A Visually Grounded Natural Language Interface for Reference to Spatial Scenes

Peter Gorniak
MIT Media Laboratory
20 Ames Street
Cambridge, MA, 02142, USA
pgorniak@media.mit.edu

Deb Roy
MIT Media Laboratory
20 Ames Street
Cambridge, MA, 02142, USA
dkroy@media.mit.edu

ABSTRACT

Many user interfaces, from graphic design programs to navigation aids in cars, share a virtual space with the user. Such applications are often ideal candidates for speech interfaces that allow the user to refer to objects in the shared space. We present an analysis of how people describe objects in spatial scenes using natural language. Based on this study, we describe a system that uses synthetic vision to "see" such scenes from the person's point of view, and that understands complex natural language descriptions referring to objects in the scenes. This system is based on a rich notion of semantic compositionality embedded in a grounded language understanding framework. We describe its semantic elements, their compositional behaviour, and their grounding through the synthetic vision system. To conclude, we evaluate the performance of the system on unconstrained input.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language parsing and understanding*; I.2.0 [**Artificial Intelligence**]: General—*Cognitive Simulation*

General Terms

Human Factors, Experimentation

Keywords

natural language understanding, computational semantics, cognitive modelling, vision based semantics

1. INTRODUCTION

Human-computer interfaces often share a geometric space with the user. For example, drawing applications provide a two dimensional drawing surface; modeling applications provide a three dimensional space. Global positioning system (GPS) based devices share a map of the physical terrain. While there have been interfaces that allow users to point in shared virtual spaces and speak about objects, here we consider the situation where pointing alone is either extremely inaccurate, such as in a complex three dimensional scenes, or undesirable, such as when driving a car. Our focus, thus, is to understand how users refer to objects in spatial scenes using only language (e.g., "the cylinder at the back under a cube" in the 3-D design case, or in a car navigation domain, "you are looking for the first house immediately after the cluster of shops on your right"). Based on an analysis of spatial language, we have developed a visually-grounded language understanding system that can interpret spatial expressions and bind them to objects in a scene.

We first present a study of how people describe objects in visual scenes of the kind shown in Figure 1. Based on this study, we propose a computational model of visually-grounded language understanding. A typical referring expression for Figure 1 might be, "the far back purple cone that's behind a row of green ones". In such tasks, speakers construct expressions to guide listeners' attention to intended objects. Such referring expressions succeed in communication because speakers and listeners find similar features of the visual scene to be salient, and share an understanding of how language is grounded in terms of these features. This work is a step towards our longer term goals to develop a conversational robot that can fluidly connect language to perception and action.[1]

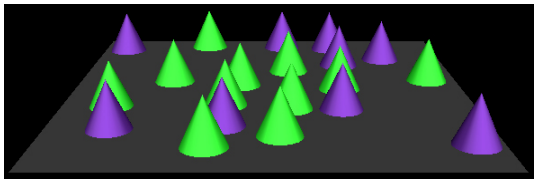


Figure 1: A sample scene used to elicit visually-grounded referring expressions (if this figure has been reproduced in black and white, the light cones are green in colour, the dark cones are purple)

To study the characteristics of descriptive spatial language, we collected several hundred referring expressions based on scenes similar to Figure 1. We analysed the descriptions by cataloguing the visual features that they referred to within a scene, and the range of linguistic devices (words or grammatical patterns) that they used to refer to

[1]An extend report on this work is currently in review for a journal publication

ICMI'03, November 5–7, 2003, Vancouver, British Columbia, Canada.

those features. The combination of a visual feature and corresponding linguistic device is referred to as a *descriptive strategy*.

We propose a set of computational mechanisms that correspond to the most commonly used descriptive strategies from our study. The resulting model has been implemented as a set of visual feature extraction algorithms, a lexicon that is grounded in terms of these visual features, a robust parser to capture the syntax of spoken utterances, and a compositional engine driven by the parser that combines visual groundings of lexical units. We use the term *grounded semantic composition* to highlight that both the semantics of individual words and the word composition process itself are visually-grounded. We propose processes that combine the visual models of words, governed by rules of syntax. In designing our system, we made several simplifying assumptions. We assumed that word meanings are independent of the visual scene, and that semantic composition is a purely incremental process. As we will show, neither of these assumptions holds in all of our data, but our system still understands most utterances correctly.

To evaluate the system, we collected a set of spoken utterances from three speakers. The model was able to correctly understand the visual referents of 59% of the expressions (chance performance was $1/30 \sum_{i=1}^{30} 1/i = 13\%$). The system was able to resolve a range of linguistic phenomena that made use of relatively complex compositions of spatial semantics. We provide an analysis of the sources of failure in this evaluation, based on which we propose a number of improvements that are required to achieve human level performance.

