Corrective Processes in Modeling Reference Resolution

Constantine Nakos (cnakos@u.northwestern.edu)
Irina Rabkina (irabkina@u.northwestern.edu)
Samuel Hill (SamuelHill2022@u.northwestern.edu)
Kenneth D. Forbus (forbus@northwestern.edu)
Qualitative Reasoning Group, Northwestern University
2233 Tech Drive, Evanston, IL 60208, USA

Abstract

Reference resolution is one of the core components of language understanding. In spite of its centrality, psychological evidence has shown that the reference resolution process is prone to errors and egocentric bias. In this work, we propose an extension to Analogical Reference Resolution, a computational model based on analogical retrieval, which accounts for such errors. We test the extended model on a study by Epley et al. (2004) and replicate human patterns of bias and correction.

Keywords: reference resolution; perspective taking; analogy

Introduction

Reference resolution is the process of identifying the entities a speaker intends to refer to in an utterance. Although common ground (Clark & Carlson, 1981, 1982) has long been held as the basis for successful reference resolution, experimental evidence paints a more complex picture (Epley, Morewedge, & Keysar, 2004; Keysar et al., 1998, 2000; Keysar, Lin, & Barr, 2003; Lin, Keysar, & Epley, 2010). Listeners adopt a two-stage strategy when resolving referring expressions: an initial, automatic judgment that only makes use of the listener's knowledge followed by a slower corrective process that takes common ground into account. This two-stage process arguably leads to greater efficiency by avoiding the cognitive overhead of keeping explicit track of common ground when unnecessary (Keysar et al., 1998). Given that interlocutors' situation models are often aligned and become more so over the course of a conversation (Pickering & Garrod, 2004), coupled with the fact that corrections in an interactive setting are relatively cheap, this two-stage strategy is an efficient way to make use of the cognitive resources at the listener's disposal.

The processes involved in reference resolution are quite powerful. For instance, people are remarkably adept at understanding *near misses*, referring expressions which fail to accurately describe their intended referent (Donnellan, 1968). For example[1]:

A: "Do you see *the man drinking champagne*?"

B: "He's actually drinking sparkling water."

B is able to interpret A's referring expression even though it is an erroneous description of the man in question. Furthermore, according to Keysar et al. (1998), B is able to do so without a full, explicit model of A and A's potential misconceptions. What can account for this ability?

Nakos, Rabkina, and Forbus (in press) argue that *structure mapping* lies at the heart of reference resolution. According to Structure-Mapping Theory (SMT; Gentner, 1983), structure mapping underlies analogical comparison. It has been proposed as the process responsible for human similarity judgments (Gentner & Markman, 1997). SMT offers a flexible, domain-independent, and psychologically plausible explanation for the ability to match descriptions to referents even in the presence of errors. It also provides an explanation for the gradations of correctness exhibited by erroneous descriptions. Analogical comparison has been shown to be invoked spontaneously by people (Day & Gentner, 2007), making it a plausible fit for the automatic process proposed by Keysar et al. (1998).

In this paper, we extend our model of Analogical Reference Resolution (ARR) to incorporate the two stages proposed by Keysar et al. (1998). Analogical retrieval provides the initial, automatic judgment, while a corrective process, introduced here, identifies common ground violations and triggers re-representation if needed. We test our implemented model on the stimuli used in Epley et al.'s (2004) perspective-taking experiment.

We begin by reviewing SMT and the computational models built on it. Next, we recap ARR and how it makes use of structure-mapping. Then we discuss Keysar et al.'s two-stage model of reference resolution, the experiments used to test it, and a body of related research. From there, we describe an extension to ARR, a corrective process that provides a flexible way for the model to handle perspective-taking and re-representation. We then test our extended model, briefly discuss other models of reference resolution, and conclude.

