IPOWER: Incremental, Probabilistic, Open-World Reference Resolution

Will Culpepper (wculpepper@mines.edu)
Thomas Bennett (tbennett@mines.edu)
Lixiao Zhu (lixiao.zhu@outlook.com)
Rafael Sousa Silva (rsousasilva@mines.edu)
Ryan Blake Jackson (rbjackso@mines.edu)
Tom Williams (twilliams@mines.edu)
Department of Computer Science, Colorado School of Mines
Golden, CO 80401 USA

Abstract

Referring expression understanding and generation are critical for robots to communicate about the world around them. Recently there have been significant advances on the problem of referring expression understanding, also known as *reference resolution*, with researchers presenting approaches to both incremental reference resolution (i.e., processing referring expressions word by word in real-time as they are spoken) and open-world reference resolution (i.e., resolving references both to known and previously unknown entities). In this work, we combine insights from these approaches to present IPOWER: the first algorithm for performing reference resolution incrementally in open-world environments.

Introduction

A robot designed to help bullied children asks a child about their relationship with their classmate. A robot at a public bus terminal helps a new immigrant to find their way across the city that is their new home. A robot assisting a person with Parkinson's issues a gentle reminder of which medications still need to be allocated to their pillboxes. Like all of these cases, the robotics applications of the future will rely on situated natural language understanding and generation capabilities in which robots must be able to talk about the people, places, and things that are found in the environments they share with their human teammates.

One of the key capabilities of situated language understanding is *reference resolution*: the process of determining which entities are being referred to in an utterance one has heard (Van Deemter, 2016). Within the AI community, especially within the field of Robotics, significant attention has been paid to the problem of reference. While much of this work has been focused on general *language grounding*, the problem of associating words with concepts or stimuli (Tellex, Gopalan, Kress-Gazit, & Matuszek, 2020), substantial recent work has specifically focused on the *reference resolution* problem of associating complete referring expressions with mental representations of specific entities. In particular, a variety of recent work has sought to enable the resolution of referring expressions *in a way that is uniquely tailored* to the nuances of situated interaction. Situated communication presents unique challenges for language understanding due to its temporal and mnemonic characteristics.

First, situated language understanding unfolds over time. This means that robots, for example, must be able to understand language *as it is coming in* and cannot simply wait until an entire sentence has been heard to start processing it. Accordingly, some researchers have begun to research *incremental* reference resolution (i.e., processing referring expressions word by word in real-time as they are spoken) (Kennington & Schlangen, 2015). Second, in situated dialogue, people regularly introduce new entities into the dialogue, and it cannot be assumed that a robot will know a priori of all entities that could be described. Accordingly, some researchers have begun to research *open-world* reference resolution (i.e., resolving references both to known and previously unknown entities) (Williams & Scheutz, 2015b).

In this work, we present the first approach that combines these two capabilities. Our approach, *IPOWER*, is capable of Incremental, Probabilistic, Open-World Reference Resolution of referring expressions to representations stored in a set of distributed, heterogeneous knowledge base (DHKB, cf. (Williams & Scheutz, 2016)), and is implemented within the Distributed, Integrated, Affect, Reflection, Cognition (DIARC (Scheutz et al., 2019)), a component-based cognitive architecture with rich language understanding and generation capabilities and a goal-driven approach to action selection and execution. DIARC components work asynchronously and are able to exchange information with other components in operation. **The central claim of this paper is that by developing reference resolution algorithms that are both incremental and open-world compatible, we should achieve the best of both worlds, being able to handle open worlds while increasing performance relative to non-incremental algorithms like POWER (Williams & Scheutz, 2015b).**

In the rest of the paper, we will: (1) summarize the related work that IPOWER builds upon, (2) define and justify our algorithmic approach, (3) provide a proof-of-concept demonstration of IPOWER's operation in a natural language instruction scenario, (4) evaluate the time performance of IPOWER compared to its non-incremental predecessors, and (5) conclude with directions for future work.

Related Work

In this section we describe the space of recent work performed on reference resolution, with especial attention paid towards work performed in situated domains.

