Social Norms Guide Reference Resolution

Mitchell Abrams and **Matthias Scheutz**
Human-Robot Interaction Laboratory, Tufts University, Medford, MA 02155
{mitchell.abrams, matthias.scheutz}@tufts.edu

Abstract

Humans use natural language, vision, and context to resolve referents in their environment. While some situated reference resolution is trivial, ambiguous cases arise when the language is underspecified or there are multiple candidate referents. This study investigates how pragmatic modulators external to the linguistic content are critical for the correct interpretation of referents in these scenarios. In particular, we demonstrate in a human subjects experiment how the social norms applicable in the given context influence the interpretation of referring expressions. Additionally, we highlight how current coreference tools in natural language processing fail to handle these ambiguous cases. We also briefly discuss the implications of this work for assistive robots which will routinely need to resolve referents in their environment.

1 Introduction

Humans interacting in natural language need to resolve referential expressions often referring to referents in their environment; utterances like *pick up the green box* or *pick it up*, for instance, highlight some referring expressions that point to a referent. These expressions appear in various forms, from clear and specific—*the green box*—to underspecified and ambiguous—*it*. But reference resolution, especially situated reference resolution, also requires vision and pragmatic context to disambiguate references. In the linguistically underspecifed example of *pick it up*, a listener may have to look for the candidate objects in the environment to figure out what *it* refers to. Additionally, the social setting can modulate what referent is intended, given the same referring expression and objects in the environment; in a dining room, for instance, a spoon on the ground may be the more likely candidate than a pencil. In this paper we investigate the role of pragmatic modulators like this in reference resolution.

The psycholinguistics literature has leveraged eye tracking to infer what referents humans resolve in various contexts (Tanenhaus et al., 1995; Spivey et al., 2001). Pragmatic modulators outside of the linguistic content can further constrain the referential domain and affect referent interpretation, such as task-relevant constraints (Hanna and Tanenhaus, 2004). There is a gap, however, in understanding how other pragmatic modulators, such as social norms and conventions affect the interpretation of referents. For example, while there has been work on modeling what social norms are activated in various contexts and settings (Malle et al., 2020), it is unclear how norms guide humans to interpret referring expressions. Similarly, conventions such as standing on the right side of an escalator while walking on the left, or sitting in the back of cab, can have modulatory influence on reference resolution and object selection.

The aim of this paper is to demonstrate the role of pragmatic modulators, especially social norms, in guiding situated reference resolution. First we provide background on reference resolution and context, with a focus on situated reference resolution in particular. Then, we show how referents are guided by social norms in certain contexts through a human-subjects experiment. We proceed to compare results from this experiment—the referents selected given the situational context and referring expressions—against several coreference tools that attempt to resolve these referents. Lastly, with an eye towards assistive robots, we conclude by outlining an approach for teaching robots to leverage social norms and context for object selection.

2 Background

2.1 Reference Resolution in NLP

Reference resolution is a key task in natural language processing. State-of-the-art approaches in NLP—through significant strides in deep learning—

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1 - 11

perform well on text-based reference resolution by learning important syntactic and semantic features. Indeed, coreference phenomenon are naturally guided by several linguistic phenomena as discussed in (Jurafsky and Martin, 2009). Among them are gender agreement, number agreement, person agreement, recency, binding constraints, verb semantics, and selectional restrictions—features often useful for coreference models in NLP such as CoreNLP (Finkel et al., 2005). A more recent end-to-end neural model, part of AllenNLP (Lee et al., 2017), moves away from traditional engineered features and syntactic information and instead relies on word embeddings within and around potential coreferent mention spans as well as the distance between spans, among other approaches.

However, while powerful models encode important linguistic cues for reference interpretation, and use word embeddings to capture word similarity, they fail to take into account contextual knowledge (Emami et al., 2018). This renders current NLP tools insufficient for situated reference resolution.

