Advances in Cognitive Systems 9 (2020) 91–107 Submitted 8/2019; published 12/2020

An Analogical Account of Reference Resolution

Constantine Nakos CNAKOS@U.NORTHWESTERN.EDU
Irina Rabkina IRABKINA@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 the problem of interpreting a referring expression to identify the speaker's intended referent. People are capable of correctly resolving referents even in the presence of ambiguous or incorrect descriptions. We propose analogy as the process underlying this ability. This paper presents ARR, an implemented computational model of analogical reference resolution and evaluates its performance on the TUNA Corpus of definite descriptions. We show that the model performs as well as an exact matching baseline on error-free descriptions and that it is more robust to errors.

1. Introduction

One of the many ways that humans use language is to refer to entities in the world. *Referring expressions* take a variety of forms depending on the context of the utterance, the knowledge of the speaker and hearer, and the cognitive status of the intended referent. Reference has many subtleties that have occupied linguists for over a century. Chief among these puzzles is the issue of *definiteness*, the intuitive but elusive property shared by noun phrases that mark known information and commonly seen in definite descriptions such as "the woman in the corner" (Birner, 2012). Formal theories of definiteness link it to *uniqueness* (e.g., Russell, 1905), where a definite description is true of a uniquely identifiable entity, familiarity (e.g., Christopherson, 1939), where a definite description marks an entity familiar to the hearer, or *individuability* (e.g, Birner & Ward, 1998), where a definite description marks an entity as *distinguishable* within the discourse model.

But one aspect of the way humans use definite descriptions has remained hard for theorists to explain: the *near miss problem* (Donnellan, 1968). Near misses are definite descriptions that do not perfectly match their intended referent but that are close enough to be understood by the hearer. For example, consider the following exchange, adapted from Strawson (1952):

A: "Did you hear about the man who jumped off Swift Hall?"
B: "He didn't jump. He was pushed."

A's description is false to facts. There is no man who jumped off Swift Hall. Yet B was able to determine A's intended referent and correct A's description in spite of the mismatch. Moreover, B was able to do so in a world of potential distractors: other men, other jumping events, and other events involving Swift Hall, not to mention the countless entities known to B. This comes naturally to people, but reference resolution is a subtle problem that involves world knowledge, semantic similarity, and context.

This example and others like it show that the mechanism humans use to resolve referring expressions must be more flexible than exact matching. Formal theories of reference have a hard time accounting for near misses, especially ones rooted in formal logics where an exact match between referent and description is assumed. Linguists like Donnellan (1968) have characterized the types of near misses and when they arise but have not proposed a mechanism by which people resolve them. The issue remains open, a capability of human language understanding that is not reflected in linguistic formalisms.

At the same time, reference resolution is a practical problem for computer systems that interact with people using natural modalities. As natural language interfaces become more sophisticated, it becomes increasingly important to interpret a user's utterances in relation to a shared context. Reference resolution is therefore relevant to human-robot interaction and human-computer interaction. Prior work in these fields has modeled it using Bayesian techniques (Kennington & Schlangen, 2017; Tellex et al., 2011), graph matching (Liu, Fang, & Chai 2012), best-first search with uncertainty (Williams & Scheutz, 2015), and search over experiential knowledge (Mohan, Mininger, & Laird, 2013). Related work in symbol grounding (Harnad, 1990; Gorniak & Roy, 2004; Steels & Hild, 2012) has tackled the problem of mapping between a robot's sensor values and symbolic representations suitable for language interpretation.

We propose *analogy* as the mechanism behind reference resolution. Building on Gentner's (1983) *Structure-Mapping Theory* (SMT), we describe how analogical matching and retrieval can be used to resolve a definite description against a set of candidate referents. This provides a notion of semantic similarity that takes relational information into account. SMT is domain independent and can thus be used with any symbolic, structured representation of entities and associated descriptions.

This paper presents ARR, an implemented computational model of *analogical reference resolution*, and evaluates it on a preexisting data set of definite descriptions. We show that the model achieves comparable performance to an exact matching baseline on error-free descriptions and that it is more robust to errors. The model does not need to be trained. It is extensible and can be modified to incorporate greater contextual information, account for other aspects of reference, and fit into a broader model of language understanding. In the next section, we describe the analogy process in greater detail. In Section 3, we describe our model of analogical reference resolution, while Section 4 evaluates our model, presents its results, and discusses their implications. In Section 5, we recap related work on reference resolution. We conclude and discuss future work in Section 6.

2. Background

Analogical reference resolution builds on a family of cognitive process models based on Gentner's (1983) *Structure-Mapping Theory* (SMT), which models analogy as a mapping process between structured descriptions of entities or situations. Similarity judgments depend on the strength of an analogical mapping, including its depth and connectedness. Structure mapping provides a plausible mechanism by which people judge the semantic similarity between a definite description and a potential referent. It is independently motivated and able to operate over arbitrary structured representations, making it domain independent and capable of working with a variety of semantic knowledge. SMT has been used to model other linguistic phenomena, such as word sense disambiguation (Barbella & Forbus, 2013), learning linguistic constructions (McFate, Klein, & Forbus, 2017), and multimodal knowledge capture (Lockwood & Forbus, 2009).

SMT operates over *cases* that contain structured descriptions of an entity or situation. Each case consists of a set of *expressions* that capture relations (e.g., “X is on top of Y”) or attributes (e.g., “X is green”, “Y is a table”). Expressions can take either atomic terms or other expressions as arguments, allowing the model to handle higher-order semantic information, such as causality. The model matches from a *base* case to a *target* case by searching for a consistent set of correspondences between the expressions of the base and the expressions of the target, as well as the entity mappings that arise from them. A *mapping* consists of a set of structurally consistent correspondences, along with a *structural similarity score* encoding the strength of the match, and a set of *candidate inferences*, expressions in the base that do not correspond to any in the target but that are suggested by the structure.

