Interactive Symbol Grounding with Complex Referential Expressions

Interactive Symbol Grounding with Complex Referential Expressions

Rimvydas Rubavicius and **Alex Lascarides**
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
rimvydas.rubavicius@ed.ac.uk, alex@inf.ed.ac.uk

Abstract

We present a procedure for learning to ground symbols from a sequence of stimuli consisting of an arbitrarily complex noun phrase (e.g. "all but one green square above both red circles.") and its designation in the visual scene. Our distinctive approach combines: a) lazy few-shot learning to relate open-class words like `green` and `above` to their visual percepts; and b) symbolic reasoning with closed-class word categories like quantifiers and negation. We use this combination to estimate new training examples for grounding symbols that occur *within* a noun phrase but aren't designated by that noun phase (e.g, `red` in the above example), thereby potentially gaining data efficiency. We evaluate the approach in a visual reference resolution task, in which the learner starts out unaware of concepts that are part of the domain model and how they relate to visual percepts.

1 Introduction

The subfield of robotics known as Interactive Task Learning (ITL, see Laird et al. (2017) for a survey) addresses scenarios where a robot must learn to adapt its behaviour to novel and unforeseen objects, relations, and attributes that are introduced into the environment after deployment. The ITL agent learns its novel task via evidence from its own actions and reactive guidance from a teacher. This paper focuses on *symbol grounding* (Harnad, 1999) in the context of ITL (Matuszek, 2018): the learner must use the teacher's embodied natural language utterance and its context to learn a mapping from natural language expressions to their denotations, given the visual percepts.

There are two challenges in learning symbol grounding models (grounders) in ITL. Firstly, in contrast to many grounders (Ye et al., 2019; Datta et al., 2019), ITL requires *incremental* learning: knowledge is acquired piecemeal via an extended interaction, and it must influence planning as and when it occurs. Secondly, previous work limits the teacher's language to bare nouns (e.g., square) or very short phrases (e.g., blue square, square above circle) (Hristov et al., 2018, 2019). But there's evidence from Dale and Reiter (1995) that speakers use complex referring expressions even when simpler ones would successfully refer. Such language creates the possibility that novel symbols—neologisms—are introduced in a context where their denotation is not designated by the teacher. In this work we study the natural language of complex referential expressions (REs) like "a blue square behind both red circles" which teachers can use when designating an object.

Our aim is for the learner to extract knowledge that improves their domain representation and state estimates—a necessary condition for successful planning. Contemporary grounders miss learning opportunities that complex REs afford: the RE example above not only entails that its referents are `blue` and `square`, but also that there exists two objects that are both `red` and `circle` and that they are `above` the designated objects, and everything else in the domain is either not `red` or not a `circle` (thanks to the meaning of `both`). Thus, a complex RE and its designation can be used to gather multiple (noisy) training exemplars (both positive and negative) for several symbols at once, even if they have not been designated.

In this work, we develop a method to integrate knowledge from interactively gathered evidence in the form of complex RE-designation pairs to aid data acquisition for a (neural) few-shot grounder. We explore the effect of such a method on data efficiency and the overall grounder's performance. A major novel component to our procedure is that we exploit the formal semantics of closed class word categories (e.g., quantifiers and negation) to boost the data efficiency of few-shot neural grounding models. Our experiments show these symbolic inductive biases are successful.

2 Related Work

Symbol Grounding. Contemporary grounders extensively utilize *batch* learning (e.g. Shridhar and Hsu (2018)). Yet, ITL requires *incremental* learning because without it the teacher guidance cannot influence the learner's inferences about plans as and when the advice is given. Further, many grounders assume that the learner starts out with a complete and accurate conceptualisation of the domain using pre-defined visual features and a known vocabulary (Kennington et al., 2015; Kennington and Schlangen, 2017; Wang et al., 2017). In ITL, both of these assumptions are unrealistic; therefore in this paper we explore models for which these assumptions don't apply. Finally, in contrast to all prior grounders, we support incremental learning when the training exemplars feature REs that are linguistically complex: e.g., "two red circles that aren't to the right of both green squares".

Representation Learning. Models for jointly learning a representation for vision and language utilize either explicit alignment via bounding boxes or instance segmentation (Lu et al., 2019; Chen et al., 2020; Tan and Bansal, 2019; Kamath et al., 2021; Yu et al., 2021), or a large-volume of weakly labeled data in the form of image-caption pairs (Radford et al., 2021). These models rely on offline learning with large datasets. This work, on the other hand, explores how to incrementally extract knowledge from few-shot learning, using sequentially observed evidence that includes neologisms.

Visual Questions Answering (VQA). This is a task of answering free-form questions about an image (Antol et al., 2015). VQA has reached impressive performance in recent years (Fukui et al., 2016; Li et al., 2020), yet VQA models struggle with out-of distribution generalization for new types of questions, requiring multi-step reasoning, with analysis revealing that they often rely on shortcuts (Jiang and Bansal, 2019; Subramanian et al., 2019, 2020). Grounded VQA models like (Yi et al., 2018) and Bogin et al. (2021) tackle these shortcomings by grounding parts of the question and then learning to compose those parts via the question's syntax to compute the answer. They thus estimate denotations of linguistic parts that are not denoted by the answer to the question. These 'compositional' models help to achieve out-of-distribution generalization for novel questions. But they lack ITL's requirement for incremental learning: model training relies on batch learning. Furthermore, while their performance is impressive, error analysis shows that it makes mistakes when language includes logical concepts like quantifiers and negation (e.g. Bogin et al. (2021) Figure 9 shows that the determiner `most` incorrectly denotes an arbitrary subset of entities). Our view is that there is little benefit in trying to learn to ground logic concepts as they are domain independent and can be interpreted using formal semantics. In our experiments, we are testing the extent to which knowing and reasoning with the logical meanings of these symbols helps incremental grounding, and in particular estimating denotations of symbols within an RE that are not designated.