Background

Analogy

Structure-Mapping Theory (SMT; Gentner, 1983) argues that analogical comparison involves computing mappings between structured descriptions of objects or situations. In particular, mapping occurs between a *base case*, typically a set of facts retrieved from memory, and a *target case*, which describes a new entity or situation to compare to. The outputs of the mapping process are a set of *correspondences* indicating which items in the base correspond to which items

[1] Adapted from (Kripke, 1977).

in the target, a *similarity score* proportional to the depth and connectedness of the mapping, and a set of *candidate inferences*, facts which are projected from the base to the target on the basis of shared structure.

The *Structure-Mapping Engine* (SME; Forbus et al., 2016) is a computational implementation of SMT that has been used to model a wide range of cognitive phenomena, including conceptual change (Friedman & Forbus, 2010), visual similarity (Lovett & Forbus, 2017), and problem-solving by analogy to prior experience (Klenk & Forbus, 2007). Entities and expressions are represented with predicate calculus.

SME has also been used as the basis for a model of analogical retrieval called *MAC/FAC* ("many are called/few are chosen"; Forbus, Gentner, & Law, 1995). Given a *probe case*, MAC/FAC searches a *case library* for the case most similar to it. Inspired by the dichotomy between human recall (which favors surface-level matches) and similarity judgments (which favor deeper, structural matches), MAC/FAC has two stages. The first, MAC, uses cheap, vector-based representations to identify the cases that are most likely to be similar to the probe. The second, FAC, performs full analogical comparison between each of these candidates and the probe. The case with the highest structural similarity score is returned.

Analogical Reference Resolution

The central claim of the ARR model (Nakos et al., in press) is that the automatic stage of reference resolution relies on analogical retrieval. The hearer constructs a representation of the referent based on the speaker's description and then searches through a set of potential referents for the one that is most similar to it. In this model, structural similarity accounts for the ability to resolve near misses. The greater the structural overlap between a referent and its description, the stronger the reference, even if the description is not completely accurate.

In computational terms, ARR consists of constructing a probe case based on the speaker's description and then calling MAC/FAC over a case library of relevant entities. For example, if a speaker used the description "the red apple", the probe case might consist of `(red apple123)` and `(apple apple123)`, where `apple123` is a token representing the entity described by the phrase.

The case library constitutes the model's representation of the visual scene. It contains facts about whatever entities are visible. So if there was actually a green apple on a table, the case library would contain a case consisting of `(green apple1)`, `(apple apple1)`, `(table table2)`, and `(on apple1 table2)`. The cases in the case library must use the same representations as the probe case for SME to capture their similarities. But unlike the probe case, the cases in the case library are not necessarily derived from language. The green apple found in the case library may be an apple that is visually salient, one that was remembered from a previous situation, or one that was described hypothetically. As long as the representations are encoded by processes using the same vocabulary of predicates as the probe case, ARR is agnostic to their source.

MAC/FAC typically retrieves one case, the case that has (approximately) the highest structural similarity score with the probe. MAC/FAC will only fail to retrieve a case if none of the cases in the case library have any overlap with the probe. This means that it will always return a match if one is possible, even if the best one it can find is extremely tenuous.

ARR imposes a similarity cutoff on the matches returned by MAC/FAC, below which retrieval is considered to have failed. This puts a limit on near misses and prevents a grossly erroneous description from matching an arbitrary referent for lack of a better option. When MAC/FAC returns a single match above the threshold, the description is considered uniquely identifying and the reference succeeds. When MAC/FAC returns no match above the threshold, the description is too inaccurate to pick out the intended referent and the reference fails.

A more interesting situation crops up when MAC/FAC returns more than one match for a description. This can happen when the top matches for the probe have similarity scores that are nearly equal. MAC/FAC returning more than one case indicates that the probe was not able to strongly distinguish between them; the reference is ambiguous. This definition of ambiguity goes a step beyond the traditional one, where an underspecified description applies equally well to multiple referents. ARR also identifies when an erroneous description could apply to multiple entities (e.g., "the man drinking champagne" when there is a woman with champagne and a man with sparkling water) or when a correct description fits a distractor closely enough to cause confusion (e.g., "the man drinking champagne" when there are two men, one with champagne and one with sparkling water).