SMT has been implemented computationally in Forbus et al.’s (2016) Structure-Mapping Engine (SME). This performs structure mapping between cases represented in predicate calculus by using a greedy merge algorithm that finds near-optimal mappings in polynomial time. The system can interface with existing knowledge bases, making it a useful tool for testing and applying SMT and letting it serve as a central component of the Companion cognitive architecture (Forbus & Hinrichs, 2017).

Forbus, Gentner, and Law’s (1995) MAC/FAC (“many are called/few are chosen”) is a model of *analogical retrieval* built on top of SME. It searches a *case library* for the cases that best match a given *probe case*. Retrieval consists of two processing stages. The first, MAC, performs a fast, shallow match between the probe case and every case in the case library. This does not take structure into account; it uses a coarse measure of similarity to filter the case library down to a manageable size. The second stage, FAC, then uses SME to perform analogical matching between the probe case and each of the cases chosen by MAC. This returns the match with the highest structural similarity score. When there are other cases whose score is close to that of the best match, it returns them as well.

For the purposes of ARR, the probe case is a representation of the definite description and the case library is a set of cases representing the entities in the hearer’s discourse model. The contents of the case library depend on context, salience, and other factors, but we assume that it contains the speaker’s intended referent (if it is known to the hearer) and a set of plausible distractors. MAC/FAC uses the probe case to retrieve the entity that is most similar to the definite description, which it takes to be the referent. The mapping between the description and the retrieved case can be used for further reasoning, such as clarification or correction.

3. A Model of Analogical Reference Resolution

In this section, we lay out the ARR model of reference resolution. We begin with an overview of the system that explains how it uses MAC/FAC to perform resolution. We then list the potential outcomes of this process and what they mean for interpreting a description. Finally, we discuss the role of analogy in the model and the advantages it offers.

Analogical reference resolution treats the interpretation of definite descriptions as a process of analogical retrieval. The inputs are a probe case containing the semantics of a definite description and a case library containing representations of the possible referents. The ARR model runs MAC/FAC to retrieve the cases that have the highest structural similarity to the description. As MAC/FAC always returns at least one match, no matter how weak, ARR uses a cutoff threshold to filter matches that are unlikely to be the true referent. The resulting matches, if any, are analyzed to determine whether the reference succeeded or how it failed. A successful reference can be integrated into the situated interpretation of the speaker's utterance. A failed reference can trigger action on the part of the hearer to resolve the issue, say by requesting clarification from the speaker or inferring the existence of an unknown entity.

3.1 Outcomes of the Resolution Process

We have identified four potential outcomes of the retrieval process, depending on the number of matches found and the strength of their scores. These outcomes correspond to the intuitive ways an act of reference can succeed or fail in natural dialogue. The outcomes listed here are a set of defaults. Depending on the context, linguistic theory, or computational application, the categories below can be merged or split.

Figure 1 shows the potential outcomes of ARR for a sample description (a). The probe case contains the semantics of the definite description. The case library contains representations of each entity in the discourse model. For example, the stored case for the referent in Figure 1 would contain facts stating that the man is young, dark haired, and wears glasses. Note that the probe is derived from a linguistic description, which is typically a subset of the properties known for the intended referent. Cases in the library consist of all known properties of each entity, although the approach should also apply to relational descriptions.

An *exact match* occurs when MAC/FAC returns a single case that matches the description perfectly, indicating that the description was sufficient to uniquely identify an entity in the hearer's discourse model. The retrieved case may contain additional information not found in the probe, since the hearer will likely know more about the referent than has been conveyed in that description, but the information that is in the probe must be correct. In SMT terms, there should be no candidate inference from the probe to the retrieved case that contradicts a fact in the retrieved case. For example, in Figure 1(b), the man shown is the only entity in the case library that matches the description, so it is the one retrieved.

Reference fails when MAC/FAC returns *no matches* above the specified threshold. Either the intended referent is not present in the discourse model or the description of it is so malformed as to be useless. Coping with this situation is a difficult problem in its own right. The hearer may interpret the speaker's description as an introduction of a new entity, just like an indefinite description, and update the discourse model accordingly.[1] The hearer may also initiate a search for matching entities that might be inferrable from context but were not stored explicitly in the discourse model during the first pass. However, without a way to assume or acquire the missing information, the lack of a suitable match simply means that reference has failed. For example, in Figure 1(c), no entity in memory matches the description closely enough, so nothing is retrieved.

[1] This use of the definite introduction is more common in narrative prose. "The detective stood under the streetlight" is a valid opening line for a novel, even though no detective or streetlight has been mentioned before.