Reference Resolution

A crucial aspect of natural language communication is the ability to *refer* (Green, 1996). Humans commonly use expressions that "pick out" some entity about which we want to make some claim, request some information, or issue some command. These so-called *referring expressions* come in a variety of forms (Strawson, 1950), including demonstrative pronouns (e.g., 'this' and 'that'), personal and impersonal pronouns (e.g., 'I', 'you', 'he', 'it'), proper names (e.g., 'Mount Evans', 'Angela Davis'), and definite and indefinite noun phrases (e.g., "I have eaten *the plums that were in the icebox*", "There is *a house in New Orleans*").

Perhaps the most popular approach toward understanding referring expressions is *co-reference resolution* (Ng, 2010; Soon, Ng, & Lim, 2001), in which new referring expressions are "linked" with previously heard referring expressions. For example, for the sentence pair "The commander needs the medical kit. He says that he left the medkit in the atrium", a co-reference resolution system should identify that [The commander], [He], and [he] all co-refer, as do [the medical kit] and [the medkit].

While co-reference resolution has been popular in textual domains, it is typically insufficient in situated contexts like robotics, where referring expressions must be understood by robots to refer to entities in "the real world". The robotics community has thus emphasized the problem of identifying what real-world entities are the referents of referring expressions, formulating this problem in different ways, such as "language grounding" (Steels & Hild, 2012), "reference resolution" (Popescu-Belis, Robba, & Sabah, 1998), and "entity resolution" (Meyer, 2013).

Moreover, in realistic robotics contexts, this problem is made more difficult due to the tenuous connection between language and physical reality. While *referring* is something assumed to happen between linguistic expressions and *real world entities*, people commonly refer not only to things that exist in the real world but which their interlocutors have no knowledge of, but also to things that are explicitly understood not to exist (e.g., hypothetical or imaginary entities). Accordingly, the problem of reference resolution may be best viewed as the identification of the *mental representations* that may or may not actually be associated with real world entities; a process that may require *creation of new mental representations as part of the reference resolution process*. This broader, representation-focused view of reference resolution, is what is known as *Open-World Reference Resolution* (Williams & Scheutz, 2015a).

Open-World Reference Resolution

Classic computational models of reference resolution operate under a closed world assumption, i.e., such approaches are only able to resolve references with respect to a set of entities whose identities and properties are known *a priori*. A situated agent cannot, however, be expected to know of every object, location, and person in its environment, especially while exploring new environments (e.g., in search-and-rescue scenarios). Open-world reference resolution algorithms address this limitation by determining which parts of referring expressions refer to known versus unknown entities, and by updating the listener's world model when unknown entities are presented or when new knowledge is received. This allows the listener to re-identify the new entity when it is referred to again in the future or to ground the entity when it is observed in the world.

Approaches towards *open-world* reference resolution have been presented by Williams and Scheutz (2015b, 2016), Duvallet et al. (2016), and Tucker, Aksaray, Paul, Stein, and Roy (2020). Williams and Scheutz (2016), for example, show how new object representations can be hypothesized and asserted into memory during reference resolution, allowing agents to communicate about entities in their environment without needing to have previously observed those entities, and allowing agents to thus operate in incompletely known environments. The DIST-POWER algorithm (Williams & Scheutz, 2015b) is also notable as it is the first *distributed* reference resolution algorithm, operating on knowledge distributed across multiple architectural components on multiple machines, while staying agnostic to knowledge representation details.

Incremental Reference Resolution

Traditional reference resolution algorithms, including these open-world variants, operate by first listening to an entire sentence, parsing that sentence into a set of semantic constraints, and then identifying the objects described in the sentence from a knowledge base of potential candidates in the current environment and the semantic constraints that apply to them (Chai, Hong, & Zhou, 2004). Critically, this approach is inconsistent with psycholinguistic accounts of reference resolution, which suggest that humans *incrementally* resolve references as an utterance unfolds word by word, rather than waiting to hear an entire sentence (Poesio & Rieser, 2011). Substantial evidence for this has come from eye tracking experiments conducted through the visual world paradigm (Allopenna, Magnuson, & Tanenhaus, 1998; Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995). The resolution speed and backchanneling are critical for robot designers seeking to achieve smooth and natural human-robot interactions (Stivers et al., 2009; Thomaz & Chao, 2011).