1.1 Related Work

There are a number of systems that allow users to talk about a virtual space shared with a computer [9]. Usually, speech is used to augment another selection modality such as pen based input. Spatial referring expressions are thus relatively simple, usually only incorporating a single descriptive strategy such as a spatial relation (e.g. "east of") [10]. Here, we are interested in situations where people use only speech to describe objects, and thus produce far more complex object descriptions.

Winograd's SHRDLU is a well known system that understands and generates natural language referring to objects and actions in a simple blocks world [20]. Like our system, it performs semantic interpretation during parsing by attaching short procedures to lexical units. However, SHRDLU has access to a clean symbolic representation of the scene and only handles sentences it can completely parse. The system discussed here works with a synthetic vision system, reasons over geometric and other visual measures, and processes verbatim transcriptions of natural (and often ungrammatical) speech.

Partee provides an overview of the formal semantics approach exemplified by SHRDLU and the problems of context based meanings and meaning compositionality from this perspective [11]. Our work reflects many of the ideas from this work, such as viewing adjectives as functions, as well as ideas about compositional behaviours of lexical items from Pustejovsky's theory of the Generative Lexicon (GL) [12]. However, these formal approaches operate in a symbolic domain and leave the details of non-linguistic influences on meaning unspecified, whereas we take the computational modelling of these influences as our primary concern.

Word meanings have been approached by several researchers as a problem of associating visual representations, often with complex internal structure, to word forms. Models have been suggested for visual representations underlying colour [5] and spatial relations [13]. Models for verbs include grounding their semantics in the perception of actions [18], and grounding in terms of motor control programs [8]. Landau and Jackendoff provide a detailed analysis of additional visual shape features that play a role in language [6]. For example, they suggest the importance of extracting the geometric axes of objects in order to ground words such as "end", as in "end of the stick". Shi and Malik propose an approach to performing visual grouping on images [17]. Their work draws from findings of Gestalt psychology that provide many insights into visual grouping behaviour [19], which also inspired further work on grouping in our laboratory [3].

SAM [2] and Ubiquitous Talker [7] are language understanding systems that map language to objects in visual scenes. Similar to SHDRLU, the underlying representation of visual scenes is symbolic and loses much of the subtle visual information that our work, and the work cited above, focus on. Both SAM and Ubiquitous Talker incorporate a vision system, phrase parser and understanding system. The systems translate visually perceived objects into a symbolic knowledge base and map utterances into plans that operate on the knowledge base. In contrast, we are primarily concerned with understanding language referring to the objects and their relations as they appear visually.

We have previously proposed methods for visually-grounded language learning [16], understanding [15], and generation [14]. We have also applied these grounded learning algorithms to learning in multimodal interfaces before [4]. However, the treatment of semantic composition in these efforts was relatively primitive. While this simple approach worked in the constrained domains addressed in the past, it does not scale to the present task.

2. A SPATIAL DESCRIPTION TASK

We designed a task that requires people to describe objects in computer generated scenes containing up to 30 objects with random positions on a virtual surface. The objects were all of identical shape and size, and were either green or purple in colour. Each of the objects had a 50% chance of being green, otherwise it was purple. Due to the indistinguishability of same-colored objects, speakers were led to make reference to spatial aspects of the scene. We refer to this task as the Bishop task, and to the resulting language understanding model and implemented system simply as Bishop.

2.1 Data Collection

Participants in the study ranged in age from 22 to 30 years, and included both native and non-native English speakers. Pairs of participants were seated with their backs to each other, each looking at a computer screen which displayed identical scenes such as that in Figure 1. In each pair, one participant served as describer, and the other as listener. The describer wore a microphone that was used to record his or her speech. The describer used a mouse to select an object from the scene, and then verbally described the selected object to the listener. The listener was not allowed to communicate verbally or otherwise at all, except through object selections. The listener's task was to select the same object on their own computer display based

on the verbal description. If the selected objects matched, they disappeared from the scene and the describer would select and describe another object. If they did not match, the describer would re-attempt the description until understood by the listener. Using a describer-listener dyad ensured that speech data resembled natural communicative dialogue. Participants were told they were free to select any object in the scene and describe it in any way they thought would be clear. They were also told not to make the task trivial by, for example, always selecting the leftmost object and describing it as "leftmost". The scene contained 30 objects at the beginning of each session, and a session ended when no objects remained, at which point the describer and listener switched roles and completed a second session (some participants fulfilled a role multiple times). We found that listeners in the study made extremely few mistakes in interpreting descriptions, and seemed to generally find the task easy to perform.