These features make analogical retrieval a natural fit for the automatic stage of reference resolution. Analogical retrieval is an existing process used in other modes of cognition, rather than an ad hoc algorithm designed specifically for reference resolution. It accounts for hearers' robustness to near misses by appealing to structural similarity. Finally, analogical retrieval is conducted in parallel (Forbus et al., 1995) and does not rely on conscious attention (Day & Gentner, 2007), making it suitable for use in an automatic process.

Perspective Taking in Reference Resolution

Keysar et al. (2000) examined the role of common ground in language understanding using a pair of perspective taking experiments. Their experiments were designed to determine whether people include common ground in their initial search for a referent or whether it is factored in later, after an initial judgment has been made. They set up a grid of boxes containing everyday objects between two people, a director and a subject. Several of the boxes were blocked on the director's side so that only the subject could see their contents. The director gave the subject verbal instructions to move certain objects to new locations in order to reach a goal configuration. The subject's eye gaze was tracked to

determine which potential referents were considered and with what timing.

The key factor in the experiment was that some of the referring expressions used by the director best described an object that could only be seen by the participant, a *distractor*. For example, the director might refer to "the small truck" when there are three trucks: a small one that can only be seen by the subject, and a large truck and a medium truck that can be seen by both people. If subjects take common ground into account during their initial search for a referent, they should ignore the distractor (i.e., the small truck), and their gaze should move immediately to the medium-sized truck.

On the contrary, Keysar et al. (2000) found that participants typically looked at the small truck first and only later corrected to the medium-sized truck, in some cases going so far as to reach for the wrong object. This finding is consistent with the Perspective Adjustment model proposed by Keysar et al. (1998). The Perspective Adjustment model holds that the initial stage of reference resolution ignores common ground and instead searches over the set of objects that are relevant to the hearer, regardless of whether the speaker is aware of them. This initial search is followed by a slower process that is responsible for checking that the speaker is, in fact, aware of the selected referent, triggering a search for a more suitable one if necessary.[2]

Several other studies have replicated these findings with minor changes to the experimental setup. Keysar et al. (2003) show that the effect persists even in the absence of direct visual perception of the distractor object. They also show that hearers still show an egocentric bias when they are explicitly made aware of the speaker's ignorance. Epley et al. (2004) confirm the findings in adults and demonstrates that the same egocentric bias exists in children. However, children are even slower to correct themselves after selecting a referent that is not in common ground, suggesting that the second process has not fully developed yet. Lin et al. (2010) delve deeper into the nature of the corrective process and show that the ability to correct for the egocentric bias in reference resolution is impaired by cognitive load. Together, these findings paint a picture of an effortful secondary process which, when executed successfully, can correct the snap decisions made by the initial reference resolution process.

Reference Correction

ARR cannot model these experimental results by itself. Restricting the case library to entities known to the hearer can replicate the egocentric bias but not the process used to correct it, since ARR performs a single retrieval over a fixed set of entities with no way to check its output or change its results. To address this issue, we extend ARR with a corrective process that checks the selected referent against common ground, reconfigures ARR if needed, and kicks off another round of analogical retrieval to obtain a final result. The addition of this corrective process brings ARR in line with the Perspective Adjustment model described by Keysar et al. (1998) and provides a principled way for ARR to take perspective into account.

The timing data from Keysar et al. (2000) suggests that the corrective process is comparatively slow and frequently does not complete in time to prevent subjects from reaching for the wrong item. Per Lin et al. (2010), the corrective process draws on cognitive resources that are shared across a variety of tasks. Subjects under increased cognitive load are more prone to making egocentric errors. This suggests that the corrective process is more cognitively demanding than the automatic initial judgment. Here we posit three steps in the corrective process that account for the increased cognitive demand: Theory of Mind reasoning, suppression, and re-representation. We consider each in turn.