While there has also been previous work on incremental reference resolution (Brick & Scheutz, 2007; Kennington & Schlangen, 2015, 2017; Schlangen, Baumann, & Atterer, 2009), those works have operated under closed-world conditions. As such, no computational model of reference resolution has yet been presented which is both open world and incremental.

In this paper, we thus present a model that combines the benefits of these previously presented approaches in order to enable incremental understanding of referring expressions in uncertain and open worlds. This model assumes that the final intended meanings and inferences intrinsic to each word can

be derived as soon as that word has been completely uttered. In this way, there is no risk of starting the incremental parsing and creating incorrect assumptions before the sentence is completed. We refer to this algorithm as IPOWER: *Incremental, Probabilistic, Open-World Reference Resolution.*

Technical Approach

Due to the intended *incremental* nature of IPOWER, our algorithm is designed to be called repeatedly as a series of semantic constraints are sequentially provided by an incremental parser. When each of these semantic constraints is provided, IPOWER (Alg. 1) is called with six parameters $\langle S, M, H, P, U, E \rangle$ as described below:

1. S: A set of semantic constraints imposed by the most recently heard portion of a referring expression, such as $cup(X)$, $on(X,Y)$, or $table(Y)$.
2. P: An initially empty set of variables used to keep track of the semantic history of the utterance.
3. H: An initially empty set of hypotheses for bindings between variables V appearing in S and entities known of by the robot. These hypotheses are made up of three components: (1) a set of bindings mapping a variable to a candidate referent; (2) a list of constraints imposed on those referents by the referring expression; and (3) the incrementally computed likelihood of the hypothesis, calculated as the joint probability of each constraint imposed thus far holding for the hypothesis' candidate referents under a naive independence assumption.
4. M: A consultant[1] able to provide information about the entities within the environment, their properties, and their relationships to each other in the form of semantic constraints.
5. U: An initially empty set of set of semantic constraints that is used to store variables that need to be hypothesized at the end of the utterance.
6. E: An end-of-clause flag indicating that no further constraints are expected.

When IPOWER is called with these parameters, incremental, probabilistic, open-world reference resolution is achieved as follows (as described in Alg. 1):

If there are no working hypotheses for resolution (Line 1), an initial set of hypotheses are created by taking the first variable appearing in the first semantic constraint in S (Line 2), and creating hypotheses in which this variable is bound to each entity known of by consultant M (Lines 3-7).

Next, this new or pre-existing set of working hypotheses is pruned using the Distributed Closed-World Reference Resolution (DIST-CoWER) algorithm (Williams, 2017b), which uses the set of semantic constraints S to guide a search through the space of possible variable-entity assignments, pruning branches whose incrementally computed probability

Algorithm 1 $IPOWER(S,P,H,M,U,E)$

1: **if** $H = \emptyset$ **then**
2: $\quad v = S_0^{V_0}$
3: $\quad$ **for all** $m \in M$ **do**
4: $\quad\quad b = (v \rightarrow m)$
5: $\quad\quad H = H \cup \{\{b\}, P, 1.0\}$
6: $\quad$ **end for**
7: **end if**
8: $H' = DIST-CoWER(S^V, S, H, M)$
9: **if** $H' = \emptyset$ **then**
10: $\quad S' = \{s \in (P \cup S) | s^V = s_0^{V_0}\}$
11: $\quad$ **return** $IPOWER((P \cup S \setminus S'), \emptyset, \emptyset, M, (U \cup S'), E)$
12: **else**
13: $\quad$ **if** E = true **then**
14: $\quad\quad$ **return** $(P \cup S, H', M.update(U), \emptyset, E)$
15: $\quad$ **else**
16: $\quad\quad$ **return** $(P \cup S, H', M, U, E)$
17: $\quad$ **end if**
18: **end if**

falls below a given threshold. The set of remaining hypotheses are then stored in H' (Line 8).