Grounded Language Acquisition. This task is often realized as grounded grammar induction from image-caption pairs (Shi et al., 2019; Zhao and Titov, 2020), or as learning (neural) semantic parsers from a reward signal (Williams, 1992) in VQA (Mao et al., 2019; Yi et al., 2018) or in planning (Azaria et al., 2016; Wang et al., 2016, 2017; Karamcheti et al., 2020). There, the main objective is to learn to map natural language to logical forms, which in turn get associated with visual percepts during the learning process. This paper does not aim to learn a semantic parser. Instead, we obtain logical forms from an existing broad-coverage grammar which is hard to engineer, but is robust on lexical variation (Curran et al., 2007). Our focus instead is on exploiting the *logical consequences* of those logical forms during symbol grounding—i.e., our focus is to utilise the *interpretation* of logical forms, and in particular the truth functional meanings of close-class words like quantifiers and negation, to inform the learning of mappings from (open-class) symbols like `red` to their denotations, given the visual percepts.

Visual Reference Resolution. In previous experiments, it is often assumed that there is a unique referent in the visual scene for the given RE in the test phase (Kazemzadeh et al., 2014; Whitney et al., 2016). We aim to cope with situations where the RE has multiple referents: identifying all the referents that satisfy an RE enables efficient planning, because it affords free choice when executing certain commands—e.g., "move a square above both red circles" when there is more than one square affords choosing a control policy so that resources are optimized.

3 Background

3.1 Formal Semantics of Natural Language

Predicate logic with generalized quantifiers $\mathfrak{L}$ (Barwise and Cooper, 1981; van Benthem, 1984) is a canonical meaning representation for natural languages. $\mathfrak{L}$-sentences ϕ are constructed recursively from predicates P, terms T (i.e., variables V and constants C), logical connectives $O = \{\neg, \wedge, \vee, \rightarrow\}$ and quantifiers Q (see Table 1 column 1):

$$\begin{aligned}\phi ::= \quad & \mathtt{p}(\mathtt{t}_1, \ldots, \mathtt{t}_n) \equiv \mathtt{p}(\mathtt{t}^n) \\ & |(\neg\phi)|(\phi_1 \wedge \phi_2)|(\phi_1 \vee \phi_2)|(\phi_1 \rightarrow \phi_2) \\ & |(\mathtt{Qx}(\phi_1, \phi_2))\end{aligned}$$

where $\mathtt{p}$ is an n-place predicate, $\mathtt{t}_i \in T$ are terms, $\mathtt{Q} \in Q$ is a quantifier, and $\mathtt{x} \in V$ is a variable (in $\mathtt{Qx}(\phi_1, \phi_2)$, ϕ_1 is the restrictor and ϕ_2 the body). We also introduce λ-expressions of the form $\lambda\mathtt{x}.\phi$, where $\mathtt{x} \in V$ is free or absent in ϕ.

3.2 Model-theoretic Interpretation

$\mathfrak{L}$-sentences are interpreted using a domain model $\mathcal{M} = (E, I)$ consisting of a set of entities E (universe of discourse), and an extension function I that maps non-logical symbols $P \cup C$ to denotations (tuples of entities). For convenience, we assume $I : C \mapsto E$ is one-to-one. Variables are interpreted via an assignment function $g\colon V \mapsto E$.

The interpretation function $[\![\cdot]\!]^{\mathcal{M},g}$ specifies the *semantic value* of well-formed expressions of $\mathfrak{L}$:

$$[\![\mathtt{a}]\!]^{\mathcal{M},g} = \begin{cases} I(\mathtt{a}) & \text{if } \mathtt{a} \in P \cup C \\ g(\mathtt{a}) & \text{if } \mathtt{a} \in V \end{cases}$$

$$[\![\mathtt{p}(\mathtt{t}^n)]\!]^{\mathcal{M},g} = 1 \text{ iff } ([\![\mathtt{t}_1]\!]^{\mathcal{M},g}, \ldots, [\![\mathtt{t}_n]\!]^{\mathcal{M},g}) \in [\![\mathtt{p}]\!]^{\mathcal{M},g}$$

$$[\![\neg\phi]\!]^{\mathcal{M},g} = 1 \text{ iff } [\![\phi]\!]^{\mathcal{M},g} = 0$$

$$[\![\phi \wedge \psi]\!]^{\mathcal{M},g} = 1 \text{ iff } [\![\phi]\!]^{\mathcal{M},g} = 1 \text{ and } [\![\psi]\!]^{\mathcal{M},g} = 1$$

$$[\![\phi \vee \psi]\!]^{\mathcal{M},g} = 1 \text{ iff } [\![\phi]\!]^{\mathcal{M},g} = 1 \text{ or } [\![\psi]\!]^{\mathcal{M},g} = 1$$

$$[\![\phi \rightarrow \psi]\!]^{\mathcal{M},g} = 1 \text{ iff } [\![\phi]\!]^{\mathcal{M},g} = 0 \text{ or } [\![\psi]\!]^{\mathcal{M},g} = 1$$

$$[\![\lambda\mathtt{x}.\phi]\!]^{\mathcal{M},g} = \{e \in E\colon [\![\phi]\!]^{\mathcal{M},g[\mathtt{x}/e]} = 1\}$$

$$[\![\mathtt{Qx}(\phi_1, \phi_2)]\!]^{\mathcal{M},g} = \mathrm{Q}([\![\lambda\mathtt{x}.\phi_1]\!]^{\mathcal{M},g}, [\![\lambda\mathtt{x}.\phi_2]\!]^{\mathcal{M},g})$$

where $g[\mathtt{x}/e]$ is just like g, except $g[\mathtt{x}/e](\mathtt{x}) = e$ and Q is a specific relation between the restrictor $[\![\lambda\mathtt{x}.\phi_1]\!]^{\mathcal{M},g}$ and body $[\![\lambda\mathtt{x}.\phi_2]\!]^{\mathcal{M},g}$, as defined in Table 1 column 3. $[\![\cdot]\!]^{\mathcal{M},g}$ is directly related to satisfiability for $\mathfrak{L}$-sentences:

$$\mathcal{M}, g \models \phi \text{ iff } [\![\phi]\!]^{\mathcal{M},g} = 1$$

$$\mathcal{M} \models \phi \text{ iff } [\![\phi]\!]^{\mathcal{M}} = 1$$

where $[\![\phi]\!]^{\mathcal{M}} = 1$ iff $[\![\phi]\!]^{\mathcal{M},g} = 1$ for all g. Further, if $\mathtt{x}$ is the *only* free variable in ϕ, then $[\![\lambda\mathtt{x}.\phi]\!]^{\mathcal{M},g} = [\![\lambda\mathtt{x}.\phi]\!]^{\mathcal{M},g'}$ for all g, g'; so without a loss of generality, this is expressed as $[\![\lambda\mathtt{x}.\phi]\!]^{\mathcal{M}}$.[1] If all variables in $\mathtt{Qx}(\phi, \psi)$ are bound by quantifiers, then this $\mathfrak{L}$-sentence is true iff Q is true for all g.