Theory of Mind (ToM) is one person's understanding of another person's mental states. For example, children who have yet to develop full ToM are unable to distinguish their beliefs from another person's (Wimmer & Perner, 1983). The ability to track others' beliefs is crucial for determining what is in common ground, but the egocentric stage of reference resolution suggests that people do not store this information explicitly. Instead, hearers must perform at least some ToM reasoning on the fly. Such reasoning can be arbitrarily difficult due to both the open, flexible nature of common ground and the infinite regress of the mutual knowledge paradox (Clark & Marshall, 1981). Copresence heuristics can alleviate some of the difficulty by limiting the search to a plausible set of scenarios, but in principle, the corrective process may have to perform a large amount of work to catch an error. This explains the need for a fast, automatic process in the first place: to form an initial judgment quickly enough for real-time understanding, at the cost of occasional errors that will have to be corrected.

The corrective process for our model begins with an error identification step, an inferential process that includes ToM reasoning. The hearer applies a set of ToM rules to determine whether common ground has been violated (i.e., the speaker does not know about the selected referent).[3] In principle, error identification can include other types of reasoning, such as making sure the referent is suitable for the task at hand. But without further experimental evidence, it is unclear whether these checks are performed as part of error identification or another step in the resolution process. As such, we limit our error identification to common ground violations for the current work.

The second step of the corrective process is suppression. Referents that are deemed unsuitable by the error identification step are removed from further consideration. This impacts ARR in two ways. First, a temporary case

[2] Keysar et al. (1998) suggest that the two stages may operate in cascade, with the corrective process beginning as soon as partial results from the initial search are available. For simplicity, we treat the two stages as sequential here and leave a cascade model for future work.

[3] Our model makes no claims as to how these rules are acquired. One possibility is that the rules are learned from experience by way of analogical generalization (Gentner & Medina, 1998). Rabkina, McFate, Forbus, and Hoyos (2017) argue that ToM rules are learned this way.

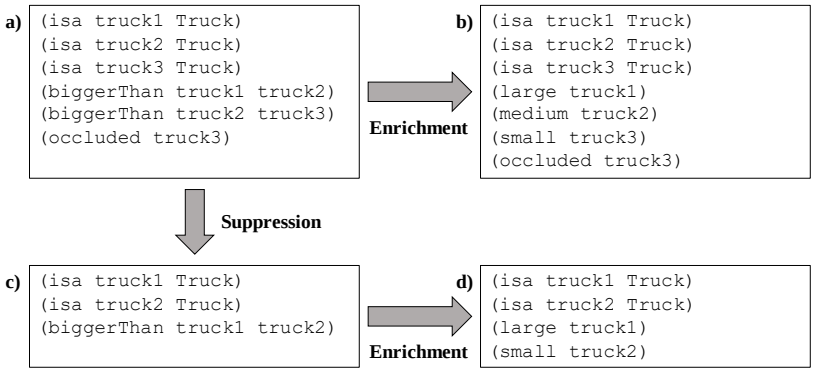

Figure 1: Interaction between enrichment and suppression in the corrective process. Given a representation of a scene (a), enrichment produces attributes for relative size and position (b). These are used for initial retrieval. If error identification determines that the retrieved referent is occluded, suppression produces a reduced scene without the occluded entity (c). This scene is then enriched, yielding a modified set of attributes (d).

library is constructed without the suppressed entities. This achieves the desired effect of ruling out referents that are not in common ground. Second, the removed entities are ignored when determining the meaning of gradable adjectives like "small" or "large". This enables ARR to select a new referent that matches the description, as described below.

The third step of the corrective process deals with re-representation. Suppression alone may not be enough for ARR to find the correct referent. Sometimes the cases themselves must be altered to reflect a new interpretation of the scene. Consider our running example: an occluded small truck, along with a medium truck and a large truck in common ground. Once the small truck has been suppressed, the hearer adjusts his or her interpretation of the sizes of the remaining trucks. Since there are two trucks, one larger and one smaller, the medium-sized truck is re-represented as the "small" one, allowing it to match the description "the small truck" without issue.

