A Probabilistic Model for Understanding Composite Spoken Descriptions

Enes Makalic, Ingrid Zukerman, Michael Niemann, and Daniel Schmidt

Faculty of Information Technology, Monash University
Clayton, VICTORIA 3800, AUSTRALIA
{enes,ingrid,niemann,dschmidt}@csse.monash.edu.au

Abstract. We describe a probabilistic reference disambiguation mechanism developed for a spoken dialogue system mounted on an autonomous robotic agent. Our mechanism receives as input referring expressions containing intrinsic features of individual concepts (lexical item, size and colour) and features involving more than one concept (ownership and location). It then performs probabilistic comparisons between the given features and features of objects in the domain, yielding a ranked list of candidate referents. Our evaluation shows high reference resolution accuracy across a range of spoken referring expressions.

1 Introduction

In this paper, we describe the reference disambiguation mechanism of *Scusi?* — the spoken language interpretation module of a robot-mounted dialogue agent. Our mechanism interprets referring expressions such as "the blue mug on the table near the lamp" by performing probabilistic comparisons between the requirements stated in a referring expression and the features of candidate objects (e.g., those in the room).

The contributions of our mechanism are (1) probabilistic procedures that perform feature comparisons; and (2) a function that combines the results of these comparisons. These contributions endow our mechanism with the ability to handle imprecise or ambiguous referring expressions. For instance, the expression "the bag near the green table" is ambiguous if there is a bag *on* a green table, and there is a bag next to a table that isn't green. Such candidate objects are ranked according to how well they match the specifications in an utterance. Our system handles the following feature types: lexical item, colour, size, ownership and location. Our evaluation shows that our mechanism exhibits high resolution accuracy for different types of referring expressions.

This paper is organized as follows. Section 2 outlines the interpretation process and the estimation of the probability of an interpretation. Section 3 describes the probabilistic feature comparison. The results of our evaluation appear in Section 4. Related research and concluding remarks are given in Sections 5 and 6 respectively.

2 Interpretation Process

Scusi? processes spoken input in three stages: speech recognition, parsing and semantic interpretation (Figure 1). First, it runs Automatic Speech Recognition (ASR) software (Microsoft Speech SDK 5.1) to generate candidate texts from a speech signal.

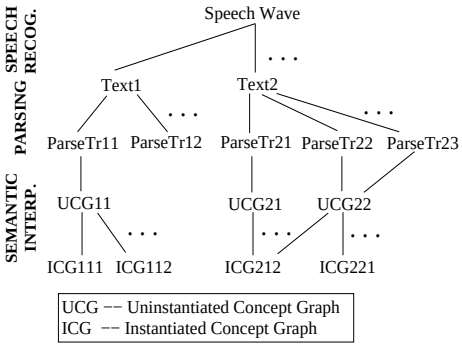

Fig. 1. Stages of the interpretation process

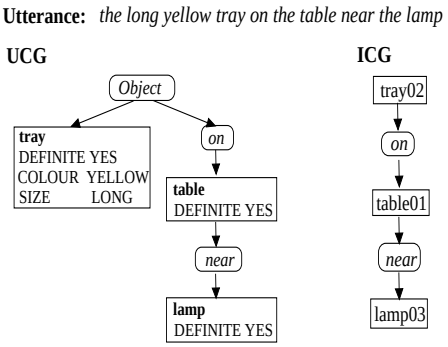

Fig. 2. UCG and ICG for a sample utterance

Each text is assigned a score that reflects the probability of the words given the speech wave. Next, *Scusi?* applies Charniak's probabilistic parser (`ftp://ftp.cs.brown.edu/pub/nlparser/`) to generate parse trees from the texts. The parser produces up to N ($= 50$) parse trees for each text, associating each parse tree with a probability.