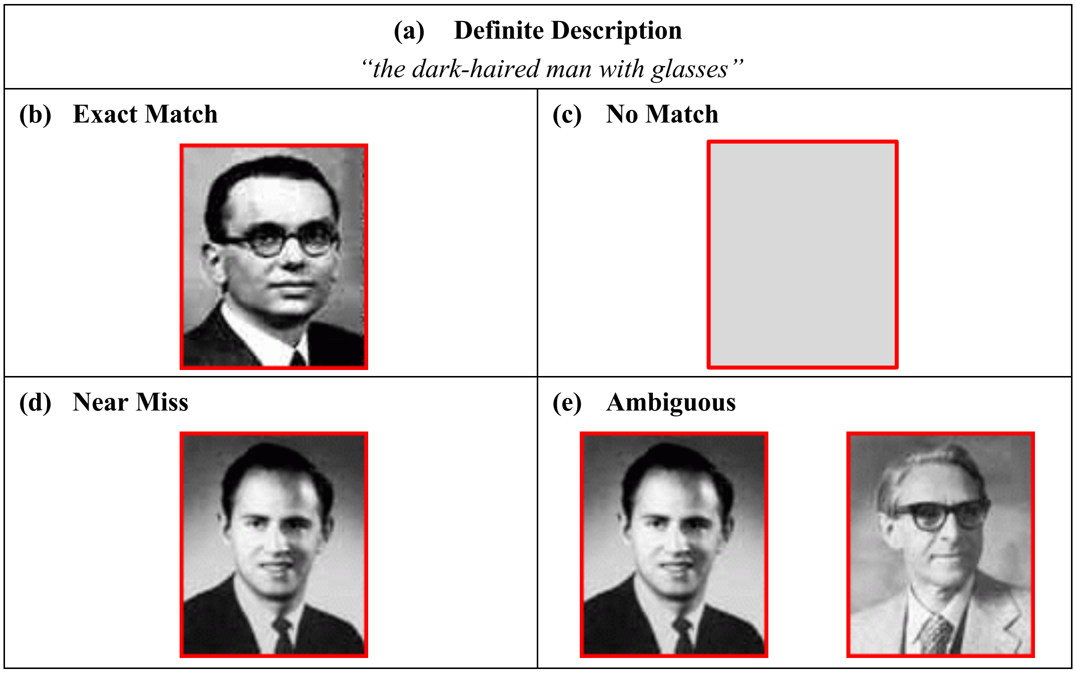

Figure 1. For a given (a) definite description, examples of (b) an exact match, (c) no match, (d) a near miss, and (e) ambiguous retrievals. Images are for illustrative purposes only. Cases consist of predicate calculus representations of the speaker's description (probe case) or entities in the discourse model (case library).

Near misses[2] occur when MAC/FAC returns a single case whose score clears the threshold, but there are discrepancies between it and the description given. For example, a near miss would occur if the speaker says "the man who jumped off Swift Hall" but the hearer only knows of a man who was pushed. Whether a near miss should be considered a success or a failure will depend on judgment. The hearer may decide to tacitly accept a near miss, offer a correction, or request clarification to ensure the discrepancy is not the result of some deeper misunderstanding. For example, in Figure 1(d), the closest match to the description is a dark-haired man with no glasses. The discrepancy may prompt further action from the hearer.

Ambiguity occurs when MAC/FAC returns multiple matches with similar scores that all clear the threshold. Because there are multiple potential referents that fit the description, the hearer cannot distinguish among them without additional information. This is the classic case of referential ambiguity. For example, the phrase "the ball" applies just as well to a red ball as to a blue one. The speaker may have intended to refer to either. The ambiguity is resolved by requesting clarification, making an assumption based on a principle like salience, recency of mention, or suitability for a given task, or allowing reference to fail. For example, in Figure 1(e), there are two entities in the case library that nearly match the description: a dark-haired man without glasses and a light-haired man with glasses. The ambiguity would typically prompt the hearer to ask for clarification, since the speaker could have meant either entity.

[2] Although conceptually similar to Winston's (1970) definition of near misses in the context of learning, the use of the term "near miss" here refers specifically to the linguistic phenomenon described by Donnellan (1968).

3.2 Role of Analogy in ARR

We should also explore the properties of analogical mapping and retrieval that are useful for reference resolution. In particular, here we look at properties that lead to human-like judgments of semantic similarity, facilitate the kinds of error correction seen in people, and contribute to ARR's use as a component in natural language systems.

Exact matching. Chief among the useful properties of analogical retrieval is that it includes exact matching as a special case. When there is one exact, unambiguous referent, ARR will behave no differently than traditional, logic-based reference resolution. The referent is identified through analogical matching rather than logical inference, but the result is the same. This makes it broadly compatible with most formal theories of reference. All that changes is the mechanism by which the description is checked against the referent, accounting for human robustness to error while leaving the behavior on nondeviant cases unchanged.

Nested descriptions. Due to SME's use of structural descriptions, ARR can handle nested definite descriptions the same way it handles descriptions of a single entity. For example, "the ball on the table" would translate to a case containing two entities linked by the *on* relation, one with the attribute *ball* and the other with the attribute *table*. The case for the ball in the discourse model would contain all known, relevant facts about it, including the fact that it was on the table. The mapping produced by SME would align the discourse variables representing the ball and the table in the nested description with the entities denoting the ball and the table in the retrieved case, resolving the two descriptions simultaneously.[3]

Systematicity. Two particular properties of SME help explain human intuitions regarding near misses. The first is the systematicity preference, where an analogical mapping will be preferred if it is deep and structural rather than broad and shallow. Systematicity predicts that hearers will favor near misses that have relational structure in common with the definite description over ones that just share surface features. For example, a hearer would be more likely to interpret the referent of "the man who jumped off Swift Hall" as a woman who jumped off Swift Hall than a man who painted Swift Hall. The former changes an attribute (i.e., the gender of the jumper) while keeping the relation between entities (i.e., the type of event) the same. The latter keeps the attribute the same but changes the relation. The systematicity preference accounts for the human capacity to pick out meaningful near misses rather than superficial ones.

Tiered identicality. The other property that helps explain near misses is tiered identicality. By default, SME only allows matches between identical attributes and relations. The identicality constraint serves in part to limit computational complexity, as otherwise all expressions could match to all other expressions with the same number of arguments. It also captures the intuition that, when it comes to analogy, it is not meaningful to map arbitrary pairs of relations, only ones that share a meaning. But there are times when identicality is too strict. Tiered identicality allows

[3] A subtlety arises if the case for the table also contains the fact that the ball is on it, in which case MAC/FAC may return the table instead. As long as the mappings are the same, it does not matter which case is retrieved. This suggests that, if two cases score similarly to one another, they should be checked to ensure they have different mappings before ambiguity is declared.

matches between nonidentical attributes or relations as long as they are ontologically related. Returning to the Swift Hall example, SME would match jumping off a building to being pushed off because they are semantically close, but it would not match jumping off a building to an arbitrary event with the same entities.[4] Tiered identicality explains how a hearer can infer the intended referent from an incorrect relational description without allowing implausible matches.