Some quantifiers, like "both",[2] are presupposition triggers: "exactly two blocks are blue" is different from "both blocks are blue" in that the latter is true only if there are exactly two individuals that are blocks. We've adopted a Russellian interpretation (Russell, 1917) of these in Table 1.

3.3 Logical Forms of Referential Expressions

We now define the logical forms of REs and their interpretations with respect to a domain model $\mathcal{M}$. Noun phrases like "a block" are represented as $\langle\mathtt{\_a\_qx.block(x)}\rangle$. More generally, let $\langle\mathtt{Qx}.\phi\rangle$ be the logical form of an RE, where $\mathtt{Q} \in Q$ and ϕ is an $\mathfrak{L}$-sentence with $\mathtt{x} \in V$ being the only free variable in ϕ. The referents $\langle\mathtt{Qx}.\phi\rangle^{\mathcal{M}}$ of this logical form with respect to $\mathcal{M}$ are computed as follows:

$$\langle\mathtt{Qx}.\phi\rangle^{\mathcal{M}} = \langle\mathtt{Q}\rangle^{\pi(\mathcal{M},\phi,\mathtt{x})} \tag{1}$$

where $\pi(\mathcal{M}, \phi, \mathtt{x})$ is an $\mathcal{M}$-projection, giving a new domain model $\mathcal{M}'$ with entities $E' = E \cap [\![\lambda\mathtt{x}\phi]\!]^{\mathcal{M}}$ and $\langle\mathtt{Q}\rangle^{\mathcal{M}}$ is a quantifier referent—a quantifier-specific subset of the power set of E. Table 1 column 4 gives the list of quantifier referents.

To illustrate, consider the domain model where:

$$E = \{a, b, c, d, f\}$$
$$I(\mathtt{cat}) = \{a, b\} \quad I(\mathtt{dog}) = \{c, d, f\}$$
$$I(\mathtt{bit}) = \{(c, a), (c, b), (d, b), (f, a), (f, b)\}$$

The RE "a dog that bit both cats" has logical form $\langle\mathtt{\_a\_q\ x.\_both\_q\ y(cat(y), dog(x) \wedge bit(x, y))}\rangle$. By Equation 1, its referent is:

$$\langle\mathtt{\_a\_q}\rangle^{\pi(\mathcal{M},\mathtt{\_both\_q\ y(cat(y),dog(x)\wedge bit(x,y)),x})}$$

The semantic value of the λ-expression formed from this RE is a set of entities $e \in E$ for which the following quantifier condition is true: $\mathrm{both\_q}(R, B)$ where $R = [\![\lambda\mathtt{y.cat(y)}]\!]^{\mathcal{M},g[\mathtt{x}/e]}$ and $B = [\![\lambda\mathtt{y.dog(x)} \wedge \mathtt{bit(x, y)})]\!]^{\mathcal{M},g[\mathtt{x}/e]}$. Only $c, f \in E$ satisfy this condition, defining a new model $\mathcal{M}'$ with $E_{\mathcal{M}'} = \{c, f\}$; this leads to the set of possible referents as $\{\{c\}, \{f\}\}$, given the quantifier referent $\langle\mathtt{\_a\_q}\rangle^{\mathcal{M}'}$.

[1] This fact will be used when defining referents.

[2] This is not an English specific phenomena: Finnish *molempi* has the same condition as *both*.

quantifiers Q	surface form	condition $\mathrm{Q}(R, B)$	referent $\langle \mathtt{Q} \rangle^{\mathcal{M}}$
`_exactly_n_q`	exactly n	$\lvert R \cap B \rvert = n$	$\{A \subseteq E \colon \lvert A \rvert = n\}$
`_at_most_n_q`	at most n	$\lvert R \cap B \rvert \leq n$	$\{A \subseteq E \colon \lvert A \rvert \leq n\}$
`_at_least_n_q`	at least n	$\lvert R \cap B \rvert \geq n$	$\{A \subseteq E \colon \lvert A \rvert \geq n\}$
`_a_q`	a/an	$\lvert R \cap B \rvert \neq n$	$\{A \subseteq E \colon \lvert A \rvert \leq 1\}$
`_every_q`	all/every	$\lvert R \cap B \rvert = \lvert R \rvert$	$\{A \subseteq E \colon \lvert A \rvert = \lvert E \rvert\}$
`_the_n_q`	the n	$\lvert R \cap B \rvert = n) \wedge \lvert R \rvert = n$	$\{A \subseteq E \colon \lvert A \rvert = \lvert E \rvert \wedge (\lvert E \rvert = n)\}$
`_both_q`	both	$\lvert R \cap B \rvert = 2) \wedge \lvert R \rvert = 2$	$\{A \subseteq E \colon \lvert A \rvert = \lvert E \rvert \wedge \lvert E \rvert = 2\}$
`_all_but_n_q`	all but n	$\lvert R \cap B \rvert = \lvert R \rvert - n$	$\{A \subseteq E \colon \lvert A \rvert = \lvert E \rvert - n \wedge \lvert E \rvert \geq n\}$
`_n_of_the_m_q`	n of the m	$\lvert R \cap B \rvert = n \wedge \lvert R \rvert = m$	$\{A \subseteq E \colon \lvert A \rvert = n \wedge \lvert E \rvert = m\}$

Table 1: Quantifiers (column 1), their surface forms (column 2), condition Q between the restrictor R and body denotations B, used to compute a semantic value for $\mathfrak{L}$-sentences of the form $\mathtt{Qx}(\phi, \psi)$ (column 3); and quantifier referents $\langle \mathtt{Q} \rangle^{\mathcal{M}}$ used to compute references of the logical form of REs (column 4).