During semantic interpretation, parse trees are successively mapped into two representations based on Conceptual Graphs [1]: first *Uninstantiated Concept Graphs (UCGs)*, and then *Instantiated Concept Graphs (ICGs)* (Figure 2). UCGs are obtained from parse trees deterministically — one parse tree generates one UCG. A UCG represents syntactic information, where the concepts correspond to the words in the parent parse tree, and the relations are derived from syntactic information in the parse tree and prepositions. Each UCG can generate many ICGs. This is done by nominating different instantiated concepts and relations from the system's knowledge base as potential realizations for each concept and relation in a UCG.

Our interpretation process applies a selection-expansion cycle to build a search graph, where each level of the graph corresponds to one of the stages of the interpretation process (Figure 1). In each selection-expansion cycle, our algorithm selects an option for consideration (speech wave, textual ASR output, parse tree or UCG). At any point after an expansion, *Scusi?* can return a list of ranked interpretations (ICGs) with their parent sub-interpretations (text, parse tree(s) and UCG(s)).

Figure 2 illustrates a UCG and an ICG for an utterance containing the composite referring expression ("the long yellow tray on the table near the lamp"). The *intrinsic* features of an object (e.g., colour and size of the tray) are stored in the UCG node for this object. In contrast, *structural* features, which involve at least two objects (e.g., "the *table* near the *lamp*"), are represented as sub-graphs of the UCG (and then the ICG). This distinction is made because intrinsic features can be compared directly to features of objects in the knowledge base, while features that depend on the relationship between several objects require the identification of these objects and the verification of this relationship. In our example, all the tables and all the lamps in the room need to be considered, and the table/lamp combination that best matches the given specification is eventually selected. The procedures for selecting objects that match intrinsic and structural features are described in Section 3.

2.1 Estimating the probability of an ICG

Scusi? ranks candidate ICGs according to their probability of being the intended meaning of a spoken utterance. Given a speech signal W and a context $\mathcal{C}$, the probability of an ICG I is represented as follows.

$$\Pr(I|W,\mathcal{C}) \propto \sum_{\Lambda} \Pr(I|U,\mathcal{C}) \cdot \Pr(U|P) \cdot \Pr(P|T) \cdot \Pr(T|W) \tag{1}$$

where U, P and T denote a UCG, parse tree and text respectively.

The summation is taken over all possible paths $\Lambda = \{P, U\}$ from a parse tree to the ICG, because a UCG and an ICG can have more than one parent. As mentioned above, the ASR and the parser return an estimate of $\Pr(T|W)$ and $\Pr(P|T)$ respectively; and $\Pr(U|P) = 1$, since the process of generating a UCG from a parse tree is deterministic. The estimation of $\Pr(I|U,\mathcal{C})$ is described in detail in [2]. Here we present the final equation obtained for $\Pr(I|U,\mathcal{C})$, and outline the ideas involved in its calculation.

$$\Pr(I|U,\mathcal{C}) \approx \prod_{k \in I} \Pr(u|k) \Pr(k|k_p, k_{gp}) \Pr(k|\mathcal{C}) \tag{2}$$

where k is an instantiated node in ICG I, u is the corresponding node in UCG U, k_p is the parent node of k in ICG I, and k_{gp} the grandparent node. For example, *near* is the parent of `lamp03`, and `table01` the grandparent of `lamp03` in the ICG in Figure 2.

- $\Pr(u|k)$ is the "match probability" between the specifications for node u in UCG U and the features of the corresponding node k in ICG I, e.g., how similar an object in the room is to the "long yellow tray" (Section 3.1).
- $\Pr(k|k_p, k_{gp})$ represents the structural probability of ICG I, simplified to node trigrams, e.g., whether `table01` is *near* `lamp03` (Section 3.2).
- $\Pr(k|\mathcal{C})$ is the probability of a concept in light of the context, which at present includes only domain knowledge.