Candidate inferences and alignable differences. Another benefit of using analogical retrieval is that the mapping it produces makes it easy to extend hearer knowledge and identify sources of error. The candidate inferences from the retrieved cases to the probe include facts known about the referent that were not present in the description. These are commonplace and simply reflect the fact that humans know more about the world than what is conveyed in language. Candidate inferences from the probe to the retrieved case provide new information about which the hearer is not aware. For example, "the green fruit on the table" may inform the hearer that the green object on the table is a fruit if this was not known already. These inferences make it natural to incorporate new knowledge into the appropriate case for an entity. Alignable differences (Markman & Gentner, 1993) occur when candidate inferences for the probe and for the retrieved case contradict each other. This indicates that either the speaker's description or the hearer's knowledge is incorrect and that clarification is needed. Again, the nature of analogical mappings makes it easy to identify and articulate where the error occurred.

Ambiguity. The ARR model incorporates a subtler notion of ambiguity than uniqueness-based approaches. In addition to the classic type of ambiguity, where an underspecified description exactly matches two or more referents, analogical retrieval can return multiple near misses, none of which match the description exactly but all of which are plausible referents of similar strength. For instance, in the Swift Hall example, suppose there was a man who jumped off Mudd Hall and a man who was pushed off Swift. The speaker could have meant the former, mistaking the location of the event, or the latter, mistaking the type of event. An even subtler case arises when one of the returned matches is an exact match and the rest are near misses. Whether reference succeeds in this case is indeterminate.

4. Empirical Evaluation

Reference resolution is a task where perfect accuracy is possible when the descriptions are error-free and uniquely identifying. Under these circumstances, an *exact matching* system will always be able to pick out the intended referent. However, this is not the case when the descriptions are erroneous. An exact matching system has no way to recover from an error in a description. On the other hand, ARR allows for partial matching and can accommodate errors. As long as the intended referent is the best fit for the description, ARR should be able to identify it.

Therefore, to evaluate our system's performance, we conducted two experiments comparing it to an exact matching baseline. In Experiment 1, we tested it on a data set of correct, uniquely identifying descriptions to show that, like the baseline, it performs accurately on error-free inputs.

[4] This example assumes a relational representation of events with an ontology to support them. In practice, a neo-Davidsonian representation (Parsons, 1990) that treats events as entities is more flexible, but it misses out on the full structural strength of SME.

In Experiment 2, we tested its robustness to noise by altering the contents of the descriptions, simulating a near miss situation. For both experiments, we used the TUNA Corpus (Gatt, van der Sluis, & van Deemter, 2007), described below. Although our experiments test ARR's ability to perform partial matches, they do not test its preference for the relational matches that SME supports. We leave testing this aspect of the model for future work.

4.1 The TUNA Corpus

For our experiments, we evaluated our system on the TUNA Corpus (Gatt et al., 2007), an existing data set originally designed as a benchmark for Referring Expression Generation (REG) systems. Each item in the corpus consists of a set of entities and a human-authored description that refers to one of them. The researchers presented participants with a 3×5 grid containing pictures of seven entities and asked them to describe the one marked. They then annotated the participants' responses with the attributes mentioned, providing a set of attributes intended to pick out a unique target among a field of distractors.

The TUNA Corpus contains 360 descriptions for the *people* domain, where the entities are black-and-white photographs of people, and 420 descriptions for the *furniture* domain, where the entities are digitally generated pictures of household items.[5] A stimulus from the people domain can be seen in Figure 2(a). Entities are characterized by simple visual characteristics such as color or orientation. The corpus contains annotations for the attributes most commonly used by experiment participants. Additional characteristics found in descriptions are marked as "other". References to an entity's relative location are marked with the corresponding absolute position.

For use with ARR, we converted the XML files of the corpus into predicate calculus notation, as shown in Figure 2(b). Each attribute-value pair was mapped to a unary predicate (e.g., `<ATTRIBUTE NAME="age" TYPE="literal" VALUE="young"/>` became `age-Young`). We did not include attributes that lacked semantic information (i.e., "other" or "unknown"). This resulted in 14 descriptions from the people domain and three from the furniture domain containing no attributes, so we removed these examples from the data set. We gave each entity and description its own case, as can be seen in Figure 2(c). To test the resolution of a description, we queried its case against the appropriate set of entities using MAC/FAC and checked whether the retrieved entity was the correct one. Because the data set consists only of plausible distractors, we used a cutoff threshold of zero. In more realistic applications, where the entities are more varied and the cases contain more information, we may need to set the threshold higher to filter out spurious matches.

4.2 Experiment 1

Table 1 shows the results of the first experiment. The baseline system attains 92 percent accuracy on the people domain and 78 percent accuracy on the furniture domain. As the baseline represents a literal interpretation of the description, the examples that it gets wrong are by definition the ones to which the description does not uniquely refer. These non-uniquely identifying descrip-

[5] The corpus contains a similar number of plural descriptions. For the sake of simplicity, we only examine the singular.

(a)

(b)

```
<ENTITY ID="9" IMAGE="Godel.gif" TYPE="target">
<ATTRIBUTE NAME="hairColour" TYPE="literal"
  VALUE="dark"/>
<ATTRIBUTE NAME="hasBeard" TYPE="boolean"
  VALUE="0"/>
<ATTRIBUTE NAME="hasGlasses" TYPE="boolean"
  VALUE="1"/>
```

(c)

```
(isa s101t26-9 Entity)
(hairColor-Dark s101t26-9)
(noBeard s101t26-9)
(glasses s101t26-9)
```

Figure 2. Example stimulus from the TUNA Corpus: (a) a picture from the corpus in the people domain; (b) a snippet of its XML description; (c) the symbolic representations used in our experiments.

tions arise from a combination of error by the annotator, error by the author, and reliance on attributes that were not cataloged in the corpus.

