this section.

3.1 Cartoon-driven Annotation Setup

Previous work in question-answering and dialog typically asks annotators to complete text-based input boxes (Rajpurkar et al., 2016; Choi et al., 2018; Rajpurkar et al., 2018; Reddy et al., 2019; Eric et al., 2020). We employ a novel cartoon-bubble completion method, aiming to immerse annotators in the dialog setting to obtain more natural and informal REs. We start with a brief overview of the setup, and then explain the steps in detail.

Figure 1 shows the first (of our two) annotation screens. Annotators are shown a cartoon with two characters (*Bob* and *Alice*) in a fictional conversation, and asked (as Bob) to complete the last speech bubble. This pictorial depiction, and the casting of the dialog as a casual chat between friends encourage the annotators to produce friendly, short, and dialogic responses. However, annotators are generally unlikely to know details about entities sampled from a collection. Therefore, we also provide background information on the entities (bottom of Figure 1), corresponding to *common knowledge* that the two characters could share on the topic.

After annotators are shown this information, they proceed to a second screen (Figure 2). It indicates one of the entities (books in this example). They are asked to describe that entity (indirectly) with 3 to 5 responses: We found eliciting more entries encourages diversity and depth in the responses. Our data consists of the entity pairs, their descriptions, the target entity, and annotator expressions.

From Figure 2, note that once on the response screen, annotators cannot re-read descriptions. This encourages recall from memory. The reasoning behind this, and many other aspects of this design, are explained in the next sections.

3.2 The Conversational Cartoon

The cartoon has three cells as shown in Figure 1. The first is a domain-specific utterance intended to set context. For example, *'Remember that book we saw at the store?'* sets up the dialog as one recalling a specific book. These utterances are from a set of five manually written expressions for each domain, with one selected at random for each conversation. Examples in the RECIPES and MUSIC domains are *'That recipe on today's Masterchef was too good!'* and *'You sang that song really well yesterday.'* Appendix A shows all these utterances.

The *alternative* question is presented in the second cell. This question follows a fixed template: *Do you mean 'A' or 'B'?* where 'A' and 'B' are the names of two *related* entities. Our entities are sampled from Wikipedia page titles, with any disambiguation parentheses removed. When the names are identical, we retain the Wikipedia disambiguation: For instance, one such question is *Do you mean 'The Gladiator (Turtledove novel)' or 'The Gladiator (Scarrow novel)'?*.

The third cell is completed by the crowdworkers, assuming the role of *Bob* to enter text that refers to the target entity. They enter those expressions as shown in Figure 2. Further screenshots of our interface for all domains are provided in Appendix B.

3.3 Entity Background

In real dialogs, when people differentiate between options, they draw on partial knowledge about entities that they recall. We aimed to foster a similar situation in our corpus, while doing so in a controlled manner without requiring domain-expert annotators. As such, when selected entities are shown to annotators, they are also presented with background information (bottom of Figure 1). We draw the background also from Wikipedia, biasing towards sections relevant to each domain. For BOOKS, these are the *main* (first) and *plot summary* sections. For RECIPES, we used the *main*, *preparation*, and *ingredients* sections. For each entity, up to 750 characters of *one* of these sections are shown on the interface. For RECIPES, the food's image[2] is also always shown to help the annotators quickly realize what it looks like (Figure 3).

For MUSIC, however, we found Wikipedia text to be less useful: Pages contain details and trivia (e.g., *5th single on the album* or *sold 4 million copies*), which we judged unlikely to be included

[2]We filtered out examples without any images.

In the following screen, you will be asked to refer to one of them.

The Sympathizer	The Underground Railroad
• The Sympathizer is the 2015 debut novel by Vietnamese American professor Viet Thanh Nguyen. It is a best-selling novel and recipient of the 2016 Pulitzer Prize for Fiction. • Set as the flashback in a coerced confession of a political prisoner, the book tells the story of the South Vietnamese Government in 1975 and subsequent events in American exile in Los Angeles, through the eyes of a half-Vietnamese, half-French undercover communist agent.	• The Underground Railroad is a historical fiction novel by American author Colson Whitehead, published by Doubleday in 2016. • The alternate history novel tells the story of Cora and Caesar, two slaves in the antebellum South during the 19th century, who make a bid for freedom from their Georgia plantation by following the Underground Railroad, which the novel depicts as a rail transport system with safe houses and secret routes.

Figure 1: Annotators were shown a cartoon in which they were asked to complete the final step of a conversation.

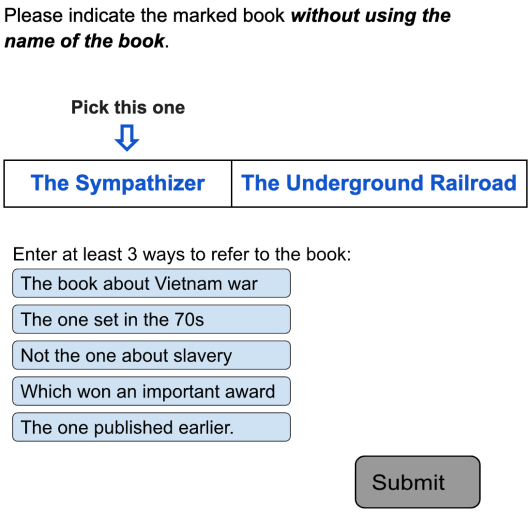

Figure 2: Annotation screen for entering expressions.

Figure 3: Background descriptions for two recipes.

in natural background knowledge about a song. On the other hand, song lyrics and music are very relevant in this domain, but are not usually found in Wikipedia. Consequently, we presented a Google search link for the song in the background section, and asked the annotators to listen to at least some of each song, and read about them before writing expressions. The search query contained the song's title and its artist, e.g., *Hello (by Adele)*. Since information about the song comes from search, we also biased our candidates towards popular songs, which have more detailed results (Section 3.4).

3.4 Generating Alternative Questions

The alternative questions (*Do you mean 'A' or 'B'?*) are generated automatically: (i) Candidate entities are extracted from English Wikipedia for each domain (Section 3.4.1), then (ii) we substitute 'A' and 'B' by sampling entity pairs (Section 3.4.2).

3.4.1 Selecting Candidate Entities

For each domain, we collect English Wikipedia articles by checking the presence of certain Wikipedia templates (infoboxes[3]), and the presence of particular sections: For RECIPES, we additionally included articles with an *ingredients* section.