3 Probabilistic Feature Comparison

Scusi? handles three intrinsic features, viz lexical item, colour and size; and two structural features, viz ownership and several types of locative references. The procedure for generating ICGs for a referring expression and calculating their probability is described in Algorithm 1. First, the intrinsic features of the objects in our world are used to calculate the probability of a match with each UCG concept (first factor in Equation 2, Step 2 in Algorithm 1). These probabilities are used to build a list of candidate objects that are a reasonable match for each UCG concept (Step 3). The objects in each list are iteratively combined into candidate ICGs, where each candidate represents an interpretation of the referring expression (Step 5). *Scusi?* then considers the structural features of each ICG to calculate its structural probability (Step 7, second factor in Equation 2), and combines the intrinsic and structural probabilities to calculate the probability of the ICG (Step 8). Finally, the ICGs are ranked according to their probability (Step 10).

For example, consider a request for "the blue mug on the table", assuming that the knowledge base contains several mugs, some of which are blue. First, for all the objects

Algorithm 1 Generate candidate ICGs for a referring expression

Require: UCG U comprising concepts and relations u, knowledge base $\mathcal{K}$ of objects

1: **for all** objects $u \in$ UCG U **do**
2: Estimate Pr$(u|k)$, the probability of the match between the features of u and those of each object $k \in \mathcal{K}$.
3: Rank the candidate objects $k \in \mathcal{K}$ in descending order of probability.
4: **end for**
5: Construct candidate ICGs by iteratively going down the list of objects generated for each concept u in UCG U — each candidate ICG contains one object from each list.
6: **for all** ICGs I **do**
7: Estimate the probabilities Pr$(k|k_p, k_{gp})$ for each *object-relation-object* trigram in I.
8: Combine these estimates with the probabilities from Step 2 to obtain the probability of I.
9: **end for**
10: Rank the candidate ICGs in descending order of probability.

in the knowledge base, we estimate the probability that they could be called 'mug' (e.g., mugs, cups), and the probability that their colour could be considered 'blue'; similarly, we calculate the probability that an object could be called 'table' (Section 3.1). The candidates for 'blue mug' and 'table' are then ranked in descending order of probability. Candidate ICGs are built by iteratively combining each candidate blue mug with each candidate table. The structural probability of each ICG is then calculated on the basis of the location coordinates of the mug and table instances in the ICG (Section 3.2).

At present, we make the following simplifying assumptions: (1) the robot is co-present with the user and the possible referents of an utterance; and (2) the robot has an unobstructed view of the objects in the room and up-to-date information about these objects. This information could be obtained through a scene analysis system [3] activated upon entering a room. These assumptions obviate the need for planning physical actions, such as moving to get a better view of certain objects, or leaving the room to seek objects that better match the given specifications.

3.1 Estimating the Probabilities of Intrinsic Features

The probability of the match between a node u specified in UCG U and a candidate instantiated concept $k \in \mathcal{K}$ (Step 2 of Algorithm 1) is estimated as follows.

$$\Pr(u|k) = \Pr(\mathbf{u}_{f_1}, \ldots, \mathbf{u}_{f_p} | \mathbf{k}_{f_1}, \ldots, \mathbf{k}_{f_p}) \tag{3}$$

where $(f_1, \ldots, f_p) \in \mathcal{F}$ are the features specified with respect to node u, $\mathcal{F}$ is the set of features allowed in the system, $\mathbf{u}_{f_i}$ is the value of the i-th feature of UCG node u, and $\mathbf{k}_{f_i}$ is the value of this feature for the instantiated concept k.

Assuming that the features of a node are independent, the probability that an instantiated concept k matches the specifications in a UCG node u can be rewritten as

$$\Pr(u|k) = \prod_{i=1}^{p} \Pr(\mathbf{u}_{f_i} | \mathbf{k}_{f_i}) \tag{4}$$

In the absence of other information, it is reasonable to use a linear distance function $h : \mathbb{R}^+ \rightarrow [0, 1]$ to map the outcome of a feature match to the probability space. That is,

the higher the similarity between requested and instantiated feature values (the shorter the distance between them), the higher the probability of a feature match. Specifically,

$$\Pr(\mathbf{u}_f|\mathbf{k}_f) = h_f(\mathbf{u}_f, \mathbf{k}_f) \tag{5}$$

Below we present the calculation of Equation 5 for the intrinsic features supported by our system (lexical item, colour and size). In agreement with [4, 5], lexical item and colour are considered *absolute* features, and size a *relative* feature (its value depends on the size of other candidates).