If no hypotheses remain after this call to DIST-CoWER (Line 9), then it is presumed that at least one of the entities described in the utterance heard thus far must be a new entity not yet known of to the robot. In this case, IPOWER identifies all semantic constraints heard thus far S' that contain the first unbound variable in the first constraint in S (Line 10), associates that variable with a placeholder value ("?") and makes a recursive call to IPOWER to re-attempt reference resolution under the assumption that: (1) those constraints S' no longer need to be handled, (2) the set of working hypotheses and pre-considered semantic constraints should be re-set, and (3) the set of semantic constraints S' should be held aside as new information to later be asserted (Line 11).

Finally, if there is no evidence that the referring expression has been heard in its entirety, new intermediate values for parameters P,H,M,U,E are returned (Line 16). If there *is* evidence that the referring expression has concluded, then before these values are returned, new representations for any entities associated with placeholder values are created, and any properties relating to those variables, held aside in U, are asserted into the robot's world model (Line 14).

Demonstration

To demonstrate the behavior of IPOWER in detail, we will step through how IPOWER incrementally handles a practical example. The example below represents the actual output of IPOWER after implementation as a Component of the DIARC Architecture (Scheutz et al., 2019)[2] in a context in which the attached POWER Consultant's Knowledge Base (implemented as a DIARC component) contains the *Uncertain Persons* Knowledge Base, which consists of knowledge

[1] Cp. the consultant framework presented by Williams (2017a).

[2] Source code for this Component is available upon request.

of seventeen distinct people (Williams & Scheutz, 2015a) (with associated information about those persons and hand-coded uncertainty levels). Within the context of this Knowledge Base, we will examine how IPOWER processes the following utterance, which is a novel modification of one of the 16 predefined *Uncertain Persons* test cases:

The chemist Nicolas, Billie's father.

We assume that as this utterance is heard, it is recognized and parsed incrementally by the DIARC, producing the following constraints: $\langle profession(X, chemist), named(X, nicolas), named(Y, billie), parent(X, Y), gender(X, male)\rangle$. The following represents a trace of IPOWER as it is incrementally provided with each of these five constraints over the course of five algorithm calls.

The first constraint received is $profession(Y, chemist)$. Because H is initially empty, $IPOWER$ creates an initial set of hypotheses containing three hypotheses: $\{(X \rightarrow people\_9, 0.99), (X \rightarrow people\_10, 0.99), (X \rightarrow people\_8, 0.99)\}$.

The second constraint received is $named(X, nicolas)$. Through DIST-CoWER, IPOWER applies this constraint to each hypothesis in IPOWER's list of current hypotheses. In doing so, IPOWER considers whether each of the previously identified chemists is known to be named "Nicolas". This results in a refined hypothesis set $\{(X \rightarrow people\_9, 0.9801)\}$ as only $people\_9$ is known to be named "Nicolas".

The third constraint received is $named(Y, billie)$. IPOWER attempts to find a person in the knowledge base named Billie, and upon failure associates Y with a placeholder entity ("?") to signify the need for future creation of new mental representations, updates the (sole remaining) hypothesis to include a binding to this placeholder (i.e., $\{(X \rightarrow people\_9, Y \rightarrow ?, 0.9801)\}$), and records the fact that it will need to later assert property $named(Y, billie)$ when a new mental representation is ultimately created and bound to X. Because this property involves a previously unknown entity, there is no reason to believe that this constraint does not hold for that entity, and as such, the probability associated with its maintained hypothesis is not changed.

The fourth constraint received is $parent(X, Y)$. Again, one of the variables here is associated with a placeholder value in the hypothesis IPOWER is maintaining, so IPOWER merely sets the constraint aside to be asserted later on, and does not modify the probability associated with its sole maintained hypothesis.