This set was then filtered to exclude very short articles, or those ambiguous between domains. For MUSIC, we use article length (number of sections/subsections) as a proxy for popularity, and choose the top $\approx$ 1000 articles. To remove any sensitive or offensive content, we also filter articles whose content matches a list of sensitive words. Appendix C contains the details of the above filters. Table 2 shows the number of candidate entities.

[3]Infoboxes are fixed-format tables that consistently present articles in a given category (e.g., all books).

	BOOKS	RECIPES	MUSIC
Main	22,763	2,822	1,032
Plot Summary	5,858	-	-
Preparation	-	343	-
Ingredients	-	147	-
Total	28,621	3,312	1,032

Table 2: Number of extracted candidate items for each domain and background section.

3.4.2 Sampling Entity Pairs

Much linguistic work on alternative questions has focused on the semantics and pragmatics of these utterances (Biezma and Rawlins, 2012), but we also need to make decisions about which entity pairs could make for a challenging disambiguation problem. Entity pairs sampled uniformly at random are less likely to be interesting, since they may not share many properties, making disambiguation easier. In this work, we develop entity pair sampling techniques at different similarity levels, as a proxy for disambiguation difficulty.

Uniform sampling. Entity pairs are sampled uniformly at random from the domain.

Same name. These entities have the same name in Wikipedia followed by a disambiguation phrase within parentheses. An example is *Dawn (McLaughlin novel)* and *Dawn (Andrews novel)*.

Similar title. These entities have a similar title in terms of character edit distance (distance ≤ 3), where the title could optionally consists of a disambiguation phrase within parentheses.

Similar description. This method looks for deeper similarity within the text of Wikipedia articles: We sample a first entity uniformly, then select the second with the highest similarity using a Universal Sentence Encoder (Cer et al., 2018). The input to the encoder is the Wikipedia section shown as the background knowledge to annotators.

Similar infobox attributes. Here we take entities that share important domain-specific properties, e.g., recipe origin, or the song genre. We match entities (except BOOKS) using the 'attributes' listed in the Wikipedia infobox: {*type*} and {*type*, *country*} for RECIPES, and {*genre*}, {*artist*}, and {*genre*, *artist*} for MUSIC.

We applied the **same name** method only to BOOKS, and the **similar title** method only to BOOKS and RECIPES. The other domains did not contain enough such examples. We applied the **similar description** method to all domains. We applied the **similar infobox attributes** method to RECIPES and MUSIC, but not the BOOKS domain; however, some pairs with identical attributes were already covered by the other methods for BOOKS. Table 3 shows the number of sampled entity pairs for each domain and sampling method.

	BOOKS	RECIPES	MUSIC
Uniform	649	813	700
Same Name	282	-	-
Similar Title	497	280	-
Similar Desc	650	583	700
Similar Attrs	-	418	675
All	2,078	2,094	2,075

Table 3: Number of sampled entity pairs (questions) for each domain and sampling method.

Do
✓ Keep it casual and conversational.
✓ Varied, interesting, and creative expressions.
✓ Use alternative words, e.g., *award* instead of *prize*.
✓ Vary the phrasing: *the book about*, *I meant the*, *was thinking of*, *the one about*, *I wasn't referring to*, etc.
Don't
✗ Mention the book by name or position (e.g., the second one).
✗ Use too detailed information that *Alice* may not recall (eg. *1992* or *in the 90s* are better choices than *Sep 9 1992*).
✗ Copy whole sentences from the description.

Table 4: Actions annotators were encouraged (Do) or discouraged (Don't) to take for the BOOKS domain.

3.5 Annotator Instructions and Pilot Runs

To maximize RE naturalness, we also provided annotators different domain-specific examples. Figure 2 shows those for the book *The sympathizer*. The REs are about topic (*about Vietnam war*), timeline (*set in the 70s*), and contrasts (*Not the one about slavery*, and *The one published earlier*). They also emphasize use of general statements instead of overly specific and unrealistic ones, e.g., *set in the 70s* instead of *1975*. Table 4 shows a detailed note on desirable expressions.

We performed pilot studies to understand how annotators responded to our instructions, and used these to refine the instructions. A first study (for BOOKS) examined how annotators should use the background text, comparing designs where annotators could, or could not, go back-and-forth between the description screen (Figure 1), and the data collection screen (Figure 2). With back-and-forth possible, the responses contained excessive details, e.g., reiterating large portions of background text (*The book that was last of three juvenile novels that Wollheim wrote for Winston*). With back-and-forth removed, annotators produced shorter REs (7.99 vs 9.61 words), with fewer proper nouns and numbers per RE (0.43 vs 0.88) as they are harder

to remember. They also used more contrastives, e.g., starting with *'not the'* (21.8% vs 2.2%) which involve drawing on information about both books. Thus, we adopted the memory recall setting.[4] After the first pilot study, we performed one pilot per domain for relatively small instruction refinements.

4 The AltEntities Corpus

Our annotations were carried out using a pool of around 60 in-house crowdworkers.[5] They were all native English speakers recruited from U.S., U.K., Canada, and Australia so as to obtain a diverse set of perspectives.[6] Each question was shown to two workers to get multiple inputs per question. Around 2K entity pairs were annotated for each domain resulting in around 42K expressions in total. Table 5 shows the final corpus statistics, and Table 6 shows example expressions for the three domains. We release the dataset under the CC-BY SA 3.0 License as per the Wikipedia License.

The REs for BOOKS were on average a word longer than for other domains. They also contained more named entities per expression. Each domain contains some repeated REs (e.g., *the pop song*), that are often high-level responses, e.g., a song's genre. The BOOKS domain contains the most unique responses. The number of contrastives, estimated as REs starting with "not the", are from 8% in MUSIC up to 20% in BOOKS.[7] For MUSIC and RECIPES, we manually checked 200 random REs for references to modalities other than text. Around 10% multi-modal REs were present in the RECIPES domain (mostly color), and 20% in the MUSIC domain (mostly beat, speed, and mood).

We estimated the RE error rate by manually inspecting 40 question samples (around 250 to 300 expressions) per domain. The error rate is between 4.5% to 6.8% for the three domains. 78% of these errors were due to the RE applying to both items, not just the target entity. The remaining errors were mostly due to confusing the two entities. We also note that the rate of exact string match between REs and Wikipedia text is $< 1\%$.