Initially, we collected 268 spoken object descriptions from 6 participants. The raw audio was segmented using a speech segmentation algorithm based on pause structure [21]. Along with the utterances, the corresponding scene layout and target object identity were recorded together with the times at which objects were selected. This 268 utterance corpus is referred to as the development data set. We manually transcribed each spoken utterance verbatim, retaining all speech errors (false starts and various other ungrammaticalities). Off-topic speech events (laughter, questions about the task, other remarks, and filled pauses) were marked as such (they do not appear in any results we report). We wrote a simple heuristic algorithm based on time stamps to pair utterances and selections based on their time stamps. When we report numbers of utterances in data sets in this paper, they correspond to how many utterance-selection pairs our pairing algorithm produces.

Once our implementation based on the development corpus yielded acceptable results, we collected another 179 spoken descriptions from three additional participants to evaluate generalization and coverage of our approach. The discussion and analysis in the following sections focuses on the development set. In Section 6 we discuss performance on the test set.

2.2 Descriptive Strategies for Achieving Joint Reference

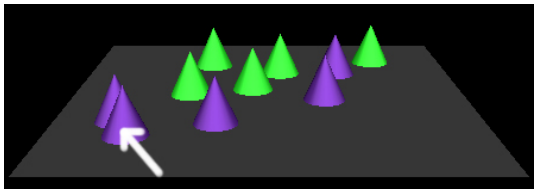
We distinguish three subsets of our development data, 1) a set containing those utterance/selection pairs that contain errors, where an error can be due to a repair or mistake on the human speaker's part, a segmentation mistake by our speech segmenter, or an error by our utterance/selection pairing algorithm, 2) a set that contains those utterance/selection pairs that employ descriptive strategies other than those we cover in our computational understanding system (we cover those in Sections 2.2.1 to 2.2.5), and 3) the set of utterance/selection pairs in the development data that are not a member of either subset described above. We refer to this last subset as the 'clean' set. Note that the first two subsets are not mutually exclusive. As we catalogue descriptive strategies from the development data in the following sections, we report two percentages for each descriptive strategy. The first is the percentage of utterance/selection pairs that employ a specific descriptive strategy relative to all the utterance/selection pairs in the development data set. The second is the percentage of utterance/selection pairs relative to the clean set of utterance/selection pairs, as described above.

2.2.1 Colour

Almost every utterance employs colour to pick out objects. While designing the task, we intentionally trivialized the problem of colour reference. Objects come in only two distinct colours, green and purple. Unsurprisingly, all participants used the terms "green" and "purple" to refer to these colours. Participants used colour to identify one or more objects in 96% of the data, and 95% of the clean data.

2.2.2 Spatial Regions and Extrema

The second most common descriptive strategy is to refer to spatial extremes within groups of objects and to spatial regions in the scene. The example in Figure 2 uses two spatial terms to pick out its referent: "front" and "left", both of which leverage spatial extrema to direct the listener's attention. Multiple spatial specifications tend to be interpreted in left to right order, that is, selecting a group of objects matching the first term, then amongst those choosing objects that match the second term.



"the purple one in the front left corner"

Figure 2: Example utterance specifying objects by referring to spatial extrema

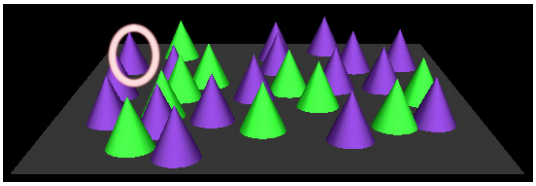
Being rather ubiquitous in the data, spatial extrema and spatial regions are often used in combination with other descriptive strategies like grouping, but are most frequently combined with other extrema and region specifications. Participants used single spatial extrema to identify one or more objects in 72% of the data, and in 78% of the clean data. They used spatial region specifications in 20% of the data (also 20% of the clean data), and combined multiple extrema or regions in 28% (30% of the clean data).

2.2.3 Grouping

To provide landmarks for spatial relations and to specify sets of objects to select from, participants used language to describe groups of objects. Figure 3 shows an example of such grouping constructs, which uses a count to specify the group ("three"). In this example, the participant first specifies a group containing the target object, then utters another description to select within that group. Note that grouping alone never yields an individual reference, so participants compose grouping constructs with further referential tactics (predominantly extrema and spatial relations) in all cases. Participants used grouping to identify objects in 12% of the data and 10% of the clean data.