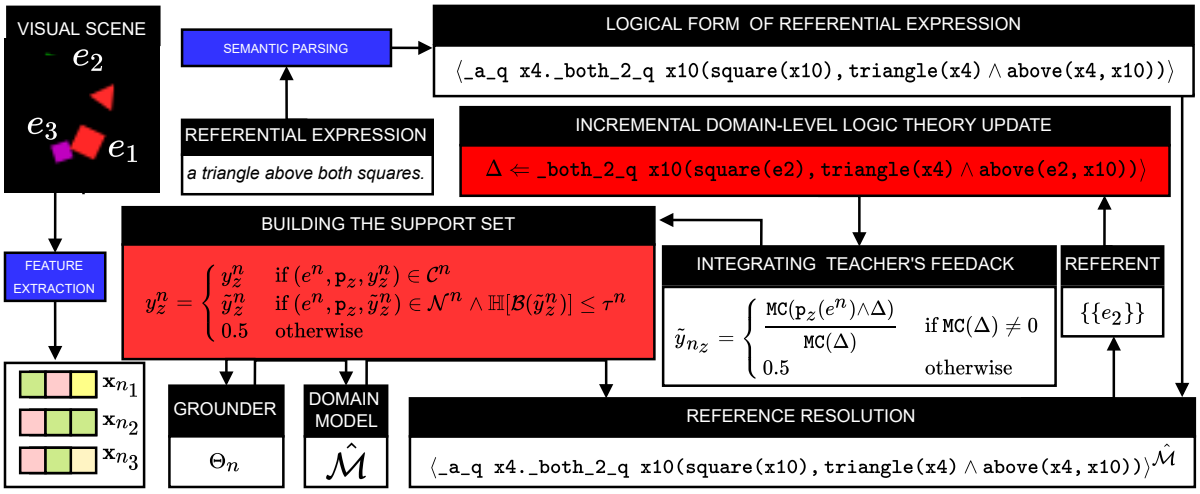

Figure 1: IGRE overview. In interaction, the learner observes an RE, which is parsed to logical form (§5.3.2) and interpreted with respect of the extracted feature vectors for denotations (§5.3.1) to perform reference resolution (§3.3) with respect to the estimated domain model $\hat{\mathcal{M}}$. In case of teacher feedback, RE and its designation is observed. This is used to build the $\mathfrak{L}$-sentence that is added to Δ to update beliefs about the underlying concept vectors (§4.3.2), which in turn are used to update the support set (§3.3), used as parameters for the grounder (§4.1). Elements in blue are pre-defined elements of IGRE while elements in red are learned through interaction.

4 Methodology

Below we present the procedure of interactive grounding with referential expressions (IGRE). The overall framework is given in Figure 1.

4.1 Grounder

Matching networks (Vinyals et al., 2016) are an extension of the k nearest-neighbour algorithm (Fix and Hodges, 1989) and has been used as a fast few-shot grounder in the ITL setting (Cano Santín et al., 2020). For predicates $P^n \subseteq P$ of the same arity n, a grounder Θ^n is parameterized by a support set $\mathcal{S}^n = \{(\mathbf{x}_i^n, \mathbf{y}_i^n)\}_{i=1}^{K^n}$, consisting of K^n pairs of feature vectors $\mathbf{x}_i^n \in \mathbb{R}^{d_n}$ for denotations $e^n \in E^n$ and concept vectors $\mathbf{y}_i^n \in [0,1]^{|P^n|}$. In $\mathbf{y}^n$, the dimension z corresponds the predicate $\mathtt{p}_z \in P^n$ and its value is the probability that $[\![\mathtt{p}_z(e^n)]\!]^{\mathcal{M},g} = 1$. Concept vectors have a one-to-one correspondence with the domain model $\mathcal{M}$.

Given a feature vector $\mathbf{x}^n$ for a denotation $e^n \in E^n$, Θ^n predicts the concept vector $\hat{\mathbf{y}}^n$, using the following inference rule:

$$\Theta^n(\mathbf{x}^n, S^n) = \sum_{i=1}^{k} \alpha^n(\mathbf{x}_i^n, \mathbf{x}^n; S^n)\mathbf{y}_i^n$$

where α^n is an attention kernel:

$$\alpha^n(\mathbf{x}_i^n, \mathbf{x}^n; S^n) = \frac{\exp\left(f^n(\mathbf{x}_i^n) \cdot h^n(\mathbf{x}^n)\right)}{\sum_{j=1}^{k} \exp\left(f^n(\mathbf{x}_j^n) \cdot h^n(\mathbf{x}^n)\right)}$$

$$f^n(\mathbf{x}^n) = \frac{\texttt{ReLU}(\mathbf{w}^n \cdot \mathbf{x}^n + b^n)}{||\texttt{ReLU}(\mathbf{w}^n \cdot \mathbf{x}^n + b^n)||_2}$$

$$h^n(\mathbf{x}^n) = \frac{\texttt{ReLU}(\mathbf{v}^n \cdot \mathbf{x}^n + c^n)}{||\texttt{ReLU}(\mathbf{v}^n \cdot \mathbf{x}^n + c^n)||_2}$$

$$\texttt{ReLU}(a) = \max(0, a)$$

with learnable parameters $\theta^n = \{\mathbf{w}^n, \mathbf{v}^n, b^n, c^n\}$, and S^n is $k = 3$ nearest exemplars to $\mathbf{x}^n$ from $\mathcal{S}^n$:

$$S^n = \{(\mathbf{x}_i^n, \mathbf{y}_i^n) \in \mathcal{S}^n : \mathbf{x}_i^n \in \mathcal{V}(k, \mathbf{x}^n, \mathcal{S}^n)\}$$

where $\mathcal{V}(k, \mathbf{x}^n, \mathcal{S}^n)$ is a set of k nearest feature vectors.

4.2 Batch Learning

Given $\mathcal{S}^n$, one can estimate Θ^n either via batch learning performed offline, or—when $\mathcal{S}^n$ is small—in real time, as outlined by Cano Santín et al. (2020). In our scenario, we learn in real time by minimizing binary cross-entropy between the ground-truth $\mathbf{y}^n$ and predicted $\hat{\mathbf{y}}^n$ concept vectors:

$$\mathcal{L}(\mathbf{y}^n, \hat{\mathbf{y}}^n) = -\sum_{z=1}^{|P^n|} l(y_z^n, \hat{y}_z^n)$$

$$l(y_i^n, \hat{y}_i^n) = y_i^n \log(\hat{y}_i^n) + (1 - y_i^n) \log(1 - \hat{y}_i^n)$$

4.3 Incremental Learning

$\mathcal{S}^n$ gets augmented whenever the teacher provides an RE–designation pair. This speech act provides two types of information: certain information $\mathcal{C}^n$ in the form of denotation-symbol-semantic value triples $(e^n, \mathtt{p}_z, y_z^n)$, corresponding to symbols and entities designated by the RE; and *noisy* information $\mathcal{N}^n$, corresponding to denotation-symbol-semantic value estimate triples $(e^n, \mathtt{p}_z, \tilde{y}_z^n)$, which are acquired from the symbols that are part of the RE and its referent inferred via (uncertain) reasoning. E.g., the RE "a circle below a square.", entails that its designation $e \in E$ is a `circle` and so $(e, \texttt{circle}, 1)$ is added to $\mathcal{C}^n$. But it also entails there exists an entity which is a `square` that is not designated by the RE, but rather this entity is in the `below` relation with the designated entity. If the grounder is sufficiently confident about the referent for `square`, then the corresponding triple is added to $\mathcal{N}^n$.