The annotators were inspired by the provided stylistic cues in the instructions (e.g., starting with *the one* or *I meant the*), but followed our guidelines to vary their responses as well. We observed that the content of REs (e.g., timeline, lyrics, singer or band information, instrument) included both the categories covered by the provided examples (e.g., timeline for books and songs) and novel categories (e.g., background information on books and songs such as *The one inspired by a Rolling Stones song*).

	BOOKS	RECIPES	MUSIC
# Questions	2,078	2,094	2,075
# Expressions	13,144	15,046	14,339
Length (words)	7.8	6.2	6.8
# Named Entities	0.7	0.2	0.4
Unique	96%	86%	76%
Contrastives	20%	9%	8%
Multi-modality	-	10%	20%
Estimated Error rate	4.5%	6.7%	6.8%

Table 5: The AltEntities corpus statistics

BOOKS
The one that is set in the 1880s
It's by a famous detective writer
The fictional one
not the one with the 12 year old boy
It's the book that has rock and politics in it
MUSIC
The one without words
It is the song sung by an Australian.
It has synthesizer sounds in it
Came out in mid of 2000.
Based on life experienced in Sheffield.
RECIPES
comes from Azerbaijan
The Japanese steamed cake
The ones eaten at Christmas
cornmeal is the main ingredient
Not the one with dried peaches.

Table 6: Random REs from crowd annotators.

5 Task and Models

Indirect reference resolution can be defined as follows: Given an alternative question with K choices[8] $C = \{c_1, \ldots, c_K\}$, and a RE r, models should disambiguate the choice $c^* \in C$ intended by r. We assume r does not *directly* mention c^* by its name or position, but does uniquely *refer* to c^*.

5.1 Information Available to Models

At a minimum, all models require the RE r and the names of the choices $C = \{c_1, \ldots, c_K\}$. In addition, models may use textual descriptions $\{s_1, \ldots, s_K\}$ to aid disambiguation. We define

[4]Note that the MUSIC entities are provided with search links which open in a new page, making back-and-forth possible, although it was discouraged in the guidelines.

[5]Paid contractors who work with our institution on such tasks.

[6]The average number of questions per annotator is 217. The minimum number of annotations was 10, and the maximum was 2015 questions, followed by 610 questions. Around 80% of annotators annotated around 100-600 questions each. We did not observe any obvious correlation between dataset artifacts and specific annotators.

[7]This estimate gives a lower bound as there are other types of contrastives expressions such as *the newer song*.

[8]In this paper, we only consider K=2.

choice text s'_i ($1 \leq i \leq K$) as: (a) The entity name c_i, or (b) the concatenation of c_i and the textual description s_i, separated by a delimiter.[9] We consider the following four experimental setups.

NAME: The entity name without further description of the entities. We use this setting as a baseline.

For the remaining models, we add the following description to the name (truncated to 512 tokens):

INFOBOX: The concatenation of all infobox key-value pairs (e.g., *'genre: pop'*).

UNSHOWN BACKGROUND: The INFOBOX text, concatenated with all the Wikipedia sections of the entity, *excluding* the section shown to the annotators as background. Since annotators were shown a search link and not a specific Wikipedia section for the MUSIC domain, we do not remove any Wikipedia section for the MUSIC entities. We note that the UNSHOWN BACKGROUND might have some overlap with the information shown to crowdworkers, but the text is not directly given to them. Hence, it is a fair setup to evaluate models in a practical system where the models might not have all the background information.

ORACLE: The same background text that was shown to the annotators (Section 3.3). Note that this only exists for BOOKS and RECIPES, as for MUSIC, annotators were only shown a search link.

5.2 Models

We evaluated 5 different models. For each, we score match to each entity choices and select c^* with the highest score value.

Universal Sentence Encoder: We calculate the cosine similarity between the universal sentence encoder (USE; Cer et al.2018) embeddings for the RE r and each choice's text s'_i.

Entailment: Using a textual entailment classifier, we classify whether a choice's text s'_i entails the RE r. We use the confidence of the 'entailment' label as the score. We use a BERT model trained on the MNLI dataset (Williams et al., 2018) as our classifier. For all models based on BERT, we use BERT large uncased.

BERT. We turn our task into binary classification: We make one example per choice (c_i, r) with label 1 if r refers to c_i; otherwise, label 0. We fine-tune BERT with a binary classification layer (with two units) on top of its [CLS] token embeddings. The LM input is the sequence [CLS]s'_i[SEP]r. During inference, for each choice c_i, we compute the probability of label 1 as its score.

BERT Joint. In contrast to the above binary setup, we encode all the K sequences [CLS]s'_i[SEP]r with BERT. We apply a linear layer (with one unit) on top of the [CLS] token embeddings from each sequence. We normalize the scores using softmax. Finally, we minimize a categorical cross entropy loss given the K scores. During inference, we directly use each choice's score.

T5. We turn our task into binary classification, as with the BERT binary model. We fine-tune a T5 XL model (3B parameters) with input sequence "expression: r entity: c_i description: s_i" and output sequence 1 or 0. For the NAME input type, the input sequence omits the "description" part.

6 Experiments

We split the questions in the `AltEntities` corpus in each domain into training (70%), development (15%), and test (15%) sets. To avoid information leaking between the sets, we allow each *target* item to be in only one of the sets. For the USE and entailment models, we do not tune any hyperparameters. For supervised models, we tune the learning rate, batch size, and number of epochs using a grid search on the development data (96 configurations for BERT and 24 configurations for T5). We report the hyper-parameter details in Appendix D.

6.1 Reference Resolution Accuracy

We compute the accuracy of each (alternative question, RE) pair, i.e. whether the correct choice is scored highest. As K=2 in our experiments, a random baseline has accuracy 50%.

We show the test set results in Table 7 for all domains and input types.[10] For each model, we also show the average results of all input types. Among the models, USE performs worst (61.03%), followed by the entailment model (66.91%). BERT Joint (73.56%) is on average 1.61% better than BERT (71.52%), confirming that modeling the choices jointly is effective. T5 has the highest average results (77.43%), as expected given that we experimented with T5 XL with 3B parameters compared to BERT large with 360M.

In the ORACLE setting for BOOKS and RECIPES, accuracy is understandably high (up to 95.10% for BOOKS and 92.60% for RECIPES). We note that

[9] It is possible to use other modalities, e.g., recipe images or music videos; however we focus on text only.