Finally, the fifth constraint received is $gender(X, male)$, with the end-of-sentence flag set. Through DIST-CoWER, IPOWER applies this constraint to the sole remaining hypothesis. In doing so, IPOWER considers whether the sole remaining chemist is known to be male. This results in a refined hypothesis set $\{(Y \rightarrow people\_9, 0.9703)\}$. Because the end of sentence flag is set, IPOWER finally creates a new mental representation to be associated with Y ($people\_18$), and requests the $people$ consultant to assert the held properties $(\{named(people\_18, billie), parent(people\_9, people\_18)\})$ as part of that new representation. Finally, the complete grounded hypothesis, $(X \rightarrow people\_9, Y \rightarrow people\_18, 0.9703)$, is returned.

Evaluation

In this section we present the results of two experimental evaluations. In the first experimental evaluation, we performed a coverage analysis, where we examined the output of IPOWER on a key set of benchmark test cases identified by researchers in previous work to identify degree of consistency with previous work. In the second experimental evaluation, we compare the speed-based performance of IPOWER compared to POWER on those benchmark test cases.

Coverage Analysis

Experimental Design To assess the consistency of IPOWER with previous work, we examined its output on the *Uncertain Persons* benchmark test cases originally presented by Williams and Scheutz (2015a). In that work, Williams and Scheutz (2015a) presented sixteen test cases that systematically examined sixteen key types of uncertainty and ignorance, in which the referent and the anchors with respect to which they are described are each either resolvable to 0 referents, 1 referent, 1 referent (but tenuously) or multiple referents. These test cases were previously used to evaluate POWER (Williams & Scheutz, 2015b), using a knowledge base of seventeen previously known persons.

Previously, Williams and Scheutz (2015b) claimed perfect accuracy of POWER on these sixteen test cases. By also evaluating IPOWER using these test cases, we are able to assess the consistency of IPOWER with previous results. To do so, we provided IPOWER with the same knowledge base previously provided by Williams and Scheutz (2015b) to POWER, and ran the same sixteen test cases, providing the same sixteen sets of predicates to I-POWER one predicate at a time.

While other tasks related to natural language processing (NLP) often benefit from evaluation under significant linguistic variation, the relatively small set of test cases used here is a suitable means of evaluation due to the nature of reference resolution. Reference resolution takes logical sentence representations from a semantic parser. As such, it is the job of the parser to deal with linguistic variation; introducing linguistic variation would only be suitable for evaluating the parser, which is not the focus of this work. In short, it would be inappropriate to vary linguistic phrasing when trying to evaluate reference resolution, and thus inappropriate to do a "corpus-based" evaluation of this work.

Results IPOWER produced identical outcomes to POWER in all test cases save one, "The chemist, Billie's father" which probes a situation in which the target has multiple possible referents when viewed on its own, but its anchors have no possible referents. While POWER interpreted this as evidence that all parties described were previously unknown, IPOWER instead interprets this as evidence that the target is unknown, but that one of the possible anchors is likely to be a

Referring Expression	POWER (ms)	IPOWER (ms)
The sister of the doctor's friend	1848.6	1759.2
Jim's friend	1583	908.2
Jim's daughter	1253.4	853.2
Tabitha's mother	1349.2	977
The chemist's neighbor	1407.2	1101.4
Craig's coworker's neighbor's son	1980	1908.6
Craig's coworker's neighbor's daughter	2082.2	2010
Marion's daughter Kristy	1346	1313.6
Troy's girlfriend	1062.6	1060.6
The baker's brother	999.4	999.4
The chemist, Billie's father	1597	1191.6
Michelle's daughter, Willie	1390.8	1372.2
Sally's wife	863.8	858.8
The Wells boy's girlfriend	1551	1553.8
Troy Wells, the podiatrist's friend	2200.2	2171
The podiatrist's friend	1468.8	1179.8
Mean (SD)	**1498.95 (380.83)**	**1326.15 (429.60)**

Table 1: From left to right: (1) Referring expression test case, (2) mean runtime for POWER, (3) mean runtime for IPOWER.

true anchor. This difference reveals an interesting philosophical difference between the two algorithms. While POWER is able to rely on a pre-provided variable ordering method that establish a sequence of referent-anchor pairs, IPOWER is not able to rely on any such variable ordering, and must instead operate under an assumption that early-referenced entities are assumed to be more well-known than late-referenced entities. This means that while IPOWER does not produce results consistent with POWER in this case, this is not necessarily an error, but rather a philosophical position that stems from a relaxed assumption. This means that IPOWER would be more likely to correctly resolve expressions such as "My dentist's neighbor, Barack Obama", correctly resolving Barack Obama even if "My dentist's neighbor" would on its own be resolved to no previously known referent.