4.3.1 Acquiring Observations and Symbols

When the learner first observes its visual scene—and so the teacher has not expressed any concepts, and so the learner is currently unaware of all concepts—the noisy support set $\mathcal{N}^n$ is populated with $(e^n, \mathtt{p}_z, 0.5)$ (0.5 is a default semantic value) for all e^n in the scene and for all known n-place predicates. Whenever the teacher's RE-designation pair features a neologism $\mathtt{p}^*$, then this expansion to the learner's vocabulary prompts adding $(e^n, \mathtt{p}^*, 0.5)$ to $\mathcal{N}^n$ for all e^n. During interaction, each RE-designation pair uttered by the teacher adds elements to $\mathcal{C}^n$ (for designated symbols) and triggers updates to the $\mathcal{N}^n$ elements for all entities in the current visual scene, as we'll now describe.

4.3.2 Integrating the Teacher's Feedback

$\mathcal{N}^n$ elements are interactively updated using an incrementally built domain-level theory Δ, which is the conjunction of $\mathfrak{L}$-sentences that are built from the logical forms of the REs that the teacher has uttered so far and their designations. To compute the beliefs about semantic values, given Δ, we model the semantic value of $\mathfrak{L}$-sentences of the form $\mathtt{p}(\mathtt{t}^n)$, in which t^n are all constants (ground atom), as a random variable with Bernoulli's distribution $\mathcal{B}$. Thus a distribution over the possible domain models can be estimated using (propositional) model counting MC (Valiant, 1979), which maps each $\mathfrak{L}$-sentence to the number of domain models satisfying it. In this way, the semantic value of any proposition can be estimated as follows:

$$\tilde{y}_z^n = \begin{cases} \frac{\texttt{MC}(\mathtt{p}_z(e^n) \wedge \Delta)}{\texttt{MC}(\Delta)} & \text{if } \texttt{MC}(\Delta) \neq 0 \\ 0.5 & \text{otherwise} \end{cases}$$

MC can be computed exactly or approximately (Samer and Szeider, 2010). In our experiments we use the ADDMC (Dudek et al., 2020) weighted model counter, with weights set to 0.5.

4.3.3 Building the Support Set

Concept vectors for $\mathcal{S}^n$ are built using information in $\mathcal{C}^n$ and $\mathcal{N}^n$: namely each denotation e^n gets associated with its feature vector $\mathbf{x}^n$, and the z-dimension of $\mathbf{y}$ corresponding to predicate $\mathtt{p}_z \in P^n$ is computed as follows:

$$y_z^n = \begin{cases} y_z^n \text{ if } (e^n, \mathtt{p}_z, y_z^n) \in \mathcal{C}^n \\ \tilde{y}_z^n \text{ if } (e^n, \mathtt{p}_z, \tilde{y}_z^n) \in \mathcal{N}^n \wedge \mathbb{H}[\mathcal{B}(\tilde{y}_z^n)] \leq \tau^n \\ 0.5 \text{ otherwise} \end{cases}$$

where $\mathbb{H}[\mathcal{P}]$ is the entropy of the probability distribution $\mathcal{P}$, and τ^n is the confidence threshold for adding noisy exemplars: in our case, it's set to 0.6 for predicates of all arities.

5 Experiments

5.1 Task: Visual Reference Resolution

To evaluate IGRE, we use a task of visual reference resolution [3]: given a visual scene (an image) with localized entities (bounding boxes) and an RE, the grounder must estimate all its referents, as defined in §3.3. The model learns its task by observing an image accompanied by a sequence of REs, with each RE paired with its designation in the image.

We measure the performance of IGRE on the task after each observed RE and its designation. Performance is measured using the precision **P**, recall **R**, and F1 score **F1** on the test set between: 1) estimated vs. ground-truth domain models, formed from the concept vectors (intrinsic evaluation) and 2) estimated vs. ground-truth referents for the RE (extrinsic evaluation). These metrics are calculated only for those symbols/concepts that the teacher has mentioned so far (since the system is unaware that the remaining concepts exist). To obtain reliable results, we repeat the experiment 10 times: i.e., 10 different visual scenes, with a sequence of 5 different teacher utterances in each scene. We record in §6 the average precision, recall and f-scores over those 10 trials.

Perhaps unusually, this training and testing regime uses very small data sets: that's because in ITL it is the initial portions of the learning curve that matters. The learner must achieve decent performance on its task via only a few teacher utterances: human teachers won't tolerate repeating the same REs many times and so the learner lacks the luxury of learning (and testing) symbol grounding on large data sets.

5.2 Data: ShapeWorld

To generate training and test sets, we construct ShapeWorld domain models (Kuhnle and Copestake, 2017), each consisting of 3-12 entities, synthesized visual scenes X (64x64 pixels), and 5 REs. Each domain model is describable using 7 shape symbols **S1** (`square`, `circle`, `triangle`, `pentagon`, `cross`, `ellipse`, `semicircle`), 6 colour symbols **C1** (`red`, `green`, `blue`, `yellow`, `magenta`, `cyan`) and 4 spatial relationships symbols **R2** (`left`, `right`, `above`, `below`).[4] In scene synthesis, the image is created from the domain model, with variation on the hue of the colour category, variation on the size, position, rotation, and distortion of the shapes, and variation on the spatial positions of the entities related by each spatial term. Note that the colour categories are not mutually exclusive—e.g., there are RGB values that count as both `red` and `magenta`.