Recently, coreference tasks such as the Winograd Schema Challenge (WSC), proposed by Levesque et al. (2012), challenge coreference models to handle world knowledge and common sense reasoning. The KnowRef dataset (Emami et al., 2018), a coreference corpus of natural texts, provides a new benchmark for coreference resolution that requires systems to reason about context. The coreference task created sentences stripped of linguistic cues from syntax, gender agreement, and number agreement, forcing systems to rely on context and world knowledge. Emami et al. (2018) fine-tuned a BERT model on the KnowRef dataset to improve its accuracy over other state-of-the-art models. This shows that reference resolution systems can encode world knowledge and common sense reasoning to an extent when trained on these Winograd Schema type datastets. Yet these powerful models remain opaque and do not explicitly model the pragmatic constraints of social norms and conventions.

3 Pragmatic Constraints and Social Norms

Work on multi-modal reference resolution gets closer to modeling pragmatic constraints, mainly by moving beyond text and considering gesture and context to help disambiguate referring expressions (Matuszek et al., 2014; Whitney et al., 2016; Chai et al., 2004). Whitney et al. (2016), in addition to speech and gesture, incorporates contextual knowledge to improve the accuracy of their model on a dataset where people refer to objects on a table. The model exploits information from the kitchen domain and uses recipes as a knowledge base to understand tools and ingredients that typically belong together. Chai et al. (2004) also uses domain knowledge in a graph-matching algorithm for multi-modal referring expressions with a map showing houses and prices. The guiding context, here, is conversational history and domain knowledge about house pricing.

Within the psycholinguistics literature, Hanna and Tanenhaus (2004) use eye-tracking in a cooking simulation to show that pragmatic constraints have modulatory influence on the interpretation of referring expressions. In this experiment, participants followed a confederate cook's instructions for a recipe, where the cook used the the definite noun phrase *the cake mix* to signal potential referents in the cooking space. The addressee's domain of interpretation changed with the task-based constraints—cued perceptually with the cook's hands being empty or full. As the addressees monitor the speaker, they tend to interpret the referent in the cook's area when the cook's hands were full and the referent in their own area when the cook's hands were empty. The results support constraint-based models, where speaker constraints are taken into account for interpretation alongside linguistic ones; indeed, this study highlights how a definite referring expression can point to a few possible candidate objects in a restricted domain, just based on its linguistic form, yet people can disambiguate which referent is being referred to from the pragmatic context. While this study focuses on speaker-based constraints, there is still a lack of knowledge about the modualtory influence of social norms and conventions in interpreting referring expressions.

A promising step in this direction are attempts to computationally model social norms with the ultimate aim of creating norm competent artificial agents (Malle et al., 2020). Malle et al. (2020) experimentally collected responses from humans to generate social norms for eight contexts, including a library, boardroom, bathroom, and restaurant, among others. While social norms can be elusive and challenging to define, since they vary by cul-

ture and appear on various levels of demand, Malle et al. (2020) follows Janoff-Bulman et al. (2009) in viewing social norms as *prescriptions* and *prohibitions* and giving attention the gradability of these norms by mapping the deontic force—how strong or weak these norms are to be followed—to the collected *prescriptions* and *prohibitions*. Malle et al. (2020) define norms more formally as such:

A norm is an instruction, in a given community, to (not) perform an action in a given context, provided that a sufficient number of individuals in the community (i) demand, to a certain degree, of each other to follow the instruction and (ii) do in fact follow it.

We will adopt this definition for *social norms* in this study, which formalizes the idea of *prescriptions* and *prohibitions* being followed by many people. We also broaden the definition of *social norms* to include descriptive norms and conventions, although Bicchieri (2005) makes a more fine-grained distinction between *social norms*, *conventions*, and *descriptive norms*. We do not consider moral norms or legal obligations in the present paper.

Equally important, this study offers an approach for teaching norms to robots for guiding actions and balancing norms with goals. They outline an enriched Markvov Decision Process (MDP) approach that uses a starting norm base, which are predefined norms collected in the experiment, and refines it through human interaction and feedback. We look at this proposal optimistically for, in a similar vein, teaching embodied agents the specific behavior of performing situated reference resolution.

A referring expression can appear in a variety of linguistic forms, but pragmatics, regardless of the linguistic form, has the potential to modulate the meaning of the sentence and referent entirely. This will be true for humans that use natural language with robots as well. Imagine a situation where someone commands an assistive robot in a home: *take it away*. The robot can use the natural language and vision input to scan the area for potential referents of *it*. If a shoe and a spoon are salient objects on a dining room table, the convention of a shoe not belonging on a table would make the shoe the more likely candidate. Alternatively if a shoe and a spoon are salient objects on the floor of a bedroom, the spoon would likely be the referent.