Lexical item. We employ the Leacock and Chodorow [6] similarity measure, denoted *LC*, to compute the similarity between the lexical feature of u and k. This measure is applied to the words in a database constructed with the aid of WordNet (the *LC* measure yielded the best results among those in [7]). The *LC* similarity score, denoted s_{LC}, is converted to a probability by applying the following h_{lex} function.

$$\Pr(\mathbf{u}_{\text{lex}}|\mathbf{k}_{\text{lex}}) = h_{lex}(s_{LC}(\mathbf{u}_{\text{lex}}, \mathbf{k}_{\text{lex}})) = \frac{s_{LC}(\mathbf{u}_{\text{lex}}, \mathbf{k}_{\text{lex}})}{s_{\max}}$$

where $s_{\max}$ is the highest possible *LC* score.

Colour. The colour model chosen for *Scusi?* is the CIE 1976 (L, a, b) colour space, which has been experimentally shown to be approximately perceptually uniform [8]. The L coordinate represents brightness ($L = 0$ denotes black, and $L = 100$ white), a represents position between green ($a < 0$) and red ($a > 0$), and b position between blue ($b < 0$) and yellow ($b > 0$). The range of L is $[0, 100]$, while for practical purposes, the range of a and b is $[-200, 200]$. Thus, the probability of a colour match between a UCG concept u and an instantiated concept k is

$$\Pr(\mathbf{u}_{\text{colr}}|\mathbf{k}_{\text{colr}}) = h_{colr}(\mathbf{u}_{\text{colr}}, \mathbf{k}_{\text{colr}}) = 1 - \frac{ED(\mathbf{u}_{\text{colr}}, \mathbf{k}_{\text{colr}})}{d_{\max}}$$

where *ED* is the Euclidean distance between the (L, a, b) coordinates of the colour specified for u and the (L, a, b) coordinates of the colour of k, and $d_{\max}$ is the maximum Euclidean distance between two colours (=574.5).

Size. Unlike lexical item and colour, size is considered a relative feature, i.e., the probability of a size match between an object $k \in \mathcal{K}$ and a UCG concept u depends on the sizes of all suitable candidate objects in $\mathcal{K}$ (those that have a reasonable match for lexical and colour comparisons). The highest probability for a size match is then assigned to the object that best matches the required size, while the lowest probability is assigned to the object which has the worst match with this size.

This requirement is achieved by the following h_{size} function, which like Kelleher *et al.*'s pixel-based mapping [9], performs a linear mapping between $\mathbf{u}_{\text{size}}$ and $\mathbf{k}_{\text{size}}$.

$$\Pr(\mathbf{u}_{\text{size}}|\mathbf{k}_{\text{size}}) = h_{size}(\mathbf{u}_{\text{size}}, \mathbf{k}_{\text{size}}) = \begin{cases} \frac{\alpha \mathbf{k}_{\text{size}}}{\max_i\{\mathbf{k}^i_{\text{size}}\}} & \text{if } \mathbf{u}_{\text{size}} \in \{\text{`large'/`big'/}\ldots\} \\ \frac{\alpha \min_i\{\mathbf{k}^i_{\text{size}}\}}{\mathbf{k}_{\text{size}}} & \text{if } \mathbf{u}_{\text{size}} \in \{\text{ `small'/`little'/}\ldots\} \end{cases}$$

where α is a normalizing constant, and $\mathbf{k}^i_{\text{size}}$ is the size of candidate object k^i (this formula is adapted for individual dimensions, e.g., length).

Combining Feature Scores. To determine how intrinsic features are used in our domain, we conducted a survey where people were asked to refer to household objects laid out in a space [10]. The results of our survey agree with Dale and Reiter's findings [4], whereby people often present features that are not strictly necessary to identify an item, and use features in the following order of frequency: *type* $\succ$ *absolute adjectives* $\succ$ *relative adjectives*, where colour is an absolute feature and size is a relative feature.