To generate REs, we sample Dependency Minimal Recursion Semantics (DMRS) (Copestake, 2009) graph templates, processed using ACE (generation mode) [5] and the English Resource Grammar (ERG) (Flickinger, 2000). Generated REs are evaluated with respect to the domain model to guarantee an existing referent. In total we generated 30 such domain models for training and 10 for testing. The data statistics for the training set is given in Table 2 for the general categories of symbols, where *certain* ($\mathcal{C}^n$) means that the designation is denoted by the symbol in the RE, and *noisy* ($\mathcal{N}^n$) means that the symbol is a part of the RE but is not designated by it. Note that the first argument to the spatial relations **R2** is always denoted by the designation while its second argument is not. Note also there is high variance in the frequencies among the individual symbols. For instance, `blue` occurs 27 and 28 times in certain vs. noisy positions respectively, while `triangle` occurs 7 and 12 times respectively.

Category	$\mathcal{C}^n$ candidates	$\mathcal{N}^n$ candidates
C1	18.67 ± 5.39	19.83 ± 5.04
S1	14.67 ± 3.98	16.50 ± 5.32
R2	0	37.75 ± 6.75

Table 2: Average symbol counts per word for colours (**C1**), shapes (**S1**), and spatial relationships (**R2**).

5.3 Implementation Details

5.3.1 Feature Extraction

To extract visual features for individuals in the scene, we utilize bounding boxes $\mathbf{b} = [x_{left}, x_{right}, y_{top}, y_{bottom}]^\top$ for each entity $e \in E$ in the visual scene by localizing them (cropping) and extracting the visual features using a pre-trained visual feature encoder (in our case,

[3] Code available at https://github.com/itl-ed/igre

[4] Entities are non-overlapping, thus we omit `on`/`behind`.

[5] http://sweaglesw.org/linguistics/ace/

DenseNet161 (Huang et al., 2017)). Additionally, for the feature vector, we add each entity's bounding box coordinates for spatial information, lost in the localization process:

$$\mathbf{x}_n = \texttt{Concat}(\{[\texttt{DenseNet161}(X[\mathbf{b}_i]], \mathbf{b}_i)\}_{i=1}^{n})$$

5.3.2 Grammar-based Semantic Parsing

To parse REs to their logical forms, we use the English Resource Grammar (ERG) and ACE (parsing mode) to produce a representation in minimal recursion semantics (MRS) (Copestake et al., 1997), which we then simplify via hand-written rules (e.g., removing event arguments from predicate symbols corresponding to adjectives and prepositions). Underspecification of the MRS was resolved using UTOOL (Koller and Thater, 2005) and the final logical form was selected based on the linear order of scope-bearing elements (quantifiers and negation): e.g. for the RE "every circle above a square", `_every_q` outscopes `_a_q`.

5.3.3 Axioms for R2

For $|E|$ entities, there are $|E|^2$ denotations to consider for each 2-place predicate—a larger search space compared to $|E|$ denotations for 1-place predicates. Moreover, these predicates can only be acquired from the noisy component $\mathcal{N}^n$ because the referent of the second argument to the relation is always latent.

To aid the learning process for **R2**, whenever a new symbol $R \in \mathbf{R2}$ is observed, domain-level axioms are added to Δ for it, making it irreflexive: $\forall x.\neg R(x,x)$ (an entity cannot be in a spatial relationship to itself) and asymmetric: $\forall x,y.R(x,y) \rightarrow \neg R(y,x)$ (reflecting the fact that entities in spatial relations take different roles (Miller and Johnson-Laird, 1976)). These axioms reduce the number of possible denotations for **R2** symbols from $|E|^2$ to $\frac{|E|^2}{2} - |E|$.

5.4 Baselines

To test the benefit of using noisy training exemplars $\mathcal{N}^n$ from the oblique symbols in the REs—in other words, those symbols that are a part of the RE but not designated by it—we implemented a HEAD grounder baseline, which uses information only from $\mathcal{C}^n$. That HEAD uses only symbol-designation pairs that are acquired when the symbol denotes the referent (in our case, that's the head noun in the RE and its pre-head modifier, if it exists).

To test the the benefit of using the precise formal semantic meanings of logical symbols (i.e., quantifiers and negation), we implemented an EXIST grounder baseline. This utilizes the information from the symbols in the oblique positions, but it does *not* utilize the precise symbolic interpretation of the logical symbols, instead simplifying the logical form of the RE by replacing all quantifiers with the existential `_a_q` and removing negation (e.g., "every cross on the left of the one circle" is equivalent to "a cross on the left of a circle"). This baseline preserves the basic linguistic structure of the formal semantic representation of the RE, but not its truth-functional interpretation.

6 Results and Discussion

Figure 2 shows the evolution of the performance of the IGRE grounder and the two baselines on the test set, as it gets exposed to more information (i.e., RE-designation pairs) over time. In the intrinsic evaluation (domain model prediction), there is no significant difference between the three grounders considered. Yet, for extrinsic evaluation (reference resolution), we observe that IGRE outperforms the HEAD and EXISTS baselines over time (both a steeper and a smoother curve). By the end of the interaction, a t-test shows significant differences in IGRE's performance compared with both baselines (p-value of 0.01).

Table 3 shows the best performance that each grounder achieved over time. When analysing their performance on particular categories, we observe that **C1** and **S1** are equally hard to learn for grounders while **R2** is easier.

We suspect that the reason why the three models performed differently in extrinsic evaluation even though they don't with intrinsic evaluation is down to the fact that IGRE uses its complete and accurate knowledge of the meanings of closed class words like quantifiers and negation at test time as well as training time in the extrinsic evaluation, but not in the intrinsic evaluation. The IGRE model can use these meanings to constrain and correct error-prone estimates of referents for open class words at test time in the reference task (as well as using their meanings to boost the training sets). For example, the RE "both squares" implies there exist exactly two squares; if the symbol grounding model has an uncertain belief that there are more (or fewer) squares than this, it will select the two most probably candidates (and infer that all other entities are