[10] The development set results (Appendix E) are slightly higher, but exhibit similar patterns.

	Books				Recipes				Music			
	Orac	Name	InBo	UnBa	Orac	Name	InBo	UnBa	Name	InBo	UnBa	Avg
USE	67.25	54.35	56.65	60.40	69.28	55.73	63.75	65.00	***57.83***	61.05	60.08	61.03
Entailment	84.95	52.15	63.65	68.80	79.98	54.08	67.14	74.41	54.52	64.49	71.84	66.91
BERT	93.30	50.55.	74.35	79.80	87.87	53.32	77.84	81.01	53.93	61.60	73.13	71.52
BERT Joint	***94.05***	**59.80**	75.35	***81.50***	88.94	54.12	75.21	80.87	***56.59***	67.48	75.24	73.56
T5	**95.10**	55.65	**78.30**	**83.40**	**92.60**	**61.97**	**83.33**	**86.76**	**58.11**	**74.28**	**82.27**	**77.43**

Table 7: Indirect reference resolution results for different models on all domains and input types: Oracle (Orac), Name, Infobox (InBo), Unshown Background (UnBa). The best result of each column is boldfaced. When the difference between the best result and another result is not statistically significant (paired t-test with p-value < 0.05), the other result is made both bold and italic (only 4 cases).

		Test Domain		
		Books	Recipes	Music
Training Domain	Books	83.40	83.55	82.54
	Recipes	81.60	86.76	82.96
	Music	82.05	84.80	82.27
	Mixed	**83.90**	**87.47**	**83.28**

Table 8: T5 results for the Unshown Background setup, when trained on one domain and tested on another domain.

	Books	Recipes	Music
Uniform	90.30	92.54	88.58
Same Name	85.02	-	-
Similar Title	83.86	86.29	-
Similar Desc	74.70	82.24	80.39
Similar Attrs	-	81.55	77.12
All	83.40	86.76	82.27

Table 9: T5 results with different sampling methods for each domain with Unshown Background input.

these results are an over-estimate of the model capabilities. On the other hand, in the Name setting, in most cases the results are slightly above 50%, with the best result being 61.97% for the Music domain with the T5 model. Here the LMs rely on their memorized entity knowledge (Petroni et al., 2019), suggesting that BERT and T5 embeddings are not sufficient to resolve arbitrary entity references.

With the Infobox input, the T5 model accuracy is 78.30%, 83.33% and 74.28% for Books, Recipes, and Music, respectively. It increases to 83.40%, 86.76%, and 82.27%, respectively, with the Unshown Background input where we add unstructured text data to the structured infobox data. This shows the text is helpful when resolving REs. In practical settings, models should work with relevant, but not necessary the same background knowledge as users because (1) it is not possible to have access to users' actual knowledge, and (2) models always have some limitation in the amount of text they can input. We thus rely on the Unshown Background setting as a realistic setting for measuring the capabilities of the different models.

6.2 Cross-Domain Experiments

Reference resolution is a semantic task, and ideally models would learn general task aspects rather than domain details. We test generalization by fine-tuning our models on one domain and testing on another. We used the Unshown Background setting for these experiments as the most realistic.

Table 8 shows the T5 model results.[11] We do not observe much difference when models are tested out of domain, supporting the hypothesis that our models are indeed generalizable. This observation is rather important since our models could be used without separate training for new choice domains.

We also create a *mixed* training (and development) set that combines the data of the three domains. The mixed training set gives better results on average, taking advantage of larger training set and cues from all the domains. However, since the dataset in each domain is relatively large, the mixed training does not increase the results substantially.

6.3 Results and Entity Similarity

Section 3.4.1 explained how we selected entity pairs to have different levels of similarity. We now examine how this affects performance. Table 9 shows the results for the T5 model with the Unshown Background input. We compute accuracy per test example subset, where each originated from a specific similarity sampling method.

As expected, when the two entities are randomly selected, disambiguation is easiest since they have little in common. The task becomes harder as entities become more similar, with entities with similar infobox attributes having the lowest performance.

6.4 Error Analysis

We analyzed the errors from the T5 model in the Unshown Background setting, to understand

[11] We observe similar results with BERT Joint and BERT models, which are not shown due to space limitations.

Error Type	Target Item	Non-Target Item	Annotator Utterance
No Textual Overlap 47%(B) 27%(R) 42%(M)	**Best Song Ever** is a song recorded by English-Irish...	**These Days** is a song by British pop group...	It has to do something with dancing all night.
	Boerewors..., a type of sausage which originated in South Africa.	**White pudding** is a meat dish popular in Ireland, Northern Ireland...	It can be stewed.
Poor reasoning 25%(B) 18%(R) 13%(M)	**Clams casino** is a clam "on the halfshell" dish...	**Buddha's delight** ... is a vegetarian dish...	The one with seafood in sauce.
	Dark Age... release_date: July 30, 2019...	**Iron Gold**... release_date: January 16, 2018...	It is the most recent one.
Multi-modality 0%(B) 25%(R) 22%(M)	**It's Not Over** is the debut single by American rock...	**Love Child** is a 1968 song released by the Motown...	Has a marriage proposal in the music video
	Pandoro appeared in remote times, the product of...	**Pandebono**... It is said that an Italian baker who lived...	Brownish-yellow in its colour.
Wrong Annotation 28%(B) 30%(R) 23%(M)	**My Story (Gillard book)** is a political memoir of Julia Gillard...	**My Story (Das book)** is an autobiographical book written by Indian author...	I mean the book that is technically an autobiography.
	Tight Connection to My Heart (by Bob Dylan)...	**Like a Rolling Stone** (by Bob Dylan)...	this song is by an American singer.

Table 10: Error analysis results. Under each error type, we report the percentage of examples from the BOOKS (B), RECIPES (R), and MUSIC (M) domains. We also show two example for each error type.

if there are systematic errors which could be improved upon in the future. We manually analyzed 40 incorrectly predicted development set examples per domain. We show four different error types and their percentages per domain in Table 10.

In most cases, there is no textual overlap between the RE and the background. This is because either the relevant text is removed (by design) since it is shown to the raters, or the Wikipedia text does not contain the information at all (e.g., music lyrics). Future research could evaluate how to adapt LMs to improve their entity knowledge to reason beyond the input textual evidence. In addition, retrieval augmented LMs could be applied to retrieve relevant information before performing the prediction (Borgeaud et al., 2022; Shi et al., 2023).