These findings prompted us to incorporate a weighting scheme into Equation 4, whereby features are weighted according to their usage in referring expressions. That is, higher ranking or more frequently used features are assigned a higher weight than lower ranking or less frequently used features. Specifically, given a match probability $\Pr(\mathbf{u}_{f_i}|\mathbf{k}_{f_i})$ and a weight w_{f_i} for feature f_i ($0 < w_{f_i} \leq 1$), the adjusted match probability for this feature is

$$\Pr'(\mathbf{u}_{f_i}|\mathbf{k}_{f_i}) = \Pr(\mathbf{u}_{f_i}|\mathbf{k}_{f_i}) \times w_{f_i} + \tfrac{1}{2}(1 - w_{f_i})$$

The effect of this mapping is that features with high weights have a wide range of probabilities (and hence a substantial influence on the match probability of an object), while features with low weights have a narrow range (and a reduced influence on match probability).

3.2 Estimating the Probabilities of Structural Features

As shown in Equation 2, the overall probability of an ICG structure can be decomposed into a product of the probabilities of the trigrams that make up the ICG. A trigram consists of a relationship k_p (e.g., ownership or location) and two instantiated concepts k and k_{gp}, e.g., `table01`–*near*–`lamp03` in Figure 2. A probability is assigned to this trigram based on the physical coordinates of `table01` and `lamp03`.

Below we present the probability calculation for the two structural features supported by our system (ownership and location). This calculation involves validating the structural feature against the information in our world, and, as for intrinsic features, performing a linear mapping from the result of this validation to a probability.

Ownership. In our world, an object is either owned by one or more people or no owner has been recorded for this object. This leads to a simple probabilistic mapping.

$$\Pr(\mathbf{k}|\textit{own}, \mathbf{k}_{gp}) = \begin{cases} 0 \text{ if } \mathbf{k} \notin \text{owner-of}(\mathbf{k}_{gp}) \\ \beta \text{ if owner-of}(\mathbf{k}) \text{ is unknown} \\ 1 \text{ if } \mathbf{k} \in \text{owner-of}(\mathbf{k}_{gp}) \end{cases}$$

where β is currently set to 0.5.

Location. At present, we assume that all the objects in our world are rigid, and hence can be represented by a circumscribing box, e.g., a lamp is represented by the smallest box that contains the lamp. As a result, each object $\mathbf{k}$ has three dimensions and one position coordinate. The dimensions are $(\mathbf{k}_l, \mathbf{k}_w, \mathbf{k}_h)$, corresponding to the object's length, width and height respectively. The position coordinate is $(\mathbf{k}_x, \mathbf{k}_y, \mathbf{k}_z)$, measured between a starting coordinate $(0, 0, 0)$ and the closest corner of the box. The system handles the following locative prepositions: *on, under, above, in (inside)* and *near (by)*.

– ***on, under, above*** **–** these prepositions have the following directional semantics.
  - ***on*** means that $\mathbf{k}_{gp_z} = \mathbf{k}_z + \mathbf{k}_h$, where $\mathbf{k}_z + \mathbf{k}_h$ represents the height of the top surface of the bounding box for object $\mathbf{k}$;
  - ***under*** means that $\mathbf{k}_{gp_z} + \mathbf{k}_{gp_h} \leq \mathbf{k}_z$; and
  - ***above*** means that $\mathbf{k}_{gp_z} > \mathbf{k}_z + \mathbf{k}_h$.