2.2.4 Spatial Relations

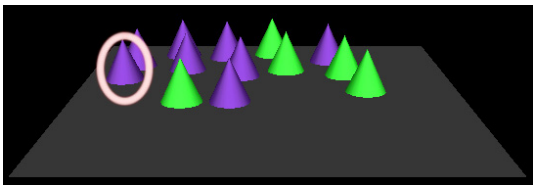
As already mentioned in Section 2.2.3, participants sometimes used spatial relations between objects or groups of objects. Examples of such relations are expressed through prepositions like "below" or "behind" as well as phrases like "to the left of" or "in front of". Figure 4 shows an example



"there's three on the left side; the one in the furthest back"

Figure 3: Example utterance using grouping

that involves a spatial relation between individual objects. The spatial relation is combined with another strategy, here an extremum (as well as two speech errors by the describer). Participants used spatial relations in 6% of the data (7% of the clean data).

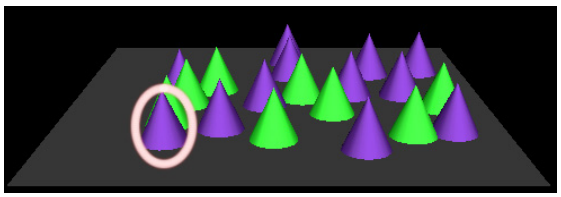


"there's a purple cone that's it's all the way on the left hand side but it's it's below another purple"

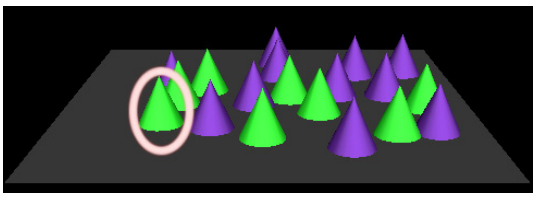
Figure 4: Example utterance specifying a spatial relation

2.2.5 *Anaphora*

In a number of cases participants used anaphoric references to the previous object removed during the description task. Figure 5 shows a sequence of two scenes and corresponding utterances in which the second utterance refers back to the object selected in the first. Participants employed spatial relations in 4% of the data (3% of the clean data).



"the closest purple one on the far left side"



"the green one right behind that one"

Figure 5: Example sequence of an anaphoric utterance

2.2.6 *Other*

In addition to the phenomena listed in the preceding sections, participants used a small number of other description strategies. Some that occurred more than once but that we have not yet addressed in our computational model are selection by distance (lexicalised as "close to" or "next to"), selection by neighbourhood ("the green one surrounded by purple ones"), selection by symmetry ("the one opposite that one"), and selection by something akin to local connectivity ("the lone one"). We annotated 13% of our data as containing descriptive strategies other than the ones covered in the preceding sections. We marked 15% of our data as containing errors.

3. THE UNDERSTANDING FRAMEWORK

3.1 Synthetic Vision

Instead of relying on the information we use to render the scenes in Bishop, which includes 3-D object locations and the viewing angle, we implemented a simple synthetic vision algorithm to ease a future transfer back to a robot's vision system. This algorithm produces a map attributing each pixel of the rendered image to one of the objects or the background. In addition, we use the full colour information for each pixel drawn in the rendered scene. We chose to work in a virtual world for this project so that we could freely change colour, number, size, shape and arrangement of objects to elicit interesting verbal behaviours in our participants.

3.2 Lexical Entries and Concepts

Conceptually, we treat lexical entries like classes in an object oriented programming language. When instantiated, they maintain an internal state that can be as simple as a tag identifying the dimension along which to perform an ordering, or as complex as multidimensional probability distributions. Each entry can contain a semantic composer that encapsulates the function to combine this entry with other constituents during a parse. These composers are described in-depth in Section 4. The lexicon used for Bishop contains many lexical entries attaching different semantic composers to the same word. For example, "left" can be either a spatial relation or an extremum, which may be disambiguated by grammatical structure during parsing.

During composition, structures representing the objects that a constituent references are passed between lexical entries. We refer to these structures as *concepts*. Each entry accepts zero or more concepts, and produces zero or more concepts as the result of the composition operation. A concept lists the entities in the world that are possible referents of the constituent it is associated with, together with real numbers representing their ranking due to the last composition operation.

3.3 Parsing

We use a bottom-up chart parser to guide the interpretation of phrases [1]. Such a parser has the advantage that it employs a dynamic programming strategy to efficiently reuse already computed subtrees of the parse. Furthermore, it produces all sub-components of a parse and thus produces a useable result without the need to parse to a specific symbol.

Bishop performs only a partial parse, a parse that is not required to cover a whole utterance, but simply takes the longest referring parsed segments to be the best guess. Unknown words do not stop the parse process. Rather, all constituents that would otherwise end before the unknown

word are taken to include the unknown word, in essence making unknown words invisible to the parser and the understanding process. In this way we recover essentially all grammatical chunks and relations that are important to understanding in our restricted task. This feature will be important in future work in which we plan to integrate speech recognition into the system.