Runtime Analysis

Experimental Design After assessing algorithmic consistency, we experimentally evaluated the speed performance of IPOWER as compared to POWER. To do so, we began by measuring the amount of time necessary for a speaker to utter each of the sixteen Uncertain Persons test case utterances, and then identified the amount of time needed to utter each constituent part of that sentence. For example, Test Case 1, "The sister of the doctor's friend" was measured to take 1748ms to utter; 793ms for "The sister of", 563ms for "the doctor's" and 392ms for "friend". We then calculated average resolution time (over five runs) for each test case, under the assumption that processing begins as soon as information becomes available, and under an assumption of instantaneous speech recognition and parsing.

For POWER, runtime was calculated as "time to speak utterance" + "time to process entire set of predicates immediately after hearing the utterance". As an example, for Test Case 1, the mean time needed for POWER to process the test case was 100.6ms, producing a total necessary time from start of utterance to end of processing of 1748+100.6=1848.6ms.

For IPOWER, runtime was calculated differently due to its incremental nature. Specifically, runtime was calculated as the time to hear the first referring expression constituent, and then the time to process each predicate, with an appropriate delay inserted if the next referring expression constituent were not yet available for processing. As an example, for Test Case 1, the first 793ms of runtime are consumed by hearing "The sister of" as described above. Processing the associated predicate takes 795.4 ms. By the end of this time, at 793+795.4=1588.4ms, the next predicate is available for processing as its associated constituent ("the doctor's") completed after 793+563=1356ms. Processing this predicate takes 72.4, bringing us to 1588.4+72.4=1660.8ms. The final constituent ("friend") would be ready at 1356+392=1748ms, meaning that IPOWER would need to wait until the utterance completed to process the final predicate, which it would accomplish in 11.2ms, for a total runtime of 1748+11.2=1759.2ms, nearly 100ms faster than POWER despite POWER's otherwise much faster processing speed in this particular case. These calculations were performed for all 16 test cases. Analysis was performed on a laptop running Ubuntu 18.04, with a 2.20 Ghz CPU and 8 GB of RAM.

Results As shown in Tab. 1, IPOWER performed, on average, 172ms faster than POWER, with IPOWER running 674ms faster in the best case (Test Case 2) and 2.8ms slower in the worst case (Test Case 14). To provide a measure of certainty for the performance of IPOWER over POWER, we performed a paired-samples t-test between the two sets of test case runtime means. Clearly the sixteen test cases were

not randomly sampled from a distribution, as they were intentionally selected by previous authors to represent distinct categories of uncertainty and ignorance. However, if these cases had resulted from random sampling, a paired samples t-test would have suggested a significant difference between the two algorithms' results (p=.004). As such, even though the sampling assumptions of this t-test are violated, we believe this allows us to straightforwardly claim with some confidence that the incrementality provided by IPOWER does indeed result in a net benefit over POWER, confirming our central research hypothesis.

Conclusion

In this paper, we have presented IPOWER, an algorithm for incremental open-world reference resolution. We discussed how this algorithm is able to combine the strengths of previous approaches to *open world reference resolution* and *incremental reference resolution* to achieve results that are more computationally efficient than previous approaches to the former, and more readily deployable in realistic open-world contexts than previous approaches to the latter. These advances are important because the success of language-capable robots depends both on (1) the ability for robots to perform as close to real-time as possible in order to mimic the incredibly short turn lengths observed in much human dialogue, and (2) the ability for robots to operate in realistic environments in which they must acknowledge their uncertainty and ignorance and leverage opportunities to learn about the environment they share with their human teammates.