Table 2 shows the types of errors made by the baseline system and the ARR model. For the baseline, ambiguous retrievals represent underspecified descriptions that match more than one referent (including the intended one), while no retrieval indicates that the description failed to apply to any referent. These errors account for the majority of the baseline system's mistakes. ARR performs better than the baseline, with 94 percent accuracy on the people domain and 80 percent accuracy on the furniture domain. However, neither improvement is statistically significant at the $p < 0.05$ level. The slight improvement comes from a few erroneous descriptions that are close enough to their referents for analogy to match, even though the literal description is incorrect. ARR performs strictly better on this data set, as it does not miss any examples that the baseline gets correct.

Further improvement can be seen in the second column of Table 1, which shows the systems' accuracy when taking ambiguity into account. For this calculation, a retrieval was considered correct if it contained the intended referent, whether or not a unique referent was found. This corresponds to situations in which a human would ask for clarification, so the system's ability to identify likely referents matters for downstream applications. Here ARR's improvements over the baseline when allowing ambiguity were significant in both domains ($p < 0.01$).

4.3 Experiment 2

For our second experiment, we permuted the descriptions to test the systems' robustness to noise. In principle, ARR provides a way to identify a referent even when inaccurate descriptions are present, as demonstrated by its improvement over the baseline in Experiment 1. To simulate a near miss situation, we performed two types of edit operations: *inserting* an incorrect attribute into a description (e.g., `color-Green` when the target is a red chair) and *substituting* a correct attribute with an incorrect one (e.g., replacing `age-Young` with `age-Old`). We also examined underspecification by *deleting* attributes from the description.

Table 1. Accuracy for ARR and the baseline system in the People and Furniture domains.

	Exact Accuracy		Accuracy w/ Ambiguity	
	ARR	**Baseline**	**ARR**	**Baseline**
People	0.94	0.92	0.99	0.95
Furniture	0.80	0.78	0.98	0.92

For each of these operations, we constructed modified sets of descriptions with one, two, and three edits. Where an edit would change the entire structure (e.g., substituting two attributes when there had been only two), we removed the example. Figure 3 shows the distribution of attribute counts. Note that the resulting descriptions are not guaranteed to help identify the correct referent. Unlike near misses in the wild, the errors are introduced at random rather than through psychologically plausible mistakes. Multiple errors can quickly overwhelm the correct attributes, whereas the restricted nature of the domain means the modified structure has a better chance of matching the wrong referent. Still, the modifications serve as a useful stress test for our system.

ARR's performance in Experiment 2 can be seen in Figure 4, with accuracy for exact retrievals in the left column and accuracy allowing for ambiguity in the right. The baseline system performed as expected, failing to retrieve the correct referent for nearly all descriptions modified by substitution or insertion.[6] For deletions, its performance closely matched that of ARR. For these reasons, baseline results are omitted from the figure.

We used paired samples two-tailed *t*-tests to compare ARR with the baseline for each operation, number of edits, and scoring measure. ARR performed significantly better than the baseline system for $n = 1, 2, 3$ insertions and $n = 1, 2$ substitutions on both accuracy measures ($p < 0.01$ in all cases). For $n = 3$ substitutions, ARR performed significantly better in the furniture domain when allowing for ambiguity ($p < 0.01$), but not when counting exact retrievals or for either measure in the people domain ($p > 0.05$). For deletions, the results are mixed. In the people domain, the only significant difference between the two systems is for one deletion when allowing ambiguities ($p < 0.05$; $p > 0.05$ in all other cases). For the furniture domain with exact retrievals, ARR outperforms the baseline for $n = 1$ ($p < 0.05$) but not for $n = 2$ or 3 ($p > 0.05$). When allowing for ambiguities, ARR does better than the baseline for all values of n ($p < 0.01$ for $n = 1, 2$; $p < 0.05$ for $n = 3$).

Table 2. Retrieval counts by type in Experiment 1 for ARR and the baseline system.

	People		Furniture	
	ARR	**Baseline**	**ARR**	**Baseline**
Exact Retrieval	324	319	333	326
Ambiguous Retrieval	20	11	77	59
Incorrect Retrieval	2	2	7	6
No Retrieval	0	14	0	26

[6] In the single-substitution condition, the system retrieves four correct referents for the furniture domain and three for the people domain. These results are anomalies caused by the erroneous descriptions found in Experiment 1.

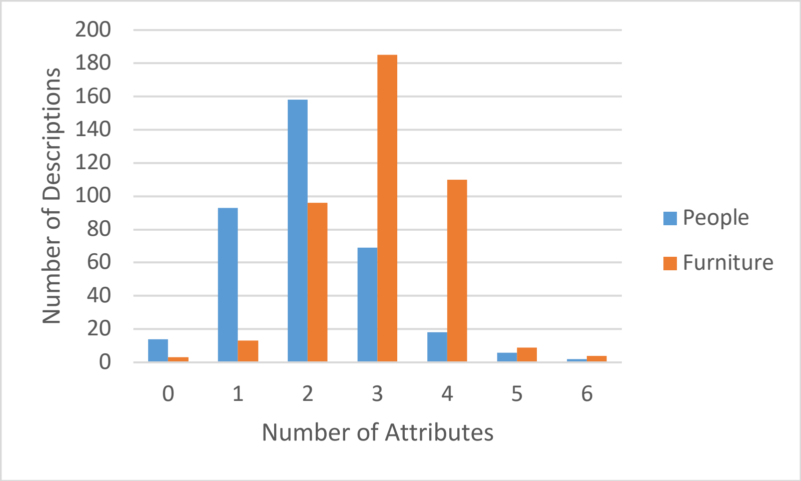

Figure 3. Frequency distribution of number of attributes per description in each domain.