We use a simple grammar containing 19 rules. Each rule is associated with an argument structure for semantic composition. When a rule is syntactically complete during a parse, the parser checks whether the composers of the constituents in the tail of the rule can accept the number of arguments specified in the rule. If so, it calls the semantic composer associated with the constituent with the concepts yielded by its arguments to produce a concept for the head of the rule.

4. SEMANTIC COMPOSITION

Most of the composers presented follow the same composition schema: they take one or more concepts as arguments and yield another concept that references a possibly different set of objects. Composers may introduce new objects, even ones that do not exist in the scene as such, and they may introduce new types of objects (e.g. groups of objects referenced as if they were one object). Most composers first convert an incoming concept to the objects it references, and subsequently perform computations on these objects. If ambiguities persist at the end of understanding an utterance (multiple possible referents exist), we let Bishop choose the one with maximum reference strength.

4.1 Colour - Probabilistic Attribute Composers

As mentioned in Section 3.1, we chose not to exploit the information used to render the scene, and therefore must recover colour information from the final rendered image. The colour average for the 2-D projection of each object varies due to occlusion by other objects, as well as distance from and angle with the virtual camera. We separately collected a set of labelled instances of "green" and "purple" cones, and estimated a three dimensional Gaussian distribution from the average red, green and blue values of each pixel belonging to the example cones. When asked to compose with a given concept, this type of probabilistic attribute composer assigns each object referenced by the source concept the probability density function evaluated at the average colour of the object.

4.2 Spatial Extrema and Spatial Regions - Ordering Composers

To determine spatial regions and extrema, an ordering composer orders objects along a specified feature dimension (e.g. x coordinate relative to a group) and picks referents at an extreme end of the ordering. To do so, it assigns an exponential weight function to objects according to $\gamma^{i(1+v)}$ for picking minimal objects, where i is the object's position in the sequence, v is its value along the feature dimension specified, normalized to range between 0 and 1 for the objects under consideration. The maximal case is weighted similarly, but using the reverse ordering subtracting the fraction in the exponent from 2. For our reported results $\gamma = 0.38$. This formula lets referent weights fall off exponentially both with their position in the ordering and their distance from the extreme object. In that way extreme objects are isolated except for cases in which many referents cluster around an extremum, making picking out a single referent difficult. We attach this type of composer to words like "leftmost" and "top".

The ordering composer can also order objects according to their absolute position, corresponding more closely to spatial regions rather than spatial extrema relative to a group. The reference strength formula for this version is $\gamma^{(1+\frac{d}{d_{\max}})}$ where d is the euclidean distance from a reference point, and $d_{\max}$ the maximum such distance amongst the objects under consideration. This version of the composer is attached to words like "middle". It has the effect that reference weights are relative to absolute position on the screen. An object close to the centre of the board achieves a greater reference weight for the word "middle", independently of the position of other objects of its kind. Ordering composers work across any number of dimensions by simply ordering objects by their Euclidean distance, using the same exponential falloff function as in the other cases.

4.3 Grouping Composers

For non-numbered grouping (e.g., when the describer says "group" or "cones"), the grouping composer searches the scene for groups of objects that are all within a maximum distance threshold from another group member. It only considers objects that are referenced by the concept it is passed as an argument. For numbered groups ("two", "three"), the composer applies the additional constraint that the groups have to contain the correct number of objects. Reference strengths for the concept are determined by the average distance of objects within the group.

The output of a grouping composer may be thought of as a group of groups. To understand the motivation for this, consider the utterance, "the one to the left of the group of purple ones". In this expression, the phrase "group of purple ones" will activate a grouping composer that will find clusters of purple cones. For each cluster, the composer computes the convex hull (the minimal "elastic band" that encompasses all the objects) and creates a new composite object that has the convex hull as its shape. When further composition takes place to understand the entire utterance, each composite group serves as a potential landmark relative to "left".

However, concepts can be marked so that their behaviour changes to split apart concepts refering to groups. For example, the composer attached to "of" sets this flag on concepts passing through it. Note that "of" is only involved in composition for grammar rules of the type NP $\leftarrow$ NP P NP, but not for those performing spatial compositions for phrases like "to the left of". Therefore, the phrase "the frontmost one of the three green ones" will pick the front object within the best group of three green objects.