  If the objects $\mathbf{k}$ and $\mathbf{k}_{gp}$ in an ICG satisfy the directional requirement of their location preposition (*loc*$\in\{$*on,under,above*$\}$), we say that Pr($\mathbf{k}|$*loc*$, \mathbf{k}_{gp}$) is proportional to the area shared by the horizontal surfaces of (the bounding boxes of) the two objects. Otherwise, Pr($\mathbf{k}|$*loc*$, \mathbf{k}_{gp}$) is set to a low probability (ϵ). Specifically, let $A(\mathbf{k})$ denote the area of the top face of object $\mathbf{k}$, and let $A(\mathbf{k}, \mathbf{k}_{gp})$ denote the overlapping area between the top faces of objects $\mathbf{k}$ and $\mathbf{k}_{gp}$ in the xy plane. The probability of a trigram involving location relations *on*, *under* or *above* is

$$\Pr(\mathbf{k}|\textit{loc}, \mathbf{k}_{gp}) = \begin{cases} \frac{A(\mathbf{k},\mathbf{k}_{gp})}{\min\{A(\mathbf{k}),A(\mathbf{k}_{gp})\}} & \text{if directional requirement is satisfied} \\ \epsilon \ll 0.1 & \text{otherwise} \end{cases}$$

  For example, consider the utterance "the book on the table", for which one of the candidate ICGs is `book01`→*on*→`table02`. The directional semantics for *on* stipulate that the z coordinate of (the bottom of) the book ($\mathbf{k}_{gp}$) must be equal to the z coordinate of the table ($\mathbf{k}$) plus the height of the table. If this condition is satisfied, then the degree of overlap between the surface of the book and that of the table is calculated. That is, a book that is entirely on a table top satisfies the *on* relationship with a higher probability than a book overhanging the table.
– ***in (inside)*** **–** the probability of an object being inside another is proportional to the volume shared by their bounding boxes (one object could be partially inside another). Formally, let $V(\mathbf{k})$ denote the volume of (the bounding box of) object $\mathbf{k}$, and let $V(\mathbf{k}, \mathbf{k}_{gp})$ denote the shared volume between (the bounding boxes of) objects $\mathbf{k}$ and $\mathbf{k}_{gp}$. The probability of an *in*-trigram is

$$\Pr(\mathbf{k}|\textit{in}, \mathbf{k}_{gp}) = \frac{V(\mathbf{k}, \mathbf{k}_{gp})}{\min\{V(\mathbf{k}), V(\mathbf{k}_{gp})\}}$$

  For example, if we are asked for "the mug inside the box", a mug that is wholly contained within a box would yield a higher probability than a mug whose top exceeds the top of a box.
– ***near*** **–** following [9], we employ a formulation inspired by the gravitational model to calculate the probability of two objects being near each other. However, since the density of objects is not specified in our world, we approximate the mass of an object by its volume. Formally, let $d(\mathbf{k}, \mathbf{k}_{gp})$ represent the shortest distance between the bounding boxes of $\mathbf{k}$ and $\mathbf{k}_{gp}$. The probability of a *near*-trigram is

$$\Pr(\mathbf{k}|\textit{near}, \mathbf{k}_{gp}) = \frac{V(\mathbf{k})V(\mathbf{k}_{gp})}{d^2(\mathbf{k}, \mathbf{k}_{gp})\, G_{max}}$$

  where G_{max}, the maximum gravitational pull in our world, is obtained when the two biggest objects in our world abut (i.e., d is arbitrarily small).
  This model enables the size of the objects to influence the nearness probability. For example, if one asks for "the ball next to the table", and there is a tennis ball a few centimeters from the table, and a beach ball farther from the table, this model will identify the ambiguity, and support the generation of a clarification question.

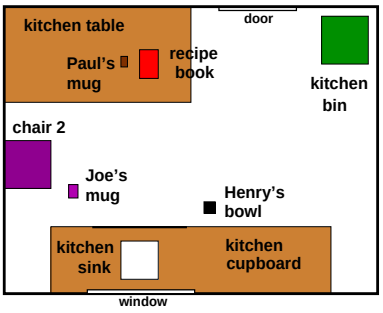

Fig. 3. Sample area from our world

1 A desk
2 The purple bowl
3 Paul's book
4 The green mug in the lounge
5 Sarah's bowl in the lounge
6 The long pants in the bathroom
7 The wardrobe under the fan
8 A bin near the small plant
9 The mug near the book on the table
10 The shirt in the bag near the plant