IPOWER does, however, have a number of limitations that present opportunities for future work. First, as we discussed in the Coverage Analysis, there is a slight difference between the behavior of IPOWER and POWER when presented with particular types of uncertainty and ignorance. This difference highlights the context-sensitive nature of many reference resolution tasks; IPOWER and POWER perform differently because they are operating under slightly different philosophical assumptions, each of which may be more or less appropriate in different contexts, based on nuanced aspects of common sense knowledge. This means that future systems may need to be sensitive to these contextual differences and intelligently decide whether or not to make the types of assumptions held by POWER but relaxed by IPOWER.

Second, while IPOWER is faster than POWER, it is not yet clear whether this timing difference is large enough to be noticed, either on its own, or when combined with the time savings of using incremental algorithms for other language processing tasks as well. It is unclear how large of a delay humans are willing to tolerate or indeed are able to notice before their perceptions of an interlocutor begin to degrade. Human subject experimentation with real interactive robots and live human participants will be needed to provide an answer to this question.

Third, in order to perform this sort of experiment, IPOWER will need to be more deeply integrated into the DIARC architecture, and will need to be able to handle knowledge provided by multiple distributed consultants at the same time, much like DIST-POWER (Williams & Scheutz, 2016). IPOWER is currently not able to handle multiple consultants as it is not currently integrated into an architectural component capable of managing these different architectural connections the way that DIST-POWER is. Similarly, the benefits of IPOWER's incrementality will only be truly realized when DIARC's downstream language understanding components such as pragmatic reasoning (Williams, Briggs, Oosterveld, & Scheutz, 2015) are also made to operate incrementally, and when the larger Givenness Hierarchy Theoretic reference resolution system into which IPOWER is integrated (Williams, 2019) is similarly made to be incremental.

Fourth, there are intriguing questions as to how the knowledge representation and reasoning systems that IPOWER interacts with should behave should it turn out that IPOWER resolved an utterance incorrectly, e.g., due to incorrect world knowledge. In particular, it is unclear how these sorts of architectural components should handle previously created knowledge representations that correspond with entities that do not turn out to exist.

Fifth, human-like reference resolution is subject to error. When listening to a sentence, humans can create hypotheses that do not correctly link references to referents. As we have mentioned, IPOWER assumes the semantics of each word can be accurately derived as soon as it has been uttered. Alternative accounts, in contrast, could adopt a more aggressive early-resolution strategy, paired with increased reliance on late-repair. Future work should investigate how IPOWER performs in situations where our assumptions are violated, and compare to these types of models.

Sixth, because IPOWER works by multiplicative combining sources of evidence, it necessarily becomes less certain as more information about the target is provided. Future work could consider alternative methods of evidence combination, such as Dempster-Shafer Theoretic approaches Shafer (1976), many of which naturally avoid this problem. We have leveraged this approach in other areas of our prior work Williams et al. (2015); Williams, Yazdani, Suresh, Scheutz, and Beetz (2019).

Finally, it is unclear how IPOWER should practically be performing if upon utterance completion it still has multiple hypotheses viewed as plausible, each of which involve a placeholder value. It is unclear in this situation whether IPOWER should create a new mental representation associated with each such hypothesis, whether it should create a single mental representation associated with simply the most probable of those hypotheses, whether it should wait until some clarification dialogue has completed before making this sort decision, or something else. Again, these types of questions are computationally important and philosophically intriguing problems that indicate the breadth of what is still unknown when it comes to open-world cognition.

Acknowledgments

This work was funded in part by grant N00014-21-1-2418 from the US Office of Naval Research.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language*, *38*(4), 419–439.

Brick, T., & Scheutz, M. (2007). Incremental natural language processing for hri. In *Proceedings of the acm/ieee international conference on human-robot interaction* (pp. 263–270).

Chai, J. Y., Hong, P., & Zhou, M. X. (2004). A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of the 9th international conference on intelligent user interfaces* (pp. 70–77).

Duvallet, F., Walter, M. R., Howard, T., Hemachandra, S., Oh, J., Teller, S., ... Stentz, A. (2016). Inferring maps and behaviors from natural language instructions. In *Experimental robotics* (pp. 373–388).