Marrying the work on pragmatic constraints on reference resolution and social norms, we conduct an experiment where humans are tasked with identifying the referent of an ambiguous referring expression across various contexts and, thus, various social norms and conventions.

This experiment relates to previous work that leverages crowd sourcing for collecting anaphora annotations and judgments (Poesio et al., 2019, 2013; Chamberlain et al., 2008; Kicikoglu et al., 2019). Although this body of work focuses on a game-with-a-purpose (GWAP) approach to crowd sourcing (Von Ahn, 2006), our study does not gamify our annotation task but it does avoid using linguistic and annotation terminology for participants. Poesio et al. (2019), specifically, collects several judgements and disagreements over ambiguous cases of anaphora. Similarly, our study captures the reasons and explanations people make when resolving a referent, although there is no adjudication process. Through overall decisions and explanations, we are then able to study the agreement and interpretations over our ambiguous scenarios.

4 Experiment and Results

Here, we report the details of an online vignette-based human-subjects experiment designed to explicate the potential role of social norms in resolving references in context. We recruited 50 participants on Prolific (see `https://prolific.co`), an online participant recruitment site which is known to provide better results on tasks like ours. Participants were free to leave the study at any point, their data was anonymized, and they received adequate payment for the study. A consent form was presented at the beginning of each study with information on how their data would be used. We restricted our recruitment to people living in the United States and at the time of the study and native speakers of English.

Each participant was presented with eight text vignettes where each vignette described a scene within a daily-life context. Each scene contains four pieces of information: an explicit mention of the setting (*The scene takes place in a library*), a description of the background—that is, the objects and people in the scene—an underspecified referring expression (e.g., *remove it*), and an actor in the scene acting on an object. There are always two salient referents that are potential candidates for the referring expression. Participants must determine whether the referent chosen by the actor in

Contexts	Items	Hypothesized Norms
Library	head seat, side seat	*do not interrupt someone at the library; give space to others*
Boardroom	head seat, side seat	*do not sit at the head of the table*
Taxicab	front seat, back seat	*you should sit in the front seat*
Friend's Car	front seat, back seat	*you should sit in the back seat*
Dining Room	shoe, spoon	*a shoe should not be on a dining table*
Leather Shop	shoe, spoon	*a spoon should not be on a non dining table*
Bookstore	magazine, toothbrush	*a toothbrush should not be on the floor of a bathroom*
Bathroom	magazine, toothbrush	*a magazine for display should not be on the ground*

Table 1: Overview of contexts in the experiment, the items mentioned in the reference task, and some hypothesized norms activated in each context.

the scene was the *correct* one.

With the information still in view, participants are asked to select the best explanation for their answer and are provided a multiple choice listing of five potential explanations and one open text response option labeled *other*. These reasons include: *typical for the setting*, *object is mentioned first*, *object is mentioned more recently*, *time sensitive option*, *more convenient option*, and *other*. We included reasons that could explain that the correct referent was the one that was intended, rather than subjective options that potentially frame the question as a personal preference. We also offered a text response if none of the options fit.[1] We summarize the contexts, candidate objects, and a hypothesized norm associated with each context in Table 1.

Each context, some of which are inspired by Malle et al. (2020), are assumed to activate their own inventory of norms to help disambiguate the referring expression. Our hypothesized norms are partly based on intuition but also inspired by previous work on norms and behavior. Aarts and Dijksterhuis (2003), for instance, conducted a survey with undergraduates to confirm the normative behavior of acting silently in a library setting. This norm is applied to our study in a library scene: there is an open seat at a table right next to someone and a seat further away from someone. Although there is no mention to noise, seating right next to someone else—a stranger—is potentially noisy and interruptive. Additionally, similar to the norm of not littering (Cialdini et al., 1991), we focused on *prohibitions* of objects not belonging in certain contexts; a shoe is not supposed to be on a clean dining room table and a toothbrush should not be on the bathroom floor. We posit other norms that tend to influence frequent behavior such as sitting with your friend in their car, as opposed to the backseat, and sitting in the back of a taxicab.