Fig. 4. Sample referring expressions

4 Evaluation

To evaluate our system, we constructed a simulated world that represents an open-plan house (in keeping with our co-presence assumption, Section 3). The world contains 54 objects distributed among four areas in the house, and five people (Figure 3 illustrates one of the areas of the house, with various objects labeled). The objects were chosen so that they had similar features, i.e., several objects could be referred to by the same lexical item (there were 2-4 instances of each type of object), had similar colours and sizes, and were placed in adjacent locations. The ownership of most objects was distributed among the five people in our world, and some objects had no known owner.

In total, 90 referring expressions of varying complexity were used for the system evaluation. Each referring expression consisted of a noun phrase comprising between one and three concepts (sample expressions are shown in Figure 4). Mean utterance length was 4.27 words, with a maximum length of 8 words. The expressions were constructed to test *Scusi?*'s ability to identify target objects (the intended book, mug, table, etc) in different situations. Specifically, objects were referred to by near synonyms (e.g., "mug" and "cup"), by colours and sizes that were shared by several objects, and by their proximity to a reference object that was adjacent to several objects. For example, the utterance "the book near Paul's mug" tests the system's ability to identify an object by its ownership and location in a world that contains several books and mugs.

Scusi? was set to generate at most 300 sub-interpretations in total (including texts, parse trees, UCGs and ICGs) for each referring expression. An ICG proposed by *Scusi?* was deemed correct if it matched the speaker's intention, which was represented by one or more Gold ICGs. These ICGs were manually constructed by one of the authors for each referring expression on the basis of the information in *Scusi?*'s knowledge base. Multiple Gold ICGs were allowed if several objects in the domain matched a specified object, e.g., "a bowl". A baseline measure of performance was obtained by executing a beam search. That is, only the top-ranked ASR result was parsed, and only the top-ranked parse tree yielded a UCG, which in turn produced only one top-ranked ICG.

Table 1 summarizes our results. Column 1 shows the procedure (*Scusi?*'s or baseline). Columns 2 and 3 show how many of the descriptions had Gold ICG referents whose probability was the highest (top 1) or among the three highest (top 3), e.g., *Scusi?* yielded 82 Gold ICGs with the top probability, and all the 90 referring expressions had

Table 1. *Scusi?*'s interpretation performance

	# Gold ICGs with prob in top 1	# Gold ICGs with prob in top 3	Average adj rank(rank)	Not found	Avg # to Gold ICGs (iters)
BASELINE	44	44	0 (0)	46	0 (4)
Scusi?	82	90	0.96 (0.11)	0	2.45 (25)

Gold referents within the top 3 probabilities. The average *adjusted rank* and *rank* of the Gold referent appear in Column 4. The rank of a referent r is its position in a list sorted in descending order of probability (starting from position 0), such that all equiprobable referents are deemed to have the same position. The adjusted rank of a referent r is the mean of the positions of all referents that have the same probability as r. For example, if we have 3 top-ranked equiprobable referents, each has a rank of 0, but an adjusted rank of $\frac{0+2}{3}$. Column 5 indicates the number of referring expressions for which a Gold ICG was not found, and Column 6 shows the average number of referents created and iterations performed until the Gold referent was found (from a total of 300 iterations).

Our results show that maintaining multiple hypotheses at each stage of the interpretation process yields a substantial improvement in interpretation accuracy in comparison to the baseline approach. *Scusi?* found the Gold interpretation for all 90 utterances tested, in contrast to the baseline approach, which found only 44 Gold ICGs. The average rank of the correct text in the output returned by the ASR was 1.5 (where the top rank is 0), and the correct text was top ranked by the ASR in 70% of the cases. This level of accuracy is higher than the accuracy of the baseline approach with respect to Gold ICGs ($44/90 = 49\%$), which indicates that even when presented with the correct text, the baseline approach may not find the intended interpretation. Furthermore, ASR accuracy is lower than *Scusi?*'s accuracy for top-1 Gold ICGs ($82/90 = 91\%$), demonstrating the robustness of the probabilistic multi-stage interpretation process in the face of ASR inaccuracy.