To handle gradable adjectives like "small" and "large" in a way that allows for re-representation, our model uses a two-stage encoding process. The first stage encodes type information and relative magnitude of the objects along any relevant dimensions (size, position, etc.). Comparisons are only made within object categories; if the scene contains trucks and candles of varying sizes, they will be compared separately, leading to a natural interpretation of phrases like "the small truck".

The second stage uses these relative magnitudes to assign attributes to the objects, a process we refer to as *enrichment*. For instance, in the original scene, the trucks will be encoded as `small`, `medium`, and `large` based on their relative sizes. Once the small truck has been suppressed, the attributes are recalculated and the remaining two trucks are marked as `small` and `large`, respectively. The separation of the underlying comparisons from the final assignment of attributes allows our model to deal with gradable adjectives in a general, flexible way. The interaction between suppression and enrichment is shown in Figure 1.

To recap, our corrective process begins by identifying common ground violations using ToM inference. It then suppresses any unsuitable entities, removing them from further consideration. This is followed by re-representation, which recomputes the semantics of attributes that correspond to gradable adjectives. With representations finalized, the corrective process then invokes ARR again to make the final choice of referent.

Evaluation

To demonstrate our model of reference resolution, we replicate the perspective taking experiment from Epley et al. (2004) using their stimuli. Our model is presented with a description from one of the trials and the corresponding set of potential referents. For the sake of simplicity, we factor out natural language processing, visual object identification, and actually obeying the speaker's direction to move the identified object. Instead, we encode the object description by hand and use CogSketch (Forbus et al., 2011) to encode the visual scene. CogSketch is a sketch understanding system that allows users to create diagrams with visual objects

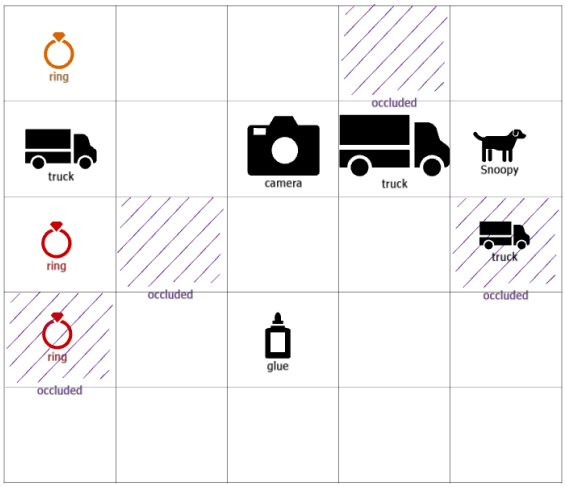

Figure 2: Sample visual scene encoded in CogSketch.

identified using menus. A sample sketch is shown in Figure 2. CogSketch automatically computes visual relationships between objects, creating a set of facts suitable for use with ARR.

Given a description and a scene, our system invokes ARR to produce an initial judgment of the intended referent. If our system is modeling reference resolution correctly, then for trials where the description is ambiguous, the initial referent should be the distractor, which is known to the hearer but not the speaker. Our system then uses a set of simple rules to check whether the speaker can see the selected entity. If not, the system suppresses the initial referent, updates size and position attributes by rerunning the enrichment process, and invokes ARR again. If our system is behaving correctly, the new referent chosen this way should be the correct one.

We find that this is indeed the case. The Epley et al. (2004) stimuli consist of four visual scenes, each with four descriptions that pick out entities in the scene. For each scene, one of the trials is a test condition where the description fits an occluded entity more closely than the intended referent. In each of the 16 trials, our model behaves as expected, immediately selecting the correct referent when no distractor is present and successfully correcting itself from the distractor when the egocentric choice fails.

There is one trial that breaks the pattern. In a scene with two rabbits, one stuffed and one made of chocolate, the description "the bunny" applies to both of them equally well. In this case, ARR notes the ambiguity before going on to correct its interpretation to the visible bunny. In another study, Keysar et al. (2003) constructed their stimuli specifically to avoid this type of ambiguity.