Similar to Winograd schema datasets, each referring expression is stripped of linguistic surface cues such as gender, number, and person that would give away the referent. Instead, these scenes are set up so that subjects in the experiment have to rely on information outside of the text to help them make a decision. The only linguistic cue we maintain in the study, however, is *recency*, where we change the ordering of the referents. These scenes include: library, boardroom, taxicab, friend's car, dining room, leather shop, bookstore, home bathroom.

Each scene has a complementary scene that shares the same referents; the library and boardroom share two seats, the taxicab and friend's car share two seats, the dining room and leather shop share a shoe and a spoon, and the bookstore and home bathroom share a toothbrush and a magazine. The purpose of creating complementary scenes with the same referent was to demonstrate how, when the referring expression is constant, the context, and thus the social norms and conventions associated with it, modulate the interpretation of the referent. We posit, for instance, that people will select the seat in the back of the cab as opposed to the front of the cab. This would be guided by the norm of sitting in the back of a cab. Alternatively, people would most likely choose to sit in the front seat in a friend's car and not the back seat, also for conventional reasons.

The following excerpts show examples of the scenes participants read during the experiment. The

[1]The experimental design of a posthoc explanation of whether the referent was "correct" was chosen after initial pilot experiments showed that asking subjects for the correct referent rather than providing them with the choice of the actor in the scene led to a confound: subjects often choose the referent they would have chosen instead of hypothesizing the referent the actor in the given context would have selected.

one below is for the dining room context:

> *The scene takes place in a dining room. There is a shoe and a spoon sitting on a dining room table. Dinner is about to be served.*
>
> *Someone says, "remove it."*
>
> *Someone else removes the shoe from the table.*

This next example, describing a leather shop, shows a complementary scene using the same candidate objects of the shoe and spoon and the same definite referring expression, *remove it*.

> *The scene takes place in a leather shop. There is a spoon and a shoe sitting on a worktable. Nothing else is on the table. A customer is coming into the store.*
>
> *Pointing to the worktable, someone in the room says, "remove it."*
>
> *The employee removes the spoon.*

While each participant sees all eight scenes, there are two conditions where the ordering of the referents mentioned in the text are flipped. Condition A lists the intended (correct) referent last and condition B lists the intended referent first. In the dining room scene, for instance, condition A lists the spoon first and then the shoe and condition B lists the shoe first and then the spoon. We create these conditions to test whether people are biased by the recency of referents and to also evaluate these texts on coreference tools which may be biased by recency in performing reference resolution.

The results in Table 2 show how many people agreed that the selected referent was correct or incorrect in a "yes-no" question. Overall, the majority of people agreed that the referent selected was the *correct* one across all scenes, and the frequency distributions seem consistent across both conditions. Stronger agreement trends towards scenes with seats as referents—that is, the library, boardroom, taxicab, and friend's car. The scenes with the most disagreement were the bookstore and home bathroom scenes, which used a toothbrush and magazine as candidate objects. For these scenes, we hypothesized contexts with a *prohibition* type norm where it is unacceptable for a toothbrush and a magazine to be on the ground. But these were, perhaps, less airtight scenarios. In a home bathroom, magazines can be stowed in the corner for casual reading,

	A		**B**		
Contexts	**yes**	**no**	**yes**	**no**	**Total**
Library	21	3	22	4	43-7
Boardroom	23	1	26	0	49-1
Taxicab	24	0	26	0	50-0
Friend's Car	23	1	25	1	48-2
Dining Room	23	1	25	1	48-2
Leather Shop	20	4	21	5	41-9
Bookstore	16	8	20	6	36-14
Bathroom	23	1	23	3	46-4

Table 2: Counts for *yes* or *no* in response asking whether the referent identified is *correct*. Results reported for conditions A and B where each condition is a different ordering of the referents mentioned in the text (e.g. ... *shoe and spoon* ... v.s. ... *spoon and shoe* ...)

but a toothbrush has a its place in a cabinet or cup holder. In a bookstore, a magazine should belong on the shelf along with other books and magazines, but a stray toothbrush in a public space can be left alone, unless a norm of not littering is competing. The explanations people chose offer some more clarity to this picture.

Table 3 provides an overview of the reasons people gave for their agreement or disagreement with the selected referent. One obvious trend that stands out is that **Convention** (displayed as *typical for the setting* in the study) outnumbers the other reasons across all scenes and conditions. Where there was high consensus on the correct referent for the seat related scenes, there was a commensurate high rate of selecting the conventional explanation. For the library scene, however, we see a tension between conventional explanation and a convenient choice for choosing a seat at the head of the table rather than a seat next to someone else. For some, it seems, the convention of keeping distance from a stranger at a library, as not to cause a disruption, is either not activated or is overruled by convenience.

Then there are the bookstore and home bathroom scene that have a lower consensus and, thus, a higher count of alternative explanations. Interestingly, when people disagree that the selected object was correct, their explanations suggest a normative reason is stronger in the other direction. For example, if a magazine is more conventional in a bookstore (*prescription*) a toothbrush is unconventional and suggests a prohibition norm. We present a sample of explanations for these scenes:

		Convention	Last	First	Time Sensitive	Convenient	Other
Library	A	10	2	0	0	9	3
Library	B	10	0	0	0	9	7
Boardroom	A	20	0	0	0	2	2
Boardroom	B	23	0	1	0	1	1
Taxicab	A	23	0	0	0	0	1
Taxicab	B	25	0	1	0	0	0
Friend's Car	A	19	0	0	0	5	0
Friend's Car	B	17	0	1	0	8	0
Dining Room	A	16	0	1	0	1	6
Dining Room	B	22	0	0	0	1	3
Leather Shop	A	12	1	0	1	1	9
Leather Shop	B	17	0	1	1	4	3
Bookstore	A	13	1	2	0	1	7
Bookstore	B	16	1	3	1	0	5
Bathroom	A	16	0	1	3	1	3
Bathroom	B	11	1	5	3	0	6

Table 3: Counts for best explanation for correct or incorrect referent selected. The shortened label **Convention** corresponds to the *typical for the setting* option in the experiment; **Last** to *object is mentioned more recently*; **First** to *object is mentioned first*; **Other** to *other* with free text response; **Time Sensitive** to *time sensitive option*; **Convenient** to *more convenient option*

Bookstore:
toothbrush doesn't belong...
the toothbrush is the more out-of-place object, and therefore, it is implied to have that removed rather than the magazine
the toothbrush does not match the setting/misplaced

Home Bathroom:
object is irrelevant to the setting and should be removed

We also note that for these scenes and others in the study, some of the explanations people articulate can be classified as norms even though they did not select the normative option in the multiple choice:

Home Bathroom:
The toothbrush should not be on the floor toothbrush does not belong on the floor

Boardroom:
The boss usually sits at the head of the table.

Library:
The head of the table doesn't have anyone sitting next to it.

Although it was unclear for some that *typical for the setting* subsumed the normative or conventional explanations, the fact that people gave normative explanations support that reasoning even more.

To summarize this experiment, given the same two referents, people interpreted one referent as correct in one context and the other as correct in another context, each according to specific norms that are activated in that context. This suggests that social norms activated by the context had enough modulatory influence to determine the interpretation of an ambiguous referring expression favored by the norm. As a consequence, not knowing the norms that apply in these context will likely lead to incorrect interpretations of referential expressions as other factors not necessarily congruent with the norm-based interpretation will be used for reference resolution, as the next section on current NLP tools will demonstrate.

4.1 Evaluating NLP Tools

To complement our empirical study, we evaluated several coreference and natural language processing tools on our experimental scenes to determine if they achieve human performance for norm-guided reference resultion tasks. These include Neural-

NLP Tool	Context	C	Answer
NeuralCoref	Dining Room	A	coordination ✗
	Dining Room	B	coordination ✗
	Leather Shop	A	[the worktable] ✗
	Leather Shop	B	[the worktable] ✗
	Bookstore	A	coordination ✗
	Bookstore	B	coordination ✗
	Bathroom	A	non-referential ✗
	Bathroom	B	non-referential ✗
CoreNLP	Dining Room	A	[dinner] ✗
	Dining Room	B	[a shoe] ✓
	Leather Shop	A	[the room] ✗
	Leather Shop	B	[the room] ✗
	Bookstore	A	[a toothbrush] ✗
	Bookstore	B	[a magazine] ✓
	Bathroom	A	[a magazine] ✗
	Bathroom	B	[a toothbrush] ✓
AllenNLP	Dining Room	A	coordination ✗
	Dining Room	B	[a shoe] ✓
	Leather Shop	A	coordination ✗
	Leather Shop	B	coordination ✗
	Bookstore	A	coordination ✗
	Bookstore	B	coordination ✗
	Bathroom	A	coordination ✗
	Bathroom	B	coordination ✗
GPT-3: Curie	Dining Room	A	coordination ✗
	Dining Room	B	coordination ✗
	Leather Shop	A	coordination ✗
	Leather Shop	B	coordination ✗
	Bookstore	A	non-referential ✗
	Bookstore	B	coordination ✗
	Bathroom	A	coordination ✗
	Bathroom	B	coordination ✗
GPT-3: Davinci	Dining Room	A	coordination ✗
	Dining Room	B	[the spoon] ✗
	Leather Shop	A	coordination ✗
	Leather Shop	B	coordination ✗
	Bookstore	A	[toothbrush] ✗
	Bookstore	B	[toothbrush] ✗
	Bathroom	A	[toothbrush] ✓
	Bathroom	B	[toothbrush] ✓

Table 4: Evaluation of coreference tools on contexts that use a definite reference. The dining room scene and leather shop scene both use the referring expression *remove it*; the bookstore scene and home bathroom scene, similarly, use the referring expression *pick it up*. We report whether these tools can detect if *it* is referential and refers to the correct object.

Coref, an extension of SpaCy (Honnibal and Johnson, 2015) and based on (Clark and Manning, 2016), Stanford CoreNLP (Finkel et al., 2005), AllenNLP (Lee et al., 2017). We also evaluated the GPT-3 base models, Davinci and Curie, by OpenAI, (Brown et al., 2020) designed for text generation and question-answering tasks. For this experiment, we specifically focus on the scenes that use a definite referring expression such as *pick it up*, in the bookstore and home bathroom scenes, and *remove it*, in the dining room and leather shop scenes. For the GPT-3 models, we prompt the Davinci model with the phrase *which one* but do not prompt the Curie model with a question. We made this decision to probe the different capabilities of these models; for the Curie model, we chose not to prompt it to see if would coherently generate the rest of the text and the resolve the correct referent likely to follow from the referring expression in the command (e.g. *pick it up*, *remove it*); for the Davinci model, we tested the question-answering capabilities by providing a question. This evaluation did not use a similar question as the main study—*Was this the correct object?*— as understanding a yes-no response is more opaque and we wanted to see if it could return the referents. We also note that all of these models were used off the shelf without fine tuning.

Results are summarized in Table 4 where we list the referents that the tools selected to match the referring expression. We represent an entity in brackets and also note when the referent is the coordination of the two referents (e.g. *a shoe and a spoon*) or the referring expression was interpreted as non-referential. A check mark ✓denotes the correct referent and an ✗denotes the incorrect referent.

The experimental conditions have a role, here, since the recency or distance of the referent serves as a traditional feature for coreference models in NLP. Swapping the ordering of the referents ensures that if a model resolves the correct referent, it is consistent and is more likely taking into account the context then the surface structure. This swapping method is similar to (Emami et al., 2018)'s evaluation of BERT on the KnowRef test set for consistency.

In the results, there are only three cases where the coreference models choose the right referent. CoreNLP and AllenNLP both correctly link [a shoe] and *it* in the dining room scene. For this condition (condition B), shoe is mentioned first in

the text. Although, once the objects are switched, the models choose the wrong object—CoreNLP selects [dinner] (none of the candidate referents) and AllenNLP selects both the shoe and the spoon in a coordination. CoreNLP seems to do the best by selecting another correct referent: [a magazine] in the bookstore scene. But it fails yet again once the objects are swapped.

The GPT-3 models perform poorly overall but the Davinci model, prompted by *which one*, gets closer to the right answer by picking out individual referents more often than the Curie model. Davinci is consistently incorrect in the bookstore scene but consistently correct in the bathroom scene, yielding the only correct result when the referents are switched. The correct referent selected in the bathroom, the toothbrush, was also selected in the bookstore for both conditions. This suggests that the model is biased towards picking the toothbrush over the magazine more generally.

For most of the tools doing reference resolution on these scenarios, we see a theme of referring back to the coordination of the two referents, when only one referent should be selected. Therefore, it is clear these results do not match human intuition for this specific reference resolution task and, more importantly, fail to understanding social norms in order to consistently infer the correct referent.

5 Discussion & Future Implementations

The coreference task performed by humans and the NLP tools show a striking difference in outcomes. Given the same context and text, people tended to agree on the correct referent. Since the examples were stripped of linguistic cues that would give away the referent, people relied on context and social norms based on the reasons they selected and the written explanations they provided. Notably, however, the inconsistent agreement across all scenes can be reconciled with the fact that social norms and conventions are not equally shared across all people. This is supported, in part, by the written responses too. Additionally, these results also suggest that not every norm is weighed the same; the deontic force—how strongly the norm is to be followed—potentially influences how the norm guides a behavior or interpretation and competes with other norms. Admittedly, a limitation of our study is that we do not explicitly categorize our hypothesized social norms and conventions in a gradable fashion, but future work will consider deontic force for a more fine-grain understanding of social norms.

NLP tools, on the other end, tell a different story. Many of the tools specifically designed for coreference resolution failed to consistently select the correct referent. The more powerful NLP engines, such as GPT-3 model, also performed poorly. This shows that relying on such a system to resolve references in these contexts would be problematic. The Davinci model when prompted by the question *which one?* justifies its response with an explanation of grammatical appropriateness: *If we use the noun that appears in the context, it is clear that the speaker is referring to the toothbrush. There is no other "it" in the sentence... We would never say, "Pick up the magazine." This is why it's important to know whether the noun is the subject or object of a sentence*. This explanation echoes something meaningful about grammar, yet is faulty and unclear. Rather, this argument is produced from statistical correlations the system extracted from large corpora. Furthermore, the system has no understanding of norms or how to apply them. The potential danger, here, is that simply employing deep learning systems without giving them a sense of norms will lead such systems to also violate norms. While the consequences of breaking norms can range in severity, at the most extreme end, they can include harm to other people.

A norm aware reference resolution system, therefore, will not only help to disambiguate referents but help a system know what *not* to do. This is especially important with embodied agents whose actions in the real world will be influenced by its reference resolution capabilities and natural language understanding.

5.1 Implementation in Embodied Agents

Inspired by our experimental results, we outline a potential methodology for robots to use social norms and conventions in performing situated reference resolution. In order to make the inferences necessary for selecting the correct referent in our scenarios, a novel pragmatic component must be tightly integrated with vision and natural language processing in a robotic cognitive architecture. All three inputs will simultaneously contribute to the interpretation of a referring expression.

A pragmatic component will serve as a knowledge base specifically for social norms and it would require a baseline representation of norms, which

can be collected experimentally for a particular domain (Malle et al., 2020). Upon hearing natural language input from a co-located speaker, a robot will begin incrementally processing the natural language and look for a referring expression. At the same time, the visual system will scan the environment for two purposes: to search for perceptually salient objects that potentially match the referring expression and to trigger the setting to activate a set of norms. For example, spotting a fork, plate, or table, the robot can infer with greater probability that it is located in a dining room and cue an inventory of social norms operationalized as prescriptions and prohibitions. Some of these prescriptions, informally, might be: *food or drinks are allowed on the table* or *you are allowed to sit at the dinner table*. Alternatively, some prohibitions might be *X items should not be on the dining table* or *food and drink should be contained on the dinner table*.

Incremental processing will allow the robot to gradually look for potential referents in the scene and, if it finds potential candidates to match the referring expression, it will also consider the joint probability of each referent given the social norms. The social norms activated from the setting should contribute to the interpretation from the start, not only when an ambiguous situation arises, since they can modulate the interpretation at any point; as seen from our experimental results, regardless of the linguistic form of the referring expression, the social norms can flip the interpretation of the referent when everything else is constant. An advantage to using an inventory of social norms in this way is that they can eliminate potential referents right away. The strength of the norm, roughly corresponding to their deontic force, must be considered for a fine-grained application of norms as some norms will compete with each other. Additionally, it will be critical to understand what norms may or may not be overruled as not to cause harm to human users. While norm activation begins early on, it can continually update through visual and natural language input. If the robot is uncertain about the setting, for instance, it can ask clarifying questions to gain more information. This approach seems applicable in preventing harm where it might be better in many instances to ask questions in uncertain contexts than to overstep boundaries.