In other cases, the model suffers from poor reasoning, e.g., that clam is seafood, or a vegetarian dish does not contain seafood. In addition, the model often misclassifies examples when entity attributes are compared (e.g., *the newer one*). Multi-modality covers around 25% of the errors in the RECIPES and MUSIC domains, e.g., annotators referenced visual aspects from music videos or recipes (e.g., *looks like shells*), or an acoustic aspect from a song (e.g., *with the piano intro* or *more upbeat*).

The remaining errors are because of wrong annotations, usually with the REs appling to both items. This wrong annotation rate (23%-30%) is much higher than the error rate in the whole dataset (less than 7% as discussed in Section 4) since the model has learned the task to a good extent.

We also analyzed correctly classified examples (for the MUSIC domain) to understand what types of REs are classified correctly. The results are shown in Appendix F.

7 Conclusion

We have revisited RE resolution with a new focus on indirect expressions, introducing `AltEntities`, a new large dataset for this task – covering BOOKS, RECIPES, and MUSIC examples. The dataset was collected using a novel cartoon completion approach to encourage conversational and causal expressions while avoiding name or position expressions. The experimental results show that in a realistic setting, LMs adapted for this task achieve 82%-87% accuracy. While an improvement on existing approaches, this also encourages further research on this important problem. Moreover, we showed that the models' performance does not drop when trained and tested on different domains, suggesting that models can learn the semantic task well and generalize to new domains.

It is notable that in practice, many entities do not have textual descriptions or rich meta-data. Future research could study resolving REs with minimal information, e.g., when we only have access to their names or limited meta-data. Future research could also use multi-modal input for training and inference. Further, to handle more complex REs such as *the newer one*, or *the happy song*, one could decompose a RE into simpler expressions and then perform the comparison. Similar data collection methodologies could be applied to collect a dataset with more number of choices and also cases where neither or multiple choices match the RE.

8 Limitations

As with any natural language understanding task, there are practical limitations and related ethical aspects that must be considered before deploying a system. In particular, our corpus and modeling approach assume that the user-provided REs *always* refer to one of the two options. If this is not the case, or if the RE is particularly contrived, undesirable or unexpected behavior may occur: For any expression, including for instance one made with arbitrary derisive language, the model would attempt to resolve this to one of the alternative entities. One approach system designers may consider could be to pre-classify any user-provided REs to avoid interpreting those that are off topic or phrased in a negative manner.

A second consideration is that of corpus representativeness. In our case, as this is a first corpus for this task, we have limited ourselves to English Wikipedia, native English speaking annotators, and particular item sampling strategies for practical reasons. However, if used for training a deployed system, the examples present may bias any model to understand specific types of references but not others. Similarly, the items in our corpus are sufficiently popular to have a relatively long Wikipedia entry, whereas items not present in Wikipedia, or with only minimal information, may exhibit different characteristics.

9 Ethics Statement

The data collection protocol was reviewed by an ethics panel to remove potential ethical concerns. A few ethical concerns were mentioned by the panel which were then judged to be handled well. These included ensuring that the entities, texts and REs were free from biased and sensitive language. We address this by filtering using a list of sensitive words (see Section 3.4.1 and Table 12). The panel also recommended a diverse representation of entities and domains. Thus our data comes from diverse domains and the entities are sampled from a large set of Wikipedia articles.

Still, we note that the limitations mentioned in Section 8 need to be considered and addressed carefully when using our dataset or models for evaluation or training of a deployed system. In addition, a biased corpus may lead to an evaluation that is unaware of RE language forms used in other cultures and languages, or that refer to other types of items. We expect this consideration to be important in practical settings.

References

Sigrid Beck and Shin-Sook Kim. 2006. Intervention effects in alternative questions. *The Journal of Comparative Germanic Linguistics*, 9(3):165–208.

Raffaella Bernardi and Sandro Pezzelle. 2021. Linguistic issues behind visual question answering. *Language and Linguistics Compass*, 15(6):elnc3–12417.

María Biezma and Kyle Rawlins. 2012. Responding to alternative and polar questions. *Linguistics and Philosophy*, 35(5):361–406.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Asli Celikyilmaz, Zhaleh Feizollahi, Dilek Hakkani-Tur, and Ruhi Sarikaya. 2014. Resolving referring expressions in conversational dialogs for natural user interfaces. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2094–2104, Doha, Qatar. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *EMNLP*.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

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.

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.

Nikos Engonopoulos, Martin Villalba, Ivan Titov, and Alexander Koller. 2013. Predicting the resolution of referring expressions from user behavior. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1354–1359.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Rui Fang, Malcolm Doering, and Joyce Chai. 2014. Collaborative models for referring expression generation in situated dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Pamela W Jordan and Marilyn A Walker. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

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.

Teruhisa Misu, Antoine Raux, Rakesh Gupta, and Ian Lane. 2014. Situated language understanding at 25 miles per hour. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 22–31, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in Context: A Pragmatic Neural Model for Grounded Language Understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Naho Orita, Eliana Vornov, Naomi Feldman, and Hal Daumé III. 2015. Why discourse affects speakers' choice of referring expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1639–1649, Beijing, China. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473.

Louisa Pragst and Stefan Ultes. 2018. Changing the level of directness in dialogue using dialogue vector models and recurrent neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 11–19, Melbourne, Australia. Association for Computational Linguistics.

Kathryn Pruitt and Floris Roelofsen. 2013. The interpretation of prosody in disjunctive questions. *Linguistic inquiry*, 44(4):632–650.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators. *arXiv preprint arXiv:2210.14986*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Svetlana Stoyanchev, Simon Keizer, and Rama Doddipatla. 2021. Action state update approach to dialogue management. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7398–7402. IEEE.

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.

Adam Vogel, Christopher Potts, and Dan Jurafsky. 2013. Implicatures and nested beliefs in approximate decentralized-POMDPs. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 74–80, Sofia, Bulgaria. Association for Computational Linguistics.

Deanna Wilkes-Gibbs and Herbert H Clark. 1992. Coordinating beliefs in conversation. *Journal of memory and language*, 31(2):183–194.

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

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. 2018. Grounding referring expressions in images by variational context. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4158–4166.