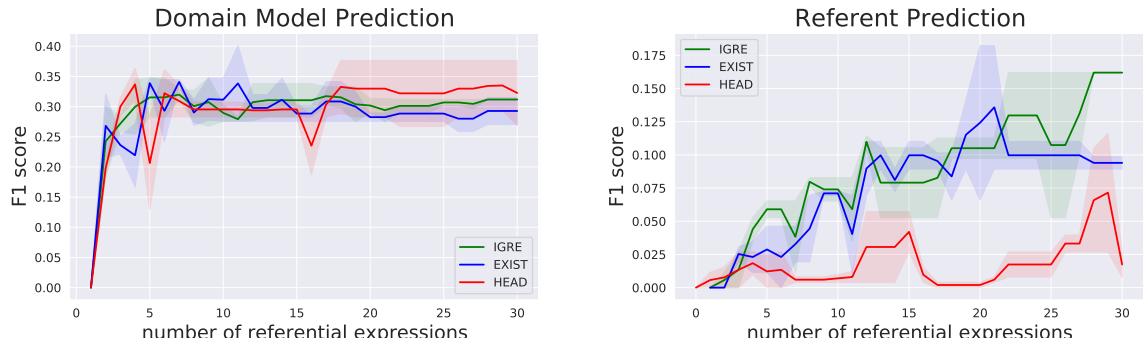

Figure 2: Evolution of F1 scores for IGRE (ours), EXISTS, HEAD grounders over the course of interaction on domain model prediction (left) and reference resolution (right)

	C1			S1			R2			Reference		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
HEAD	**0.17**	**0.54**	**0.25**	0.16	0.52	0.23	0.16	0.50	0.25	0.14	0.04	0.06
EXISTS	0.15	0.49	0.21	0.16	0.48	0.22	**0.33**	**1.00**	0.49	0.21	0.06	0.10
IGRE	**0.17**	0.51	0.23	0.17	**0.56**	**0.25**	**0.33**	**1.00**	**0.50**	**0.42**	**0.10**	**0.16**

Table 3: Precision **P**, recall **R**, and the **F1** score for symbols of different category: colour **C1**, shape **S1**, and spatial relation **R2** (intrinsic evaluation), as well as reference prediction (extrinsic evaluation). Reported metrics are averaged across the words in each category and in turn averaged across the 10 different test visual scenes.

non-squares). These experiments suggest that this sort of correction to confident but wrong estimates of the denotations of open-class symbols happens sufficiently often at test time in the reference task to make a difference in this low-data regime we are interested in, for addressing ITL tasks.

6.1 Error Analysis

The HEAD and EXISTS baselines never acquire negative exemplars: e.g., information that a particular individual is not `red`. Figure 2 shows that this severely impacts their performance, and error analysis showed that in some experiment runs it leads to model-collapse, with all denotations predicted to be in the extensions of all symbols. On the other hand, IGRE is able to acquire and use negative examples from the truth functional meanings of the logical symbols, specifically from: (a) negation ("not"); (b) the presupposition triggers"the N", "N of the M", and "all but N" where N, and M are numbers and "both"; and (c) the use of "every" when it modifies the head noun.

7 Conclusions

In this work, we presented IGRE—a grounder that supports incremental learning of the mapping from symbols to visual features whenever the teacher presents a linguistically complex RE and its designation(s) in the visual scene. The grounder starts the learning process with no conceptualisation of the domain model, and so the learner must revise its hypothesis space of possible domain models as and when the teacher introduces new and unforeseen concepts via neologisms. We showed how exploiting the model-theoretic interpretation of the formal semantic representations of REs, and in particular the truth conditions of 'logical' words like quantifiers and negation, can inform the acquisition of noisy training exemplars that in turn guide learning—IGRE reasons about the likely denotations of symbols within an RE that aren't designated by that RE, and when sufficiently confident it exploits them to update its grounding model. We showed that: 1) this grounding approach is more data efficient then a model that omits such observations and reasoning, using only the designated symbols; and 2) it is beneficial to exploit the logical consequences of the logical symbols, to gain even more data efficiency and training stability. In both cases, there was much to be gained from such reasoning because in contrast to the baselines, it contributes to acquiring *negative* exemplars: in other words, objects that get associated with *not* being red, for example.

7.1 Future Work

IGRE uses a single source of data augmentation by acquiring noisy exemplars from symbols in oblique positions. Further and parallel data gains may be obtained by exploring semi-supervised learning methods (Yarowsky, 1995; Delalleau et al., 2005).

In this work, converting $\mathcal{L}$-sentences to conjunctive normal form, which is an NP-hard problem, was a computational bottleneck. Future work needs to address this by either considering lifted inference methods (e.g., den Broeck et al. (2011)) or defining model counters that use $\mathfrak{L}$-sentences directly.

Finally, the purpose of IGRE is to aid ITL: i.e., the (incremental) updates to beliefs about symbol grounding should enhance learning to solve domain-level planning problems. Future work needs to address this by using IGRE to learn planning tasks where the learner has the physical ability to execute certain actions but starts out unaware of domain concepts that define the goal and are critical to task success. The learner must not only use IGRE to interpret the teacher's feedback, but also learn decision making strategies, both on what to say (or ask) the teacher in their extended dialogue and what actions to perform in the environment. Furthermore, the static formal semantics that we used here should be replaced with a dynamic semantics (e.g., Groenendijk and Stokhof (1991); van der Sandt (1992); Asher and Lascarides (2003)), to account for how contextual salience influences truth and reference in dialogue. Following Batra et al. (2020), we plan to test the benefits of IGRE within a system that learns to solve planning problems that focus on rearrangement tasks.

Acknowledgments

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1), UKRI Strategic Priorities Fund to the UKRI Research Node on Trustworthy Autonomous Systems Governance and Regulation (grant EP/V026607/1, 2020-2024), and the Turing 2.0 'Enabling Advanced Autonomy through Human-AI Collaboration' project funded by EPSRC and the University of Edinburgh through the Alan Turing Institute.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Amos Azaria, Jayant Krishnamurthy, and Tom M. Mitchell. 2016. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12383 Instructable intelligent personal agent. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2681–2689. AAAI Press.

Jon Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4:159–219.

Dhruv Batra, Angel X. Chang, Sonia Chernova, Andrew J. Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, Manolis Savva, and Hao Su. 2020. Rearrangement: A challenge for embodied AI. *CoRR*, abs/2011.01975.

Ben Bogin, Sanjay Subramanian, Matt Gardner, and Jonathan Berant. 2021. Latent compositional representations improve systematic generalization in grounded question answering. *Transactions of the Association of Computational Linguisticsl (TACL)*, 9.

José Miguel Cano Santín, Simon Dobnik, and Mehdi Ghanimifard. 2020. Fast visual grounding in interaction: bringing few-shot learning with neural networks to an interactive robot. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 53–61, Gothenburg. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Ann A. Copestake. 2009. Slacker semantics: Why superficiality, dependency and avoidance of commitment can be the right way to go. In *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, pages 1–9. The Association for Computer Linguistics.