Related Work

Many other computational models of reference resolution have been proposed over decades of research, beginning with Winograd's (1972) SHRDLU system, which used procedural semantics to interpret referents in a block world domain. More recent work has focused on robotics (Chai et al., 2014; Williams & Scheutz, 2015), visual scenes (Gorniak & Roy, 2004; Kennington & Schlangen, 2017), and multimodal input (Chai, Hong, & Zhou, 2004; Kehler, 2000).

Our approach has the most in common with Chai et al.'s (2014). Their system handles reference resolution using inexact graph matching between a *vision graph*, a representation of the observed scene, and a *dialogue graph*, a graph of entity mentions and coreference. Their work explicitly deals with the issue of common ground and its effect on human-robot interactions, particularly when mismatched sensory capabilities put the robot at a disadvantage.

Unlike their system, ARR is not limited to visual characteristics, and the graph-matching process it uses (SME) is believed to be employed in multiple areas of cognition. Furthermore, the corrective process outlined in this paper accounts for errors in human reference resolution which Chai et al.'s system does not address. On the other hand, their system has the advantage of working more naturally with sensor data and quantitative scales, where ARR must rely on sketched data and symbolic representations.

Discussion

In this paper, we have extended the ARR model of reference resolution to account for the pattern of egocentric bias and error correction observed in studies like Epley et al. (2004). Resolution consists of an initial call to analogical retrieval that does not take perspective into account, followed by a more elaborate corrective process which performs ToM inference, suppression, and re-representation as needed. A simulation using Epley et al.'s stimuli shows that our model behaves as predicted, matching the error pattern of human subjects. This provides preliminary evidence for ARR as a plausible cognitive model.

Analogy provides a useful starting point for further investigation. If analogical retrieval is at the heart of reference resolution, the claims of Structure-Mapping Theory should apply. In particular, human subjects should prefer referents that are structurally similar to the description, favoring systems of relations over surface attributes. Designing an experiment to test this hypothesis would shed light not only on reference resolution but the use of analogy in human similarity judgments.

Another open question is what type of analogical retrieval is at work during reference resolution. Our model currently uses MAC/FAC to retrieve the entity that is most similar to the description, but an alternative approach is SAGE-WM (Kandaswamy, Forbus, & Gentner, 2014), which models working memory as a case library ordered by recency. SAGE-WM performs retrieval by searching for the most recent case with a strong structural match to the probe. This model better captures referential phenomena that rely on temporal ordering, like anaphora, but sacrifices MAC/FAC's clean handling of ambiguity. Determining how these retrieval strategies interplay in online reference resolution is an area for future work.

Further study is also needed to clarify the specific steps involved in the corrective process. In this paper, we have proposed one possible form of correction which explains the findings of Epley et al. (2004). More work will need to be done to extend this process to correction that goes beyond simple ToM reasoning. More sophisticated knowledge about the speaker, contextual factors such as task suitability, and other forms of re-representation should be taken into account.

In particular, it is unclear from the available data whether suppression operates incrementally or in batch. Incremental operation would mean that only the selected referent is suppressed when a common ground violation occurs. Batch operation would mean that, once a common ground violation is detected, all entities are checked against the common ground and the ones the speaker does not know about are ruled out in bulk. Empirically, incremental operation would suggest that, when faced with multiple distractors, people tend to look at each one in turn before settling on the appropriate referent. On the other hand, batch operation would suggest that people skip past subsequent distractors,

an effect that should persist to consecutive trials over the same scene. Further extensions to the testing paradigm proposed by Keysar et al.'s (2000) should shed light on this matter.

Reference resolution is not a new problem in the fields of linguistics, psycholinguistics, or artificial intelligence. Formal theories of reference, psychological evidence, and practical algorithms all shed different kinds of light on one of the fundamental processes of language. We believe that Analogical Reference Resolution can help bridge the gap between these traditions, providing a model that accounts for human experimental data and can serve as a component in larger AI systems.