4.4 Discussion

Our experiments confirm that ARR performs as well as an exact matching baseline on accurate descriptions and outperforms it on inaccurate ones. Experiment 1 showed that ARR performed strictly better than the baseline on a corpus of definite descriptions. Although in principle both systems should have done perfectly, there were naturally occurring errors in the data which affected their behavior. ARR showed its robustness to noise by handling some of these errors correctly, leading to a statistically significant improvement in accuracy when taking ambiguity into account. Moreover, it correctly identified all referents that the baseline did, confirming that it is an effective model when dealing with error-free descriptions.

Experiment 2 expanded on these results by showing how artificially introduced noise affects ARR's performance. Performance is reasonable for a single deletion or insertion, with up to 48 percent accuracy in the furniture domain, but it declines rapidly as more errors are introduced. Substitution, which combines a deletion and an insertion operation, causes an even sharper decrease in accuracy. This behavior follows from the task definition, where the addition of errors quickly causes the noise to outweigh the signal. An extreme case of this can be seen in the baseline results. Introducing just one incorrect attribute prevents an exact match, making the baseline system extremely brittle. ARR's flexibility more closely resembles the way humans cope with near misses, retrieving the closest match even if it is inexact.

Deletion behaves differently from the other operations, as removal of attributes leads to an underspecified description that may no longer be unique. For exact retrieval, this results in decreased performance for both systems as the number of competing referents increases. For ambiguous retrieval, the performance remains at ceiling. Unlike insertion and substitution, deletion does not produce incorrect references, only incomplete ones.

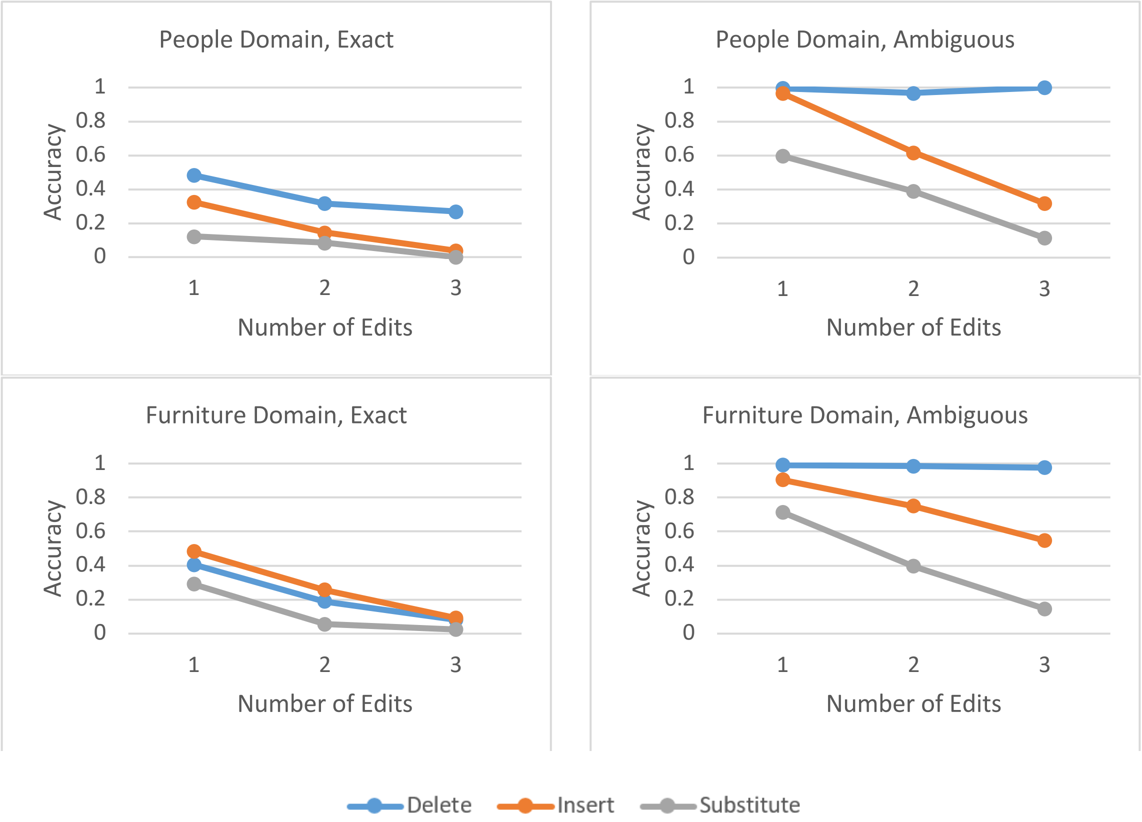

Figure 4. ARR's accuracy (exact and ambiguous) in each domain, per edit operation and total number of edits. Baseline results are omitted for all conditions because of near-zero accuracy on insertion and substitution and overlap with ARR's performance on deletion.

The contrast between performance on exact and ambiguous retrieval for each system sheds light on the type of referential errors introduced by insertion and substitution. Because the modified descriptions no longer apply to the intended referent, the baseline system shows no improvement when ambiguous retrieval is allowed. This shows that the type of ambiguity introduced is not underspecification, but rather ambiguity among near misses. ARR handles this type of ambiguity robustly, retrieving the correct referent 54 percent of the time in the furniture domain even with the insertion of three incorrect attributes. Because situated reference generally occurs in contexts where clarification is possible, the ability to retrieve a correct referent, even nonuniquely, is valuable.