To walk through a situated reference resolution scenario, and use a scenario from our experiment, imagine someone commanding the robot: *remove it*. Even if the speaker pauses after the verb *remove*, incremental processing begins to parse the utterance and the robot visually scans the environment for the setting and salient objects. The robots activates the norms stored in the social norm knowledge base and continues processing the input. Once the utterance is completely processed, the expression, *it*, is linked to either a shoe or a spoon. With no other cue from the linguistic input, the prohibition of shoes on dining room tables pushes the interpretation towards removing the shoe. The norm is determined to be strong enough for the robot to act and so it proceeds to remove the shoe. Thus, the robot successfully uses its norm knowledge base in tandem with its vision and natural language processing abilities, to handle what appears on the surface to be an underspecified referring expression.

6 Conclusions

We conducted a human subjects study to demonstrate how social norms can guide reference resolution. Given a text vignette and a referring expression stripped of linguistic cues, the majority of subjects confirmed the intended referent in each context and relied on knowledge of conventions to make their decision. In contrast, several NLP tools evaluated on the same examples consistently failed to select the correct referent. We argue that these NLP tools critically lack an understanding of conventions and social norms and should not be completely relied on for reference resolution as they can also violate norms.

Finally, we integrate our findings into designing a methodology for teaching robots to use social norms and conventions to perform situated reference resolution. In future work, we experiment with using visual scenes for activating norms and evaluate larger NLP models with fine-tuning to our task. Lastly, we will implement our methodology into a cognitive architecture and look more closely at how the gradability of social norms influences reference resolution.

Acknowledgements

We are grateful to James Pustejovsky for assisting with our experimental contexts, the anonymous reviewers for their helpful feedback, and the SMART scholarship for funding the first author.

References

Henk Aarts and Ap Dijksterhuis. 2003. The silence of the library: environment, situational norm, and social behavior. *Journal of personality and social psychology*, 84(1):18.

Cristina Bicchieri. 2005. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Joyce Y Chai, Pengyu Hong, and Michelle X Zhou. 2004. A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of the 9th international conference on Intelligent user interfaces*, pages 70–77.

Jon Chamberlain, Massimo Poesio, Udo Kruschwitz, et al. 2008. Phrase detectives: A web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics' 08)*, pages 42–49.

Robert B Cialdini, Carl A Kallgren, and Raymond R Reno. 1991. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology*, volume 24, pages 201–234. Elsevier.

Kevin Clark and Christopher D Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2018. The knowref coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. *arXiv preprint arXiv:1811.01747*.

Jenny Rose Finkel, Trond Grenager, and Christopher D Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370.

Joy E Hanna and Michael K Tanenhaus. 2004. Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive science*, 28(1):105–115.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Ronnie Janoff-Bulman, Sana Sheikh, and Sebastian Hepp. 2009. Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of personality and social psychology*, 96(3):521.

D. Jurafsky and J.H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Series in Artificial Intelligence. Pearson Prentice Hall.

Doruk Kicikoglu, Richard Bartle, Jon Chamberlain, and Massimo Poesio. 2019. Wormingo: a'true gamification'approach to anaphoric annotation. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, pages 1–7.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Bertram F Malle, Eric Rosen, Vivienne B Chi, Matthew Berg, and Peter Haas. 2020. A general methodology for teaching norms to social robots. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1395–1402. IEEE.

Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. 2014. Learning from unscripted deictic gesture and language for human-robot interactions. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(1):1–44.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. 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)*, volume 1, pages 1778–1789. Association for Computational Linguistics.

Michael J Spivey, Melinda J Tyler, Kathleen M Eberhard, and Michael K Tanenhaus. 2001. Linguistically mediated visual search. *Psychological science*, 12(4):282–286.

Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Luis Von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

David Whitney, Miles Eldon, John Oberlin, and Stefanie Tellex. 2016. Interpreting multimodal referring expressions in real time. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3331–3338. IEEE.