Acknowledgements

We would like to thank Nicholas Epley, Carey Morewedge, and Boaz Keysar for providing the stimuli used in this paper. We would also like to thank Dedre Gentner for her helpful comments. This research was supported by grant FA9550-16-1-0138 from the Air Force Office of Scientific Research.

References

Chai, J. Y., Hong, P., & Zhou, M. X. (2004). A probabilistic approach to reference resolution in multimodal user interfaces. *Proceedings of the Ninth International Conference on Intelligent User Interfaces* (pp. 70-77). Madeira, Funchal, Portugal: ACM.

Chai, J. Y., She, L., Fang, R., Ottarson, S., Littley, C., Liu, C., & Hanson, K. (2014). Collaborative effort towards common ground in situated human-robot dialogue. *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 33-40). Bielefeld, Germany: ACM.

Clark, H. H., & Carlson, T. B. (1981). Context for comprehension. In J. Long & A. Baddeley (Eds.), *Attention and performance IX*. Hillsdale, NJ: Erlbaum.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. Joshi, B. Webber, & I. Sag (Eds.), *Elements of discourse understanding* (pp. 10-63). Cambridge, UK: Cambridge University Press.

Clark, H. H., & Carlson, T. B. (1982). Speech acts and hearers' beliefs. In N. V. Smith (Ed.), *Mutual knowledge*. London: Academic Press.

Day, S. B., & Gentner, D. (2007). Nonintentional analogical inference in text comprehension. *Memory & Cognition, 35*, 39-49.

Donnellan, K. S. (1968). Putting humpty dumpty together again. *The Philosophical Review*, *77*, 203-215.

Epley, N., Morewedge, C. K., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology*, *40*, 760-768.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19*, 141-205.

Forbus, K., Usher, J., Lovett, A., Lockwood, K., & Wetzel, J. (2011). CogSketch: Sketch understanding for cognitive science research and for education. *Topics in Cognitive Science*, *3*, 648-666.

Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2016). Extending SME to handle large-scale cognitive modeling. *Cognitive Science, 41*, 1152-1201.

Friedman, S., & Forbus, K. (2010). An integrated systems approach to explanation-based conceptual change. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)* (pp. 1523-1529).

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155-170.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*, 45-56.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, *65*, 263-297.

Gorniak, P., & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429-470.

Kandaswamy, S., Forbus, K., & Gentner, D. (2014). Modeling learning via progressive alignment using interim generalizations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*.

Kehler, A. (2000). Cognitive status and form of reference in multimodal human-computer interaction. *Proceedings of the Seventeenth National Conference on Artificial Intelligence* (pp. 685-690). Austin, TX: AAAI Press.

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

Keysar, B., Barr, D. J., Balin, J. A., & Paek, T. S. (1998). Definite reference and mutual knowledge: Process models of common ground in comprehension. *Journal of Memory and Language*, *39*, 1-20.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*, 32-38.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*, 25-41.

Klenk, M., & Forbus, K. (2007). Cognitive modeling of analogy events in physics problem solving from examples. *Proceedings of CogSci-07*. Nashville, TN.

Kripke, S. (1977). Speaker's reference and semantic reference. *Midwest Studies in Philosophy*, *2*, 255-276.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*, 551-556.

Lovett, A., & Forbus, K. D. (2017). Modeling visual problem solving as analogical reasoning. *Psychological Review*, *124*, 60-90.

Nakos, C., Rabkina, I., & Forbus, K. (in press). An Analogical Account of Reference Resolution. *Advances in Cognitive Systems*.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169-190.

Rabkina, I., McFate, C., Forbus, K. D., & Hoyos, C. (2017). Towards a Computational Analogical Theory of Mind. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 2949-2954.

Williams, T., & Scheutz, M. (2015). POWER: A domain-independent algorithm for probabilistic, open-world entity resolution. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 1230-1235). Hamburg, Germany: IEEE.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function in young children's understanding of deception. *Cognition*, *13*, 103-128.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, *3*, 1-191.