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of psycholinguistic research*, *24*(6), 409–436.

Green, G. M. (1996). *Pragmatics and natural language understanding*. Psychology Press.

Kennington, C., & Schlangen, D. (2015). Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proc. acl* (pp. 292–301).

Kennington, C., & Schlangen, D. (2017). A simple generative model of incremental reference resolution for situated dialogue. *Computer Speech & Language*, *41*, 43–67.

Meyer, F. (2013). *Grounding words to objects: A joint model for co-reference and entity resolution using markov logic for robot instruction processing*. Unpublished doctoral dissertation, Hamburg University of Technology (TUHH), Hamburg, Germany.

Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the forty-eigth annual meeting of the association for computational linguistics (ACL)* (pp. 1396–1411).

Poesio, M., & Rieser, H. (2011). An incremental model of anaphora and reference resolution based on resource situations. *D&D*, *2*, 235–277.

Popescu-Belis, A., Robba, I., & Sabah, G. (1998). Reference resolution beyond coreference: a conceptual frame and its application. In *Proceedings of the seventeenth annual conference on computational linguistics (COLING)* (pp. 1046–1052).

Scheutz, M., Williams, T., Krause, E., Oosterveld, B., Sarathy, V., & Frasca, T. (2019). An overview of the distributed integrated cognition affect and reflection diarc architecture. In *Cognitive architectures* (pp. 165–193). Springer.

Schlangen, D., Baumann, T., & Atterer, M. (2009). Incremental reference resolution: The task, metrics for evaluation, and a bayesian filtering model that is sensitive to disfluencies. In *Proceedings of the sigdial 2009 conference* (pp. 30–37).

Shafer, G. (1976). A mathematical theory of evidence. In *A mathematical theory of evidence.* Princeton university press.

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, *27*(4), 521–544.

Steels, L., & Hild, M. (2012). *Language grounding in robots*. Springer Science & Business Media.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... others (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592.

Strawson, P. F. (1950). On referring. *Mind*, *59*(235), 320–344.

Tellex, S., Gopalan, N., Kress-Gazit, H., & Matuszek, C. (2020). Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, *3*, 25–55.

Thomaz, A. L., & Chao, C. (2011). Turn-taking based on information flow for fluent human-robot interaction. *AI Magazine*, *32*(4), 53–63.

Tucker, M., Aksaray, D., Paul, R., Stein, G. J., & Roy, N. (2020). Learning unknown groundings for natural language interaction with mobile robots. In *Robotics research* (pp. 317–333). Springer.

Van Deemter, K. (2016). *Computational models of referring: a study in cognitive science*. MIT Press.

Williams, T. (2017a). A consultant framework for natural language processing in integrated robot architectures. *IEEE Intelligent Informatics Bulletin*.

Williams, T. (2017b). *Situated natural language interaction in uncertain and open worlds*. Unpublished doctoral dissertation, Tufts University.

Williams, T. (2019). A givenness hierarchy theoretic approach. *The Oxford handbook of reference*, 457.

Williams, T., Briggs, G., Oosterveld, B., & Scheutz, M. (2015). Going beyond literal command-based instructions: Extending robotic natural language interaction capabilities. In *Twenty-ninth aaai conference on artificial intelligence.*

Williams, T., & Scheutz, M. (2015a). A domain-independent model of open-world reference resolution. In *Proceedings of the 37th annual meeting of the cognitive science society.*

Williams, T., & Scheutz, M. (2015b). Power: A domain-independent algorithm for probabilistic, open-world entity resolution. In *2015 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 1230–1235).

Williams, T., & Scheutz, M. (2016). A framework for resolving open-world referential expressions in distributed heterogeneous knowledge bases. In *Thirtieth aaai conference on artificial intelligence.*

Williams, T., Yazdani, F., Suresh, P., Scheutz, M., & Beetz, M. (2019). Dempster-shafer theoretic resolution of referential ambiguity. *Autonomous Robots*, *43*(2), 389–414.