5. Related Work

Reference resolution is a well-studied problem in the field of computational linguistics. AI systems that address this problem date back to Winograd's (1971) SHRDLU and its ability to identify situated referents using procedural semantics. Since then, reference resolution has remained an active area of research, with systems that handle it for collaborating with robots

(Chai et al., 2014), processing visual scenes (Gorniak & Roy, 2004), and interpreting multimodal user input (Chai, Hong, & Zhou, 2004; Kehler, 2000). Here we review the three most pertinent models of reference resolution and contrast ARR with them.

Chai et al. (2014) address the problem of establishing *common ground* (Clark, 1996) between human and robot interlocutors that have mismatched perceptual capabilities. Their system uses an inexact graph matching algorithm (Liu, Fang, & Chai, 2012; Liu et al., 2013) to ground referring expressions against a visual scene. The algorithm matches a *vision graph*, representing the scene as observed by the robot, and a *dialogue graph*, representing entities mentioned in the discourse so far. In each graph, nodes represent entities and their attributes, and edges represent the spatial relations between them. The algorithm searches for a match that minimizes a cost function measuring the degree of difference between nodes of the two graphs. The output is a match that connects each referring expression with the object in the scene to which it most likely refers.

Of the existing models of reference resolution, Chai et al.'s comes the closest to the account presented here. Both rely on graph matching to map between a description and its potential referents, but the earlier work focuses on reference resolution over visual scenes grounded in sensor data, whereas ARR is a domain-independent model that operates over symbolic representations. ARR also benefits from an independently motivated, cognitively plausible graph-matching algorithm that does not need to be altered for this task. Structure mapping has been used to model human reasoning in a variety of domains, including tasks that combine information from visual and linguistic inputs (Lockwood & Forbus, 2009; Chang, 2016).

Another notable model of reference resolution is POWER (Williams & Scheutz, 2015), a probabilistic reference resolution system capable of handling uncertainty and references to unknown entities in an open world. This system queries consultants that track entities in relation to a specific domain, such as spatial reasoning or visual processing, and calculates the probability that a given configuration of entities satisfies a semantic constraint. It conducts a best-first search to find the most likely assignment of variables to entities. The system also posits new entities as needed to extend its knowledge of the world. POWER has been tested in several domains, both in isolation (Williams & Scheutz, 2015) and as part of the DIARC robotics architectures (Williams et al., 2019). It has been extended with the Givenness Hierarchy (Gundel, Hedberg, & Zacharski, 1993) to form GH-POWER (Williams et al., 2016), which handles a variety of referring expression types by searching over different levels of memory. Recent work has relaxed the strict precedence of this search, resulting in GROWLER (Williams et al., 2018), which handles more naturalistic uses of referring expressions.

Our domain-independent model of reference resolution also operates over a wide variety of semantic knowledge. But where POWER handles uncertainty by attaching probabilities to facts, ARR uses analogy to deal with inaccurate descriptions and knowledge.[7] The two approaches differ in their focus. POWER is a key component in robotics architectures, where different modalities of data must be integrated and where quantifiable uncertainty is the norm. ARR is a psychologically plausible account of reference resolution suitable for use with symbolic reasoning systems in which probabilistic information may or may not be available.

[7] It is worth noting that SME can take probabilities into account when computing structural similarity scores, so in principle ARR can handle probabilistic uncertainty. For the purposes of the current work, we ignore this capability and focus instead on facts that are fully believed but that may be erroneous.

Kennington and Schlangen (2017) present a generative model of reference resolution that incrementally calculates the most likely referent from the tokens in a referring expression. Their model learns probabilistic mappings between objects and properties and between properties and language, letting it flexibly adapt to the meanings of words as they are used in a particular domain. The incrementality of their model matches psycholinguistic evidence of how humans process referring expressions (Spivey et al., 2002; Tanenhaus et al., 1995), while its probabilistic nature makes it able to incorporate uncertainty about the properties of objects. Their model is language independent and has been successfully applied to German, Japanese, and English.

ARR differs from Kennington and Schlangen's model in several key respects. Where their generative model must be trained on each new domain, our model requires no training and encodes semantics symbolically rather than as probability distributions. Our approach is also domain independent, where the generative model has chiefly been applied to visual domains. While the current version of ARR makes no attempt to handle the incremental processing of referring expressions, the incrementality of SME offers a path to achieving this ability.

6. Conclusions and Directions for Future Work

In this paper, we presented a model of reference resolution that uses analogical retrieval, an independently motivated cognitive process, to match definite descriptions to their intended referents. We discussed the advantages of this approach, showed that our model performs as well as an exact matching baseline on correct descriptions, and demonstrated its robustness to errors, in particular near misses and ambiguities. Although we consider this to be a promising start, we should explore the model's behavior on a wider range of inputs, extend it to handle other types of referring expressions, and provide the scaffolding needed to integrate it with a dialogue system.

Here we tested ARR in isolation on a relatively simple, nonrelational domain using pre-annotated visual attributes and semantics. To demonstrate the model's ability to interface with other components in a cognitive architecture, we plan to integrate it with EA NLU (Tomai & Forbus, 2009), a domain-independent semantic parser built on the NextKB ontology (Forbus & Hinrichs, 2017). Integrating ARR with a semantic parser will let us experiment on more naturalistic input and test the system on tasks where its strengths, such as the systematicity preference and the ability to generate clarification dialogs, play a more important role.

In addition, we should situate ARR within a broader account of language understanding. The model depends on external processes to encode its cases and populate its case library. The corpus used in the current work included structured representations that made encoding straightforward, but for more complex domains, this may not be the case. By explicitly modeling the processes on which ARR depends, we hope to account for a wider range of referential phenomena. In particular, we plan to incorporate aspects of the Givenness Hierarchy (Gundel et al., 1993) to guide case library construction and search through memory.