A Opening Utterances

The first annotation screen (Figure 1) starts with a manually written opening utterance. Table 11 shows all these utterances for the three domains..

B Annotation Guidelines

In this section, we provide the domain-specific guidelines that were shown to the annotators prior to the start of their annotation. The guidelines for each domain includes three *instruction* screens. The second and third instruction screens are then repeated for each alternative question as their first and second *annotation* screens, respectively (the two screen discussed in Section 4).

In the first instruction screen, a summary of the task based on a cartoon completion setup is shown to the annotators. Figure 4 shows the first instruction screen for the BOOKS domain. We do not show the first instruction screen for the other two domains as they are very similar to the BOOKS domain except that the text is slightly different to reflect the domain, and that the examples are from those domains.

The second instruction screen provides further information about the task and describes where the annotators should acquire the knowledge to perform the annotations. Figures 5, and 7, and 9 show the second instruction screens for the BOOKS, RECIPES, and MUSIC domains, respectively.

The third instruction screen shows which item should be referred to, and lists five examples of appropriate REs. The REs cover different aspects of the items to encourage the annotators to cover a variety of the item aspects. It also lists a number of actions that the annotators should or should not do. Figures 6, 8, and 10 show the third instruction screen for the BOOKS, RECIPES, and MUSIC domains, respectively.

C Filtering Wikipedia Articles

Table 12 shows a number of filters we applied to narrow down the extracted articles.

D Hyper-parameters Details and Computing Infrastructure

We tune the hyper-parameters using a grid search based on the accuracy of the indirect reference resolution task on the development set of each domain. For BERT and BERT multiple choice models, we select the base learning rate from $\{1e{-}4, 5e{-}5, 3e{-}5, 1e{-}5, 5e{-}6, 3e{-}6, 1e{-}6, 5e{-}7\}$, the training batch size from $\{16, 32, 64\}$, and the number of epochs from $\{1, 3, 5, 10\}$. For T5, we select the base learning rate from $\{5e{-}7, 1e{-}7, 3e{-}6, 5e{-}6, 1e{-}5, 3e{-}5, 5e{-}5, 1e{-}4\}$ and the training batch size from $\{16, 32, 64\}$. We train the T5 models for 50K steps (batches).

Table 13 shows the selected hyper-parameters for each model, domain, and input type.

We used Cloud TPU v2 accelerators for both training and inference. In our experiments, each training epoch took on average around 4 minutes for BERT, 6 minutes for BERT Multiple Choice, and 15 to 25 minutes for T5 models.

E Development Set Results

We reported the test set results in multiple settings in Section 6. In this section, we report all those results on the development sets.

Table 14 shows the development set results of different models for all domains and input types. We note that the general trends are very similar to that of the test sets. On average, the results of different models are slightly higher for the development set compared to the test set (up to 2.35%). This is expected as we have tuned the hyper-parameters on the development sets.

F Analyzing Correctly Classified Examples

We analyzed 100 correctly classified examples in the MUSIC domain and assigned one or more categories (e.g., *date* or *genre*) to each example. We used the predictions of our T5 model with the UNSHOWN BACKGROUND input. Table 15 shows the results which cover a wide range of categories.

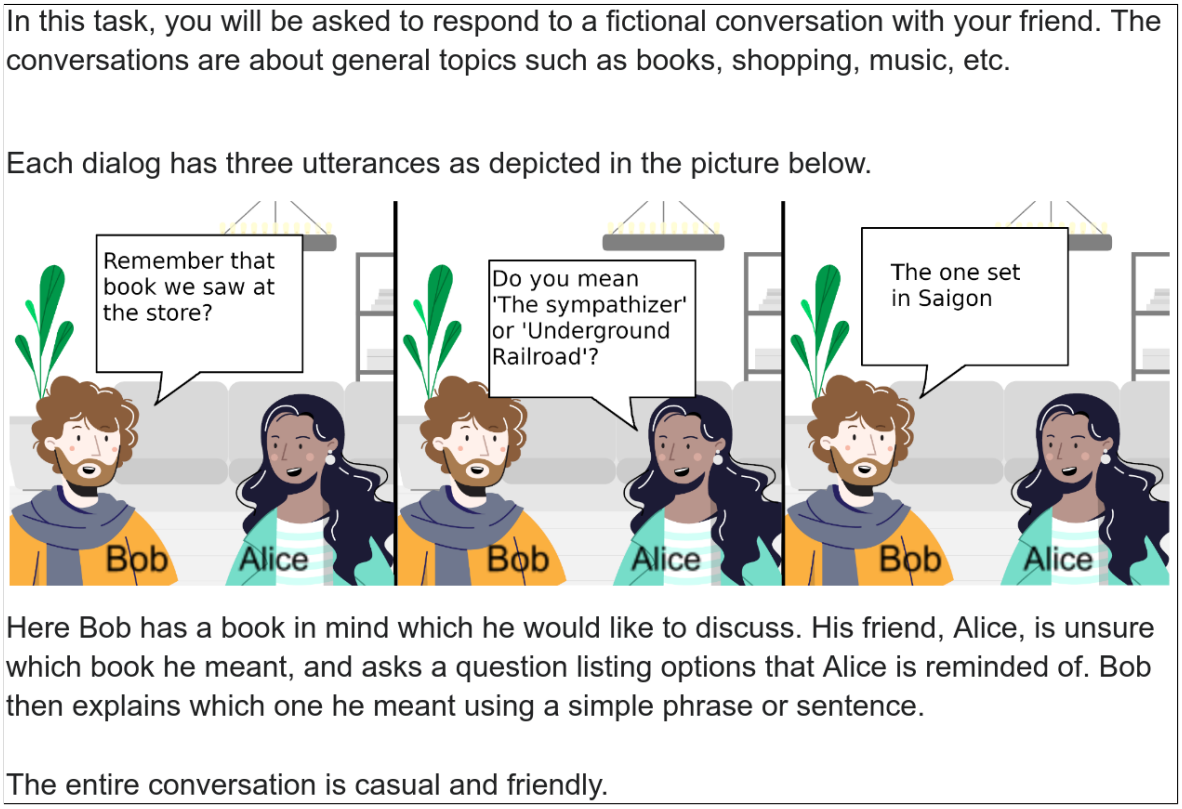

Figure 4: The first instruction screen shown for the BOOKS domain. It summarizes the task based on a cartoon completion setup.

BOOKS
"Remember that book we saw at the store?"
"Hey, about that book I lent you last month..."
"Can you get me that book on the first shelf?"
"I really liked that book from the reading club..."
"That book I got was super interesting!"
MUSIC
"So that song I keep singing..."
"One of those cool songs that Bob sang last night..."
"You sang that song really well yesterday..."
"Could you play that song from your playlist?"
"I'll now play my favorite song."
RECIPES
"Remember that fabulous stuff from Tom's party?"
"That recipe on today's Masterchef was too good!"
"Going to make that dish from Mary's potluck."
"Our favorite food blogger had a cool episode this week!"
"Does mom's cookbook have that recipe?"

Table 11: The manual utterances which are used to populate the first cell of the cartoon.

Your task is to assume the role of **Bob** in the cartoon. You have a book in mind and wish to indicate that one to Alice.

You will be asked to fill in the last bubble in the conversation.

Under the figure, you will see a description of each of the choices. These descriptions indicate *common knowledge* that both you and your friend have on the topic. Please read both of the descriptions carefully. You will use the information at a high level to refer to the book. Alice will understand which one you mean if you use this information.

In the following screen, you will be asked to refer to one of them.

The Sympathizer	The Underground Railroad
• The Sympathizer is the 2015 debut novel by Vietnamese American professor Viet Thanh Nguyen. It is a best-selling novel and recipient of the 2016 Pulitzer Prize for Fiction. • Set as the flashback in a coerced confession of a political prisoner, the book tells the story of the South Vietnamese Government in 1975 and subsequent events in American exile in Los Angeles, through the eyes of a half-Vietnamese, half-French undercover communist agent.	• The Underground Railroad is a historical fiction novel by American author Colson Whitehead, published by Doubleday in 2016. • The alternate history novel tells the story of Cora and Caesar, two slaves in the antebellum South during the 19th century, who make a bid for freedom from their Georgia plantation by following the Underground Railroad, which the novel depicts as a rail transport system with safe houses and secret routes

Figure 5: The second instruction screen shown for the BOOKS domain. It provides further information about the task and describes where the annotators should acquire the knowledge to perform the annotations.

You will then be asked to indicate one book ***without using the name of the book***.

Pick this one

⇩

The Sympathizer	The Underground Railroad

We would like you to give us **at least 3 expressions** (max 5) for the chosen book to fill your speech bubble. For example:

Enter at least 3 ways to refer to the book:

- The book about the Vietnam war
- The one set in the 70s
- Not the one about slavery
- Which won an important award
- The one published earlier.

We would like you to recall from your memory / understanding as much as possible, but Don't worry if you don't remember something exactly. Write the answer based on what you can remember. Also note the following dos and don'ts:

Do	Don't
✓ **Keep it casual and conversational** ✓ **Varied, interesting, and creative expressions** ✓ **Use alternative words eg. 'award' instead of 'prize'.** ✓ **Vary the phrasing: 'the book about', 'I meant the', 'was thinking of', 'the one about', 'not the one about', 'I didn't mean', 'I wasn't referring to', etc.**	✗ **Mention the book by name or position (eg. the second one)** ✗ **Use too detailed information that Alice may not recall (eg. '1992' or 'in the 90s' are better choices than 'Sep 9 1992')** ✗ **Copy whole sentences from the description**

Figure 6: The third instruction screen shown for the BOOKS domain. It shows which item should be referred to, and lists five examples of appropriate REs. It also lists a number of actions that the annotators should or should not do.

Your task is to assume the role of **Bob** in the cartoon. You have some food in mind and wish to indicate that one to Alice.

You will be asked to fill in the last bubble in the conversation.

Under the cartoon, there is a description to find out more about each of the choices.

Please read about both options and examine the images if provided. Then use the information at a high level to refer to the recipe.

In the following screen, you will be asked to refer to one of them.

Simnel Cake	Pandan Cake
Simnel cake is a fruitcake widely eaten in the United Kingdom, Ireland and other countries with patterns of migration from them, associated with Lent and Easter. It is distinguished by layers of almond paste or marzipan, and a set of eleven balls made of the same paste. 	Pandan cake is a light, fluffy, green-coloured sponge cake flavoured with the juices of Pandanus amaryllifolius leaves.The cake is popular in Indonesia, Malaysia, and also the Netherlands, especially among the Indo community, due to its historical colonial ties with Indonesia. 

Figure 7: The second instruction screen shown for the RECIPES domain. It provides further information about the task and describes where the annotators should acquire the knowledge to perform the annotations.

You will then be asked to indicate one ***without using the name of the food***.

Pick this one ⇩

Simnel Cake	Pandan Cake

We would like you to give us **3 to 5 expressions** for the chosen food to fill in your speech bubble. Make sure your expressions **refer only to the food pointed out to you and not to the other one**. Try to recall from memory and it is ok if you don't remember details exactly. For example:

- It looks surprisingly green in color
- Without any frosting or fruit
- It is made from some leaf
- Comes from Indonesia
- Isn't the Easter one

Think of how **you** might describe the choice to a friend.

Do	Don't
✓ **Keep it casual, creative and varied** ✓ **Read facts about the food and note any images provided.** ✓ **Vary the wording: 'the one with', 'I meant the', 'was thinking of', 'not the one made with', 'I didn't mean', 'is often done by', etc.**	✗ **Mention the food by name or position (eg. the second one).** ✗ **Use overly detailed information (eg. 'made from Pandanus amaryllifolius leaves').** ✗ **Use expressions which are not about the food. Eg. 'on a white plate', 'decorated with flowers'.** ✗ **Copy whole sentences.**

Figure 8: The third instruction screen shown for the RECIPES domain. It shows which item should be referred to, and lists five examples of appropriate REs. It also lists a number of actions that the annotators should or should not do.

Your task is to assume the role of **Bob** in the cartoon. You have a song in mind and wish to indicate that one to Alice.

You will be asked to fill in the last bubble in the conversation.

Under the cartoon, there is a link to find out more about each of the choices. It will take you to search results that let you listen to each song and read basic facts about it.

Please listen to at least some of each song and read about both options. Then use the information at a high level to refer to the song.

In the following screen, you will be asked to refer to one of them.

Easy on Me (by Adele)	**I Gotta Feeling (by The Black Eyed Peas)**
• Click here to find out about the song.	• Click here to find out about the song

Figure 9: The second instruction screen shown for the MUSIC domain. It provides further information about the task and describes where the annotators should acquire the knowledge to perform the annotations.