4.4 Spatial Relations - Spatial Composers

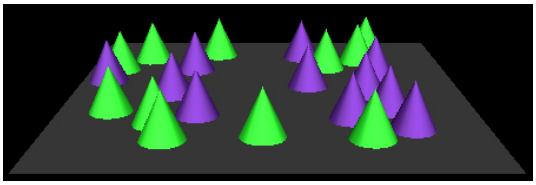
The spatial semantic composer employs a version of the Attentional Vector Sum (AVS) suggested in [13]. The AVS is a measure of spatial relation meant to approximate human judgements corresponding to words like "above" and "to the left of" in 2-D scenes of objects. Given two concepts as arguments, the spatial semantic composer converts both into sets of objects, treating one set as providing possible landmarks, the other as providing possible targets. The composer then calculates the AVS for each possible combination of landmarks and targets. Finally, the spatial composer divides the result by the Euclidean distance between the objects' centres of mass, to account for the fact that par-

ticipants exclusively used nearby objects to select through spatial relations.

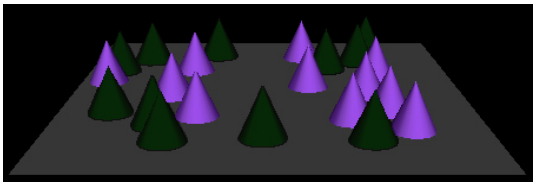
4.5 Anaphoric Composers

Triggered by words like "that" (as in "to the left of that one") or "previous", an anaphoric composer produces a concept that refers to a single object, namely the last object removed from the scene during the session. This object specially marks the concept as referring not to the current, but the previous visual scene, and any further calculations with this concept are performed in that visual context.

5. EXAMPLE: UNDERSTANDING A DESCRIPTION



"the purple one on the left"



"the purple one"

"one on the left"

"the purple one on the left"

Figure 6: Example: "the purple one on the left"

Consider the scene in Figure 6, and the output of the chart parser for the utterance, "the purple one on the left" in Figure 7. Starting at the top left of the parse output, the parser finds "the" in the lexicon as an ART (article) with a selecting composer that takes one argument. It finds two lexical entries for "purple", one marked as a CADJ (colour adjective), and one as an N (noun). Each of them have the same composer, a probabilistic attribute composer marked as P(), but the adjective expects one argument whereas the noun expects none. Given that the noun expects no arguments and that the grammar contains a rule of the form NP← N, an NP (noun phrase) is instantiated and the probabilistic composer is applied to the default set of objects yielded by N, which consists of all objects visible. This composer call is marked P(N) in the chart. After composition, the NP contains a subset of only the purple objects (Figure 6, top right). At this point the parser applies NP ← ART NP, which produces the NP spanning the first two words and again contains only the purple objects, but is marked as unambiguously referring to an object. S(NP) marks the application of this selecting composer called S.

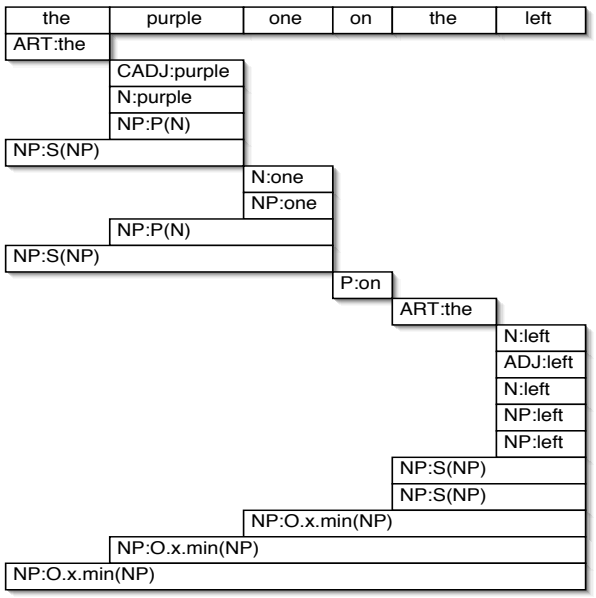


Figure 7: Sample parse of a referring noun phrase

The parser goes on to produce a similar NP covering the first three words by combining the "purple" CADJ with "one" and the result with "the". The "on" P (preposition) is left dangling for the moment as it needs a constituent that follows it. It contains a modifying semantic composer that simply bridges the P, applying the first argument to the second. After another "the", "left" has several lexical entries: in its ADJ and one of its N forms it contains an ordering semantic composer that takes a single argument, whereas its second N form contains a spatial semantic composer that takes two arguments to determine a target and a landmark object. At this point the parser can combine "the" and "left" into two possible NPs, one containing the ordering and the other the spatial composer. The first of these NPs in turn fulfills the need of the "on" P for a second argument according to NP ← NP P NP, performing its ordering compose first on "one" (for "one on the left"), selecting all the objects on the left (Figure 6, bottom left). The application of the ordering composer is denoted as $O.x.min$(NP) in the chart, indicating that this is an ordering composer ordering along the x axis and selecting the minimum along this axis. On combining with "purple one", the same composer selects all the purple objects on the left (Figure 6, bottom right). Finally on "the purple one", it produces the same set of objects as "purple one", but marks the concept as unambiguously picking out a single object. Note that the parser attempts to use the second interpretation of "left" (the one containing a spatial composer) but fails because this composer expects two arguments that are not provided by the grammatical structure of the sentence.