5 Related Research

Reference disambiguation is an essential aspect of discourse understanding to which a large research effort has been devoted. Much of the research on reference resolution has focused on the generation of referring expressions, which involves constructing expressions that single out a target object from a set of distractors, e.g., [4, 5]. Methods for understanding referring expressions in dialogue systems are examined in [9, 11] among others. Kelleher *et al.* [9] propose a reference resolution algorithm that accounts for four attributes: lexical type, colour, size and location, where the score of an object is estimated by a weighted combination of the visual and linguistic salience scores of each attribute. Like in *Scusi?*, the values of the weights are pre-defined and based on empirical observations. However, Kelleher *et al.* limit the probabilistic comparison of features to size and location, and use binary comparisons for lexical item and colour. Pfleger *et al*. [11] use modality fusion to combine hypotheses from different analyzers (linguistic, visual and gesture), choosing as the referent the first object satisfying a 'differentiation criterion'. As a result, their system does not handle situations where more than one object satisfies this criterion.

6 Conclusion

We have offered a probabilistic reference disambiguation mechanism which considers intrinsic and structural features. Our mechanism performs probabilistic comparisons between features specified in referring expressions (specifically lexical item, colour, size, ownership and location) and features of objects in the domain. Our mechanism was empirically evaluated for these features, exhibiting very good interpretation performance for a range of referring expressions.

In the future, we propose to extend the weighting mechanism devised for intrinsic features to cater for structural features and their combination with intrinsic features. We also propose to integrate our mechanism with a vision system, which will affect the type of information we can obtain from our knowledge base. Finally, we intend to remove the co-presence and unobstructed-view assumptions (Section 3), which will demand the integration of our feature comparison mechanism with planning procedures.

References

1. Sowa, J.: Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley, Reading, MA (1984)
2. Zukerman, I., Makalic, E., Niemann, M., Schmidt, D.: A probabilistic approach to the interpretation of spoken utterances. In: PRICAI 2008 – Proceedings of the Tenth Pacific Rim International Conference on Artificial Intelligence, Hanoi, Vietnam (2008)
3. Makihara, Y., Takizawa, M., Shirai, I., Miura, J., Shimada, N.: Object recognition supported by user interaction for service robots. In: Proceedings of the 16th International Conference on Pattern Recognition. Volume 3., Quebec, Canada (2002) 561–564
4. Dale, R., Reiter, E.: Computational interpretations of the Gricean maxims in the generation of referring expressions. Cognitive Science **18**(2) (1995) 233–263
5. Wyatt, J.: Planning clarification questions to resolve ambiguous references to objects. In: Proceedings of the 4th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, Edinburgh, Scotland (2005) 16–23
6. Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. In Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press (1998) 265–285
7. Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet::Similarity – measuring the relatedness of concepts. In: AAAI-04 – Proceedings of the 19th National Conference on Artificial Intelligence, San Jose, California (2004) 25–29
8. Puzicha, J., Buhmann, J., Rubner, Y., Tomasi, C.: Empirical evaluation of dissimilarity measures for color and texture. In: Proceedings of the 7th IEEE International Conference on Computer Vision. Volume 2., Kerkyra, Greece (1999) 1165–1172
9. Kelleher, J., Kruijff, G., Costello, F.: Proximity in context: an empirically grounded computational model of proximity for processing topological spatial expressions. In: COLING-ACL'06 Proceedings, Sydney, Australia (2006) 745–752
10. Zukerman, I., Makalic, E., Niemann, M.: Using probabilistic feature matching to understand spoken descriptions. In: AI'08 Proceedings – the 21st Australasian Joint Conference on Artificial Intelligence, Auckland, New Zealand (2008)
11. Pfleger, N., Alexandersson, J., Becker, T.: A robust and generic discourse model for multimodal dialogue. In: Proceedings of the 3rd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, Acapulco, Mexico (2003)