You will then be asked to indicate one ***without using the name of the song or artist***.

Pick this one

⇩

Easy on Me (by Adele)	**I Gotta Feeling (by the Black Eyed Peas)**

We would like you to give us **3 to 5 expressions** for the chosen song to fill in your speech bubble. For example:

- The one with the piano music
- The song that's not energetic
- It has something about a river
- The newer one
- It's about not having time to choose

Think of how **you** might describe the choice to a friend.

Do	**Don't**
✓ Keep it casual, creative and varied **✓ Listen to the song, read facts about it, and note how it makes you feel.** **✓ Use what you may already know about the song** **✓ Vary the wording: 'the song about', 'I meant the', 'was thinking of', 'the one about', 'not the one about', 'I didn't mean', 'I wasn't referring to', etc.**	**✗ Mention the song by name, artist or position (eg. the second one)** **✗ Use overly detailed information (eg. '5th single on the album' or 'sold 4 million copies' is too specific. 'From the 90s' is better than '9 July 1992')** **✗ Copy whole sentences from other places**

Figure 10: The third instruction screen shown for the MUSIC domain. It shows which item should be referred to, and lists five examples of appropriate REs. It also lists a number of actions that the annotators should or should not do.

Filter	Rationale
Articles with more than one infobox	Items should focus on a single topic. For example, we do not accept a movie that has a recorded song for the MUSIC domain.
Items with a selected section length ≤ 250 characters[12]	Items have enough information in the section selected to show as background knowledge to the annotators.
Books or music items that do not have genres in their infobox	Items contain important attributes for the domain
Recipes that are not a prepared food or without images (§3.3)	Items contain important attributes for the domain
Items in the MUSIC domain with ≤ 14 sections	Song should be popular to enable the annotators to also use their own background knowledge.
Items containing words on a denylist	Avoid sensitive or inappropriate items.

Table 12: List of filters applied to select candidate items from those extracted from Wikipedia articles. For each filter, we show the rationale behind it.

		BOOKS				RECIPES				MUSIC		
		ORAC	NAME	INBO	UNBA	ORAC	NAME	INBO	UNBA	NAME	INBO	UNBA
BERT	lr	3e-5	1e-5	5e-6	1e-5	5e-6	5e-7	1e-5	3e-5	1e-5	3e-6	5e-6
	bsz	16	16	32	16	16	16	32	64	64	64	32
	epochs	5	10	3	3	3	1	3	1	1	3	3
BERT Multiple Choice	lr	3e-5	5e-6	3e-5	3e-5	3e-5	1e-6	3e-5	3e-5	5e-6	1e-5	5e-6
	bsz	64	32	32	64	64	32	64	64	64	32	32
	epochs	3	3	1	1	1	1	1	1	1	1	3
T5	lr	5e-6	3e-5	3e-6	3e-6	3e-6	3e-6	3e-6	3e-6	3e-6	3e-6	3e-6
	bsz	64	32	64	64	32	32	16	64	64	64	32

Table 13: Selected hyper-parameters for the supervised models for each domain and input type. We list selected values for base learning rate (lr), Training batch size (bsz), Num training epochs (epochs).

	BOOKS				RECIPES				MUSIC			
	ORAC	NAME	INBO	UNBA	ORAC	NAME	INBO	UNBA	NAME	INBO	UNBA	AVG
USE	66.06	55.15	59.12	58.41	70.77	52.48	64.98	66.36	57.53	60.71	60.57	61.10
Entailment	85.00	50.91	63.16	70.54	81.31	**56.73**	69.41	75.58	52.68	62.42	74.32	67.46
BERT	94.34	59.58	78.27	81.91	88.87	53.99	76.15	81.07	**60.57**	63.35	74.50	73.87
BERT Joint	95.00	**61.85**	77.31	82.47	89.58	56.60	76.86	81.21	59.79	68.07	76.17	74.99
T5	**95.91**	61.04	**78.98**	**84.13**	**93.22**	56.69	**82.80**	**85.77**	59.14	**72.33**	**82.97**	**77.54**

Table 14: Indirect reference resolution development set results for different models on all domains and input types: ORACLE (ORAC), NAME, INFOBOX (INBO), UNSHOWN BACKGROUND (UNBA). The best result of each column is boldfaced.

Category	Example 1	Example 2	Percentage
Date	was released in 2012	the song that's only a few years old	25%
Content	Singer compared his new life and the old.	Not the sad song	24%
Singer or band	The one by a male singer	song is by an Irish rock band	19%
Genre	It is the song that is R&B.	it's that baroque pop ballad track	13%
Further song info	Was remixed in the late 80s	The one sampled from Shirly Bassey	10%
Comparison	The newer one	Released later	10%
Negation	Not the song about greed	No not the one with Rap	10%
Instrument or sound	It is a midtempo R&B ballad	not the one with the piano intro	7%
Album	One from their second album	The one from the album Wordshaker	5%

Table 15: Categories of correctly classified REs in the MUSIC domain. The results are based on the T5 model with the UNSHOWN BACKGROUND input.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
8

☑ A2. Did you discuss any potential risks of your work?
Section 9, as part of the Ethics Statement.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3, 4, 5, and 6

☑ B1. Did you cite the creators of artifacts you used?
3, 4, 5, and 6

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
4

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 3.4.1: "To remove any sensitive or offensive content, we also filter articles whose content matches a list of sensitive words." In addition, we did not ask the raters for any Personally Identifiable Information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sections 3 and 4.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Sections 4 and 6.

C ☑ Did you run computational experiments?

6

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 6 and Appendix D.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 6 and Appendix D.

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We did not observe meaningful differences when running the experiments multiple times in the preliminary experiments. We therefore reported the results of only one run.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3, 5, and 6. We cited LMs such as BERT and T5.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Sections 3 and 4.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 3 and Appendix A and B.

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
In section 4, we mention: "We used a pool of around 60 in-house crowdworkers. They were all native English speakers recruited from U.S., U.K., Canada, and Australia."
This work was carried out by participants who are paid contractors. Those contractors received a standard contracted wage, which complies with living wage laws in their country of employment. Due to global privacy concerns, we cannot include more details about our participants, e.g., estimated hourly wage or total amount spent on compensation.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
This was discussed with the annotators before data collection.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
9

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
4