Ann A. Copestake, D. Flickinger, C. Pollard, and I. Sag. 1997. Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3:281–332.

James R. Curran, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale NLP with c&c and boxer. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

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

Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. 2019. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2601–2610. IEEE.

Olivier Delalleau, Yoshua Bengio, and Nicolas Le Roux. 2005. Efficient non-parametric function induction in semi-supervised learning. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, AISTATS 2005, Bridgetown, Barbados, January 6-8, 2005*. Society for Artificial Intelligence and Statistics.

Guy Van den Broeck, Nima Taghipour, Wannes Meert, Jesse Davis, and Luc De Raedt. 2011. Lifted probabilistic inference by first-order knowledge compilation. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 2178–2185. IJCAI/AAAI.

Jeffrey M. Dudek, Vu Phan, and Moshe Y. Vardi. 2020. ADDMC: weighted model counting with algebraic decision diagrams. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 1468–1476. AAAI Press.

Evelyn Fix and Joseph L. Hodges. 1989. Discriminatory analysis - nonparametric discrimination: Consistency properties. *International Statistical Review*, 57:238.

Dan Flickinger. 2000. http://journals.cambridge.org/action/displayAbstract?aid=58601 On building a more effcient grammar by exploiting types. *Nat. Lang. Eng.*, 6(1):15–28.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 457–468. The Association for Computational Linguistics.

Jeroen Groenendijk and Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and Philosophy*, 14:39–100.

Stevan Harnad. 1999. The symbol grounding problem. *CoRR*, cs.AI/9906002.

Yordan Hristov, Daniel Angelov, Michael Burke, Alex Lascarides, and Subramanian Ramamoorthy. 2019. Disentangled relational representations for explaining and learning from demonstration. In *3rd Annual Conference on Robot Learning, CoRL 2019, Osaka, Japan, October 30 - November 1, 2019, Proceedings*, volume 100 of *Proceedings of Machine Learning Research*, pages 870–884. PMLR.

Yordan Hristov, Alex Lascarides, and Subramanian Ramamoorthy. 2018. Interpretable latent spaces for learning from demonstration. In *2nd Annual Conference on Robot Learning, CoRL 2018, Zürich, Switzerland, 29-31 October 2018, Proceedings*, volume 87 of *Proceedings of Machine Learning Research*, pages 957–968. PMLR.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2726–2736, Florence, Italy. Association for Computational Linguistics.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. 2021. MDETR - modulated detection for end-to-end multimodal understanding. *CoRR*, abs/2104.12763.

Siddharth Karamcheti, Dorsa Sadigh, and Percy Liang. 2020. Learning adaptive language interfaces through decomposition. *CoRR*, abs/2010.05190.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 787–798. ACL.

Casey Kennington, Livia Dia, and David Schlangen. 2015. A discriminative model for perceptually-grounded incremental reference resolution. In *Proceedings of the 11th International Conference on*

Computational Semantics, IWCS 2015, 15-17 April, 2015, Queen Mary University of London, London, UK, pages 195–205. The Association for Computer Linguistics.

Casey Kennington and David Schlangen. 2017. A simple generative model of incremental reference resolution for situated dialogue. *Comput. Speech Lang.*, 41:43–67.

Alexander Koller and Stefan Thater. 2005. The evolution of dominance constraint solvers. In *Proceedings of Workshop on Software*, pages 65–76, Ann Arbor, Michigan. Association for Computational Linguistics.

Alexander Kuhnle and Ann A. Copestake. 2017. Shapeworld - A new test methodology for multimodal language understanding. *CoRR*, abs/1704.04517.

John E. Laird, Kevin A. Gluck, John R. Anderson, Kenneth D. Forbus, Odest Chadwicke Jenkins, Christian Lebiere, Dario D. Salvucci, Matthias Scheutz, Andrea Thomaz, J. Gregory Trafton, Robert E. Wray, Shiwali Mohan, and James R. Kirk. 2017. Interactive task learning. *IEEE Intelligent Systems*, 32(4):6–21.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Cynthia Matuszek. 2018. Grounded language learning: Where robotics and NLP meet. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5687–5691. ijcai.org.

George A. Miller and Philip N. Johnson-Laird. 1976. *Language and Perception*. Belknap Press Imprint.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Bertrand Russell. 1917. Knowledge by acquaintance and knowledge by description. In *Mysticism and Logic*, pages 152–167. London: Longmans Green.

Marko Samer and Stefan Szeider. 2010. Algorithms for propositional model counting. *J. Discrete Algorithms*, 8(1):50–64.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1842–1861. Association for Computational Linguistics.

Mohit Shridhar and David Hsu. 2018. Interactive visual grounding of referring expressions for human-robot interaction. In *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*.

Sanjay Subramanian, Ben Bogin, Nitish Gupta, Tomer Wolfson, Sameer Singh, Jonathan Berant, and Matt Gardner. 2020. Obtaining faithful interpretations from compositional neural networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5594–5608, Online. Association for Computational Linguistics.

Sanjay Subramanian, Sameer Singh, and Matt Gardner. 2019. Analyzing compositionality in visual question answering. In *ViGIL@NeurIPS*.

Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics.

Leslie G. Valiant. 1979. The complexity of computing the permanent. *Theor. Comput. Sci.*, 8:189–201.

Johan van Benthem. 1984. Questions about quantifiers. *J. Symb. Log.*, 49(2):443–466.

Rob van der Sandt. 1992. Presupposition projection as anaphora resolution. *Journal of Semantics*, 9(4):333–377.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. https://proceedings.neurips.cc/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3630–3638.

Sida I. Wang, Samuel Ginn, Percy Liang, and Christopher D. Manning. 2017. Naturalizing a programming language via interactive learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 929–938. Association for Computational Linguistics.

Sida I. Wang, Percy Liang, and Christopher D. Manning. 2016. Learning language games through interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

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

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8:229–256.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10502–10511. Computer Vision Foundation / IEEE.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. http://papers.nips.cc/paper/7381-neural-symbolic-vqa-disentangling-reasoning-from-vision-and-language-understanding Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1039–1050.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3208–3216. AAAI Press.

Yanpeng Zhao and Ivan Titov. 2020. Visually grounded compound pcfgs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4369–4379. Association for Computational Linguistics.