6. RESULTS AND DISCUSSION

6.1 Overall Performance

In Table 1 we present overall accuracy results, indicating for which percentage of different groups of examples our system picked the same referent as the person describing the object. The first line in the table shows performance relative to the total set of utterances collected. The second one shows the percentage of utterances our system understood correctly excluding those marked as using a descriptive strategy that was not listed in Section 4, and thus not expected to be understood by Bishop. The final line in Table 1 shows the percentage of utterances for which our system picked the correct referent relative to the clean development and testing sets. Although there is obviously room for improvement, these results are significant given that chance performance on this task is only 13.3% and linguistic input was transcripts of unconstrained speech.

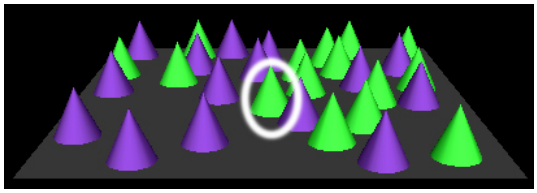
Utterance Set	Accuracy - Development	Accuracy - Testing
All	76.5%	58.7%
All except 'Other'	83.2%	68.8%
Clean	86.7%	72.5%

Table 1: Overall Results

Colour Due to the simple nature of colour naming in the Bishop task, the probabilistic composers responsible for selecting objects based on colour made no errors.

Spatial Extrema Our ordering composers correctly identify 100% of the cases in which a participant uses only colour and a single spatial extremum in his or her description. Participants also favour this descriptive strategy, using it with colour alone in 38% of the clean data. In the clean training data, Bishop understands 86.8% of all utterances employing spatial extrema. Participants composed one or more spatial region or extrema references in 30% of the clean data. Our ordering composers correctly interpret 85% of these cases, for example that in Figure 2 in Section 2.2.2. The mistakes our composers make are usually due to overcommitment and faulty ordering.

Spatial Regions Description by spatial region occurs alone in only 5% of the clean data, and together with other strategies in 15% of the clean data. Almost all the examples of this strategy occurring alone use words like "middle" or "centre". The top image in Figure 8 exemplifies the use of "middle" that our ordering semantic composer models. The object referred to is the one closest to the centre of the board. The bottom image in Figure 8 shows a different interpretation of middle: the object in the middle of a (linguistically not mentioned) group of objects. Note that within the group there are two candidate centre objects, and that the one in the front is preferred. There are also further meanings of middle that we leave out due to space constraints. In summary, we can catalogue a number of different meanings for the word "middle" in our data that are linguistically indistinguishable, but depend on visual and historical context to be correctly understood.



"the green one in the middle"

"the purple cone in the middle"

Figure 8: Types of "middles"

Grouping Our composers implementing the grouping strategies used by participants are the most simplistic of all composers we implemented, compared to the depth of the actual phenomenon of visual grouping. As a result, Bishop only understands 29% of utterances that employ grouping in the clean training data. More sophisticated grouping algorithms have been proposed, such as [3].

Spatial Relations The AVS measure divided by distance between objects corresponds very well to human spatial relation judgements in this task. All the errors that occur in utterances that contain spatial relations are due to the possible landmarks or targets not being correctly identified (grouping or region composers might fail to provide the correct referents). Our spatial relation composer picks the correct referent in all those cases where landmarks and targets are the correct ones. Bishop understands 64.3% of all utterances that employ spatial relations in the clean training data. There are types of spatial relations such as relations based purely on distance and combined relations ("to the left and behind") that we decided not to cover in this implementation, but that occur in the data and should be covered in future efforts.

Anaphora Our solution to the use of anaphora in the Bishop task performs perfectly (100% of utterances employing anaphora) in understanding reference back to a single object in the clean development data. However, there are more complex variants of anaphora that we do not currently cover, for example reference back to groups of objects.

7. FUTURE DIRECTIONS

Each of our semantic composers attempts to solve a separate hard problem, some of which (e.g. grouping and spatial relations) have seen long lines of work dedicated to more sophisticated solutions than ours. The individual problems were not the emphasis of this paper, and the solutions presented here can be improved.

If a parse does not produce a single referent, backtracking would provide an opportunity to revise the decisions made at various stages of interpretation until a referent is produced. Yet backtracking only solves problems in which the

system knows that it has failed to obtain a good answer. We presented cases of selection of word meanings by visual context in our data. In such cases, a good candidate solution according to one word meaning may still produce the wrong referent due to a specific visual context. A future system should take into account local and global visual context during composition to account for these human selection strategies.

By constructing the parse charts we obtain a rich set of partial and full syntactic and semantic fragments offering explanations for parts of the utterance. In the future, we plan to use this information to engage in clarification dialogue with the human speaker.

Machine learning algorithms may be used to learn many of the parameter settings that were set by hand in this work, including on-line learning to adapt parameters during verbal interaction. Furthermore, learning new types of composers and appropriate corresponding grammatical constructs poses a difficult challenge for the future.

8. SUMMARY

We have presented a model of visually-grounded language understanding that is able to connect natural language descriptions to objects in a scene. At the heart of the model is a set of lexical items, each grounded in terms of visual features and grouping properties. A robust parsing algorithm finds chunks of syntactically coherent words from an input utterance. To determine the semantics of phrases, the parser activates semantic composers that combine words to determine their joint reference. The robust parser is able to process grammatically ill-formed transcripts of natural spoken utterances. In evaluations, the system selected correct objects in response to utterances for 76.5% of the development set data, and for 58.7% of the test set data. On clean data sets with various speech and processing errors held out, performance was higher yet.

We suggested several avenues for improving performance of the system including better methods for spatial grouping, semantically guided backtracking during sentence processing, the use of machine learning to replace hand construction of models, and the use of interactive dialogue to resolve ambiguities. We plan to merge this work in understanding of complex spatial language semantics with other work in learning language semantics interactively from the user to provide rich understanding and adaptation facilities to multimodal user interfaces.

9. REFERENCES

[1] J. Allen. *Natural Language Understanding*, chapter 3. The Benjamin/Cummings Publishing Company, Inc, Redwood City, CA, USA, 1995.

[2] M. Brown, B. Buntschuh, and J. Wilpon. SAM: A perceptive spoken language-understanding robot. *IEEE Transactions on Systems, Man and Cybernetics*, 6(22):1390–1402, Nov/Dec 1992.

[3] S. Dhande. A computational model to connect gestalt perception and natural language. Master's thesis, Massachusetts Institure of Technology, 2003.

[4] P. Gorniak and D. Roy. Augmenting user interfaces with adaptive speech commands. In *Proceedings of the International Conference for Multimodal Interfaces*, 2003.

[5] J. M. Lammens. *A computational model of color perception and color naming*. PhD thesis, State University of New York, 1994.

[6] B. Landau and R. Jackendoff. "what" and "where" in spatial language and spatial cognition. *Behavioural and Brain Sciences*, 2(16):217–238, 1993.

[7] K. Nagao and J. Rekimoto. Ubiquitous talker: Spoken language interaction with real world objects. In *Proceeding of the International Joint Conference on Artificial Intelligence*, 1995.

[8] S. Narayanan. *KARMA: Knowledge-based Action Representations for Metaphor and Aspect*. PhD thesis, Univesity of California, Berkeley, 1997.

[9] S. Oviatt, P. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, and D. Ferro. Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions. *Human Computer Interaction*, 15(4):263–322, August 2000.

[10] S. L. Oviatt, A. DeAngeli, and K. Kuhn. Integration and synchronization of input modes during multimodal human-computer interaction. In *CHI*, pages 415–422, 1997.

[11] B. H. Partee. Lexical semantics and compositionality. In L. R. Gleitman and M. Liberman, editors, *An Invitation to Cognitive Science: Language*, volume 1, chapter 11, pages 311–360. MIT Press, Cambridge, MA, 1995.

[12] J. Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, MA, USA, 1995.

[13] T. Regier and L. Carlson. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2):273–298, 2001.

[14] D. Roy. Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, 16(3), 2002.

[15] D. Roy, P. J. Gorniak, N. Mukherjee, and J. Juster. A trainable spoken language understanding system. In *Proceedings of the International Conference of Spoken Language Processing*, 2002.

[16] D. Roy and A. Pentland. Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146, 2002.

[17] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(22):888–905, August 2000.

[18] J. M. Siskind. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research*, 15:31–90, August 2001.

[19] M. Wertheimer. Laws of organization in perceptual forms. In *A source book of Gestalt psychology*, pages 71–88. Routledge, New York, 1999.

[20] T. Winograd. *Procedures as a representation for data in a computer program for understanding natural language*. PhD thesis, Massachusetts Institute of Technology, 1970.

[21] N. Yoshida. Utterance segmenation for spontaneous speech recognition. Master's thesis, Massachusetts Institute of Technology, 2002.