The Givenness Hierarchy also suggests a mechanism for ARR to handle other types of referring expressions. Although our approach is a natural fit for definite descriptions and should extend cleanly to indefinite ones, it is less clear how analogy applies to pronouns and deixis. By specifying a series of case libraries corresponding to different levels of cognitive status, we believe that we can model the resolution of more referring expressions than with analogy alone. In this, we take the same position as Williams et al. (2016) in their work on GH-POWER.

Another benefit of adopting a more complex search strategy is the opportunity to address two factors—salience and recency—that are commonly used to help distinguish intended referents. MAC/FAC only takes structural similarity into account. The system does not consider either a case's contextual salience or how recently it was last retrieved. We can address the former by separating the case library into tiers according to salience, so that more salient cases are considered first. We can address the latter by replacing MAC/FAC with SAGE-WM (Kandaswamy, Forbus, & Gentner, 2014), a model of analogical generalization and retrieval that checks cases according to how recently they were added to memory. This would give the model a way to prioritize more recent matches, making it a better fit for reference resolution in temporal contexts.

Based on our preliminary evaluation, analogy provides a promising model for the resolution of referring expressions. This supplies the similarity judgments needed to understand near misses while using a mechanism already believed to be at work in other areas of cognition. Furthermore, its properties make it a practical choice for use in cognitive systems. Future work will extend the ARR model to more complex domains and forms of reference, including ones that involve richer, relational contents.

Acknowledgements

We would like to thank Gregory Ward for his early input on this work. This research was supported by grant FA9550-16-1-0138 from the Air Force Office of Scientific Research.

References

Barbella, D., & Forbus, K. (2013). Analogical word sense disambiguation. *Advances in Cognitive Systems*, *2*, 297–315.

Birner, B. J. (2012). *Introduction to pragmatics*. Chichester, West Sussex, UK: John Wiley & Sons.

Birner, B. J., & Ward, G. L. (1998). *Information status and noncanonical word order in English*. Philadelphia: John Benjamins Publishing.

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.

Chang, M. (2016). *Capturing qualitative science knowledge with multimodal instructional analogies*. Doctoral dissertation, Department of Computer Science, Northwestern University, Evanston, IL.

Christopherson, P. (1939). *The articles: A study of their theory and use in English*. Copenhagen: Munksgaard.

Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

Donnellan, K. S. (1968). Putting Humpty Dumpty together again. *The Philosophical Review*, *77*, 203–215.

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

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

Forbus, K. D., & Hinrichs, T. (2017). Analogy and qualitative representations in the Companion cognitive architecture. *AI Magazine*, *38*, 34–42.

Gatt, A., van der Sluis, I., & van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. *Proceedings of the Eleventh European Workshop on Natural Language Generation* (pp. 49–56). Saarbrücken, Germany: Association for Computational Linguistics.

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

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

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*, 274–307.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*, 335–346.

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* (pp. 2471–2476). Quebec City, Canada: Cognitive Science Society.

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*, 4367.

Liu, C., Fang, R., & Chai, J. Y. (2012). Towards mediating shared perceptual basis in situated dialogue. *Proceedings of the Thirteenth Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 140–149). Seoul: Association for Computational Linguistics.

Liu, C., Fang, R., She, L., & Chai, J. (2013). Modeling collaborative referring for situated referential grounding. *Proceedings of the SIGDIAL 2013 Conference* (pp. 78–86). Metz, France: Association for Computational Linguistics.

Lockwood, K., & Forbus, K. (2009). Multimodal knowledge capture from text and diagrams. *Proceedings of the Fifth International Conference on Knowledge Capture* (pp. 65–72). Redondo Beach, CA: ACM.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431–467.

McFate, C., Klein, J., & Forbus, K. (2017). A computational investigation of analogical generalization of linguistic constructions. Presented at the *Fourth Analogy Conference*. Paris, France: Institut du Cerveau et de la Moelle Epinière.

Mohan, S., Mininger, A. H., & Laird, J. E. (2013). Towards an indexical model of situated language comprehension for real-world cognitive agents. *Proceedings of the Second Annual Conference on Advances in Cognitive Systems* (pp. 153–170). Baltimore, MD.

Parsons, T. (1990). *Events in the semantics of English*. Cambridge, MA: MIT Press.

Russell, B. (1905). On denoting. *Mind*, *14*, 479–493.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, *45*, 447–481.

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

Strawson, P. F. (1952). *Introduction to logical theory*. London: Methuen & Co., Ltd.

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

Tellex, S., Kollar, T., Dickerson, S., Walter, M. R., Banerjee, A. G., Teller, S., & Roy, N. (2011). Approaching the symbol grounding problem with probabilistic graphical models. *AI Magazine*, *32*, 64–76.

Tomai, E., & Forbus, K. D. (2009). EA NLU: Practical language understanding for cognitive modeling. *Proceedings of the Twenty-Second International FLAIRS Conference* (pp. 117–122). Sanibel Island, FL: AAAI.

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* (pp. 1230–1235). Hamburg, Germany: IEEE.

Williams, T., Acharya, S., Schreitter, S., & Scheutz, M. (2016). Situated open world reference resolution for human-robot dialogue. *The Eleventh ACM/IEEE International Conference on Human Robot Interaction* (pp. 311–318). Christchurch, New Zealand: IEEE Press.

Williams, T., Krause, E., Oosterveld, B., & Scheutz, M. (2018). Towards givenness and relevance-theoretic open world reference resolution. *Proceedings of the RSS Workshop on Models and Representations for Natural Human-Robot Communication*. Pittsburgh, PA: Robotics: Science and Systems Foundation.

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

Winograd, T. (1971). *Procedures as a representation for data in a computer program for understanding natural language*. Doctoral dissertation, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA.

Winston, P. H. (1970). *Learning structural descriptions from examples*. Doctoral dissertation, Department of Electical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA.