Interpreting Pointing Gestures and Spoken Requests – A Probabilistic, Salience-based Approach

Ingrid Zukerman and **Gideon Kowadlo** and **Patrick Ye**
Faculty of Information Technology
Monash University
Ingrid.Zukerman@monash.edu, gkowadlo@gmail.com, ye.patrick@gmail.com

Abstract

We present a probabilistic, salience-based approach to the interpretation of pointing gestures together with spoken utterances. Our mechanism models dependencies between spatial and temporal aspects of gestures and features of utterances. For our evaluation, we collected a corpus of requests which optionally included pointing. Our results show that pointing information improves interpretation accuracy.

1 Introduction

DORIS (Dialogue Oriented Roaming Interactive System) is a spoken dialogue system designed for a household robot. In (Zukerman et al., 2008), we described *Scusi?* — a spoken language interpretation module which considers multiple sub-interpretations at different levels of the interpretation process, and estimates the probability of each sub-interpretation at each level (Section 2). This formalism is required for requests such as "Get me the blue cup" in the context of the scene depicted in Figure 1, where possible candidates are the three white cups, and the blue and purple tumblers, but it is unclear which is the intended object, as none of the alternatives match the request perfectly.

In this paper, we integrate pointing gestures into *Scusi?*'s probabilistic formalism. We adopt a salience-based approach, where we take into account spatial and temporal information to estimate the probability that a pointing gesture refers to an object. To evaluate our formalism, we collected a corpus of requests where people were allowed to point (Section 4). Our results show that when people point, our mechanism yields significant improvements in interpretation accuracy; and when pointing was artificially added to utterances where the people did not point, its effect on interpretation accuracy was reduced.

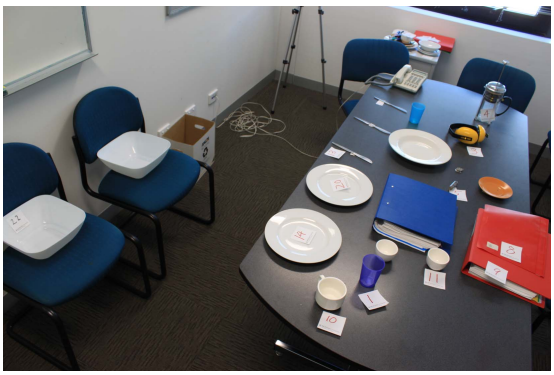


Figure 1: Experimental Setup

This paper is organized as follows. Section 2 outlines the interpretation of a spoken request and the estimation of the probability of an interpretation. Section 3 describes how pointing affects this probability. Our evaluation is detailed in Section 4. Related research is discussed in Section 5, followed by concluding remarks.

2 Interpreting Spoken Requests

Here we summarize our previous work on the interpretation of single-sentence requests (Makalic et al., 2008; Zukerman et al., 2008).

Coling 2008: Poster Volume, pages 1558–1566,
Beijing, August 2010

Scusi? processes spoken input in three stages: speech recognition, parsing and semantic interpretation. First, Automatic Speech Recognition (ASR) software (Microsoft Speech SDK 5.3) generates candidate hypotheses (texts) from a speech signal. The ASR produces up to 50 texts for a spoken utterance, where each text is associated with a probability. In the parsing stage, the texts are considered in descending order of probability. Charniak's probabilistic parser (`ftp://ftp.cs.brown.edu/pub/nlparser/`) is applied to each text, yielding up to 50 parse trees — each associated with a probability.

During semantic interpretation, parse trees are successively mapped into two representations based on Concept Graphs (Sowa, 1984). First *Uninstantiated Concept Graphs (UCGs)*, and then *Instantiated Concept Graphs (ICGs)*. 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. Instantiated concepts are objects and actions in the domain (e.g., `mug01`, `mug02` and `cup01` are possible instantiations of the uninstantiated concept "mug"), and instantiated relations are similar to semantic role labels (Gildea and Jurafsky, 2002). The interpretation process continues until a preset number of sub-interpretations (including texts, parse trees, UCGs and ICGs) has been generated or all options have been exhausted.

Figure 2 illustrates a UCG and an ICG for the request "get the large red folder on the table". The *intrinsic* features of an object (lexical item, colour and size) are stored in the UCG node for this object. *Structural* features, which involve two objects (e.g., "*folder*-on-*table*"), are represented as sub-graphs of the UCG (and the ICG).

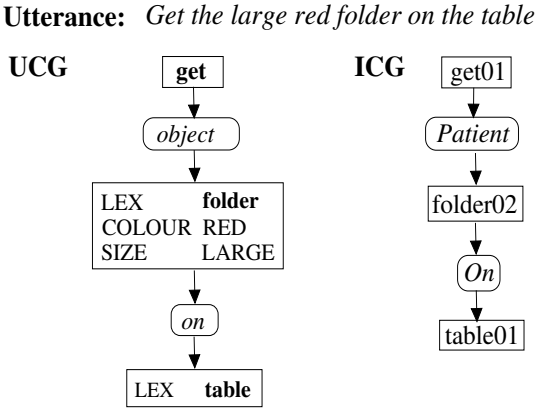


Figure 2: UCG and ICG for a sample utterance

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, $\Pr(I|W,\mathcal{C})$, is proportional to

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

where T, P and U denote text, parse tree and UCG respectively. The summation is taken over all possible paths $\Lambda = \{P, U\}$ from the speech wave to the ICG, because a UCG and an ICG can have more than one ancestor. 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 (Zukerman et al., 2008). Here we present the final equation obtained for $\Pr(I|U,\mathcal{C})$, and outline the ideas involved in its calculation.

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

where u is a node in UCG U, k is the corresponding instantiated node in ICG I, k_p is k's parent node, and k_{gp} is k's grandparent node. For example, *On* is the parent of `table01`, and `folder02` the grandparent in the ICG in Figure 2.

- $\Pr(u|k)$ is the "match probability" between the specifications for node u in UCG U and the intrinsic features of the corresponding node k in ICG I, i.e., the probability that a speaker who intended a particular object k gave the specifications in u.

- $\Pr(k|k_p, k_{gp})$ represents the structural probability of ICG I, where structural information is simplified to node trigrams, e.g., whether `folder02` is *On* `table01`.

- $\Pr(k|\mathcal{C})$ is the probability of a concept in light of the context, which includes information about domain objects, actions and relations.

Scusi? handles three intrinsic features: lexical item, colour and size; and two structural features: ownership and several locative relations (e.g., on, under, near). The match probability $\Pr(u|k)$ and the structural probability $\Pr(k|k_p, k_{gp})$ are estimated using a distance function between the requirements specified by the user and what is found in reality — the closer the distance between the specifications and reality, the higher the probability (for details see (Makalic et al., 2008)).

3 Incorporating Pointing Gestures

Pointing affects the salience of objects and the language used to refer to objects: objects in the temporal and spatial vicinity of a pointing gesture are more salient than objects that are farther away, and pointing is often associated with demonstrative determiners. Thus, the incorporation of pointing into *Scusi?* affects the following elements of Equation 2 (Section 2.1).

- $\Pr(k|\mathcal{C})$ – the context-based probability of an object (i.e., its salience) is affected by the time of a pointing gesture and the space it encompasses. For instance, if the user says "Get the cup" in the context of the scene in Figure 1, pointing around the time s/he said "cup", the gesture most likely refers to an object that may be called "cup". Further, among the candidate cups in Figure 1, those closer to the "pointing vector" have a higher probability.[1]

- $\Pr(u|k, \mathcal{C})$ – when pointing, people often use demonstrative determiners, e.g., "get me *that* cup". Also, people often use generic identifiers in conjunction with demonstrative determiners to refer to unfamiliar objects, e.g., "that thing" to refer to a vacuum tube (Figure 1).

These probabilities are estimated in Sections 3.1 and 3.2. Our calculations are based on information returned by the gesture recognition system described in (Li and Jarvis, 2009): gesture type, time, probability and relevant parameters (e.g., a vector for a pointing gesture). Since we focus on pointing gestures, we convert the probabilities expected from Li and Jarvis's system into the probability of Pointing and that of Not Pointing, which comprises all other gestures and no gesture (these hypotheses are returned at the same time).[2]

3.1 Calculating salience from pointing

When pointing is taken into account, the probability of object k is expressed as follows.

$$\begin{aligned}\Pr(k|\mathcal{C}) &= \Pr(k|\mathcal{P}, \mathcal{C}) \cdot \Pr(\mathcal{P}|\mathcal{C}) + \\ &\quad \Pr(k|\neg\mathcal{P}, \mathcal{C}) \cdot \Pr(\neg\mathcal{P}|\mathcal{C})\end{aligned} \tag{3}$$

where $\mathcal{P}$ designates Pointing, $\Pr(\mathcal{P}|\mathcal{C})$ and its complement are returned by the gesture recognition system, and $\Pr(k|\neg\mathcal{P}, \mathcal{C}) = \frac{1}{N}$ (N is the number of objects in the room, i.e., in the absence of pointing, we assume that all the objects in the room are equiprobable[3]).

As indicated above, we posit that pointing is spatially correlated with an intended object, and temporally correlated with a word referring to the intended object. Hence, we separate Pointing into two components: spatial (s) and temporal (t), obtaining $\langle \mathcal{P}_s, \mathcal{P}_t \rangle$. Thus

$$\begin{aligned}\Pr(k|\mathcal{P}, \mathcal{C}) &= \frac{\Pr(k, \mathcal{P}_t, \mathcal{P}_s, \mathcal{C})}{\Pr(\mathcal{P}, \mathcal{C})} \\ &= \frac{\Pr(\mathcal{P}_t|k, \mathcal{P}_s, \mathcal{C}) \cdot \Pr(k|\mathcal{P}_s, \mathcal{C}) \cdot \Pr(\mathcal{P}_s|\mathcal{C})}{\Pr(\mathcal{P}|\mathcal{C})}\end{aligned} \tag{4}$$

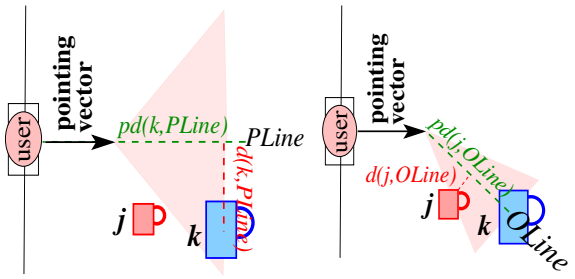
We assume that given k, $\mathcal{P}_t$ is conditionally independent from $\mathcal{P}_s$; and that $\Pr(\mathcal{P}_s|\mathcal{C}) = \Pr(\mathcal{P}|\mathcal{C})$, i.e., the spatial probability of a pointing gesture is the probability returned by the gesture system for the entire pointing hypothesis (time and space). This yields

$$\Pr(k|\mathcal{P}, \mathcal{C}) = \Pr(\mathcal{P}_t|k, \mathcal{C}) \cdot \Pr(k|\mathcal{P}_s, \mathcal{C}) \tag{5}$$

[1] At present, we assume that an utterance is associated with at most one pointing gesture, and that pointing pertains to objects. This assumption is supported by our user study (Section 4.1).

[2] Owing to timing limitations of the gesture recognition system (Section 4), we simulate its output.

[3] At present, we do not consider dialogue salience.



(a) Pointing (b) Occlusion

Figure 3: Spatial pointing and occlusion

where $\Pr(k|\mathcal{P}_s, \mathcal{C})$ and $\Pr(\mathcal{P}_t|k, \mathcal{C})$ are estimated as described in Section 3.1.1 and 3.1.2 respectively. This equation is smoothed as follows (and incorporated into Equation 3) to take into account objects that are (spatially or temporally) excluded from the pointing gesture.

$$\Pr'(k|\mathcal{P}, \mathcal{C}) = \frac{\Pr(k|\mathcal{P}, \mathcal{C}) + \frac{1}{N}}{1 + \sum_{j=1}^{N} \Pr(k_j|\mathcal{P}, \mathcal{C})} \quad (6)$$

3.1.1 Estimating $\Pr(k|\mathcal{P}_s, \mathcal{C})$

$\Pr(k|\mathcal{P}_s, \mathcal{C})$, the probability that the user intended object k when pointing to a location, is estimated using a conic Gaussian density function around *PLine*, the *Pointing Line* created by extending the pointing vector returned by the gesture identification system (Figure 3(a)).[4]

$$\Pr(k|\mathcal{P}_s, \mathcal{C}) = \frac{\alpha\theta_k}{\sqrt{2\pi}\sigma_{P_s}(pd)} e^{-\frac{d(k,PLine)^2}{2\sigma^2_{P_s}(pd)}} \quad (7)$$

where α is a normalizing constant; $\sigma_{P_s}(pd)$ is the standard deviation of the Gaussian cone as a function of $pd(k, PLine)$, the *projected distance* between the user's pointing hand and the projection of object k on *PLine*; $d(k, PLine)$ is the shortest distance between the center of object k and *PLine*; and θ_k is a factor that reduces the probability of object k if it is (partially) *occluded* (Figure 3(b)).

The ***projected distance*** pd takes into account the imprecision of pointing actions — a problem that is exacerbated by the uncertainty associated with sensing a pointing vector. A small angular error in the detected pointing vector yields a discrepancy in the distance between the pointing line and candidate objects. This discrepancy increases as $pd(k, PLine)$ increases. To compensate for this situation, we increase the variance of the Gaussian distribution linearly with the projected distance from the user's hand (we start with a small standard deviation of $\sigma_0 = 5$ mm at the user's fingers, attributed to sensor error). This allows farther objects with a relatively high displacement from the pointing vector to be encompassed in a pointing gesture (e.g., the larger mug in Figure 3(a)), while closer objects with the same displacement are excluded (e.g., the smaller mug). This yields the following equation for the variance.

$$\sigma^2_{P_s}(pd) = \sigma_0^2 + K \cdot pd(k, PLine)$$

where $K = 2.5$ mm is an empirically determined increase rate.

The ***occlusion factor*** θ_k reduces the probability of objects as they become more occluded. We approximate θ_k by considering the objects that are closer to the user than k, and estimating the extent to which these objects occlude k (Figure 3(b)). This estimate is a function of the position of these objects and their size — the larger an intervening object, the lower the probability that the user is pointing at k. These factors are taken into account as follows.

$$\Pr(j \text{ occl } k) = \frac{\gamma}{\sqrt{2\pi}\sigma_\theta(pd)} e^{-\frac{(d(j,OLine) - \frac{1}{2}\dim_{\min}(j))^2}{2\sigma^2_\theta(pd)}} \quad (8)$$

where γ is a normalizing constant; the numerator of the exponent represents the maximum distance from the edge of object j to the line between the user's hand and object k, denoted *Object Line* (*OLine*); and

$$\sigma^2_\theta(pd) = \frac{1}{2}\left(\sigma_0^2 + K \cdot pd(j, OLine)\right)$$

represents the variance of a cone from the user's hand to object k as a function of distance. In order to represent the idea that object j must be close to the Object Line to occlude object k, we use half the variance of that used for the "pointing cone", which yields a thinner "occlusion cone" (Figure 3(b)). θ_k is then estimated as 1 minus the

[4] Since this is a continuous density function, it does not directly yield a point probability. Hence, it is normalized on the basis of the largest possible returned value.

maximum occlusion caused by the objects that are closer to the user than k.

$$\theta_k = 1 - \max_{\forall j \ d(j,\text{hand}) < d(k,\text{hand})} \{\Pr(j \text{ occl } k)\} \quad (9)$$

3.1.2 Estimating $\Pr(\mathcal{P}_t|k,\mathcal{C})$

$\Pr(\mathcal{P}_t|k,\mathcal{C})$ is obtained as follows.

$$\Pr(\mathcal{P}_t|k,\mathcal{C}) = \sum_{i=1}^{n} \frac{\Pr(\mathcal{P}_t, k, W_i, \mathcal{C})}{\Pr(k,\mathcal{C})} \quad (10)$$

$$= \sum_{i=1}^{n} \frac{\Pr(k|\mathcal{P}_t, w_i, \mathcal{C}) \cdot \Pr(T(w_i)|\mathcal{P}_t, \mathcal{C}) \cdot \Pr(\mathcal{P}_t|\mathcal{C})}{\Pr(k|\mathcal{C})}$$

where n is the number of nouns in the user's utterance, and $W_i = \langle w_i, T(w_i)\rangle$ is a tuple comprising the ith noun and the mid point of the time when it was uttered.

We make the following assumptions.

- $\Pr(\mathcal{P}_t|\mathcal{C}) = 1$, as all the gesture hypotheses are returned at the same time;
- given $\mathcal{P}_t$, the timing of a word $T(w_i)$ is conditionally independent of $\mathcal{C}$; and
- given w_i, k is conditionally independent of the timing of the pointing gesture $\mathcal{P}_t$, i.e., $\Pr(k|\mathcal{P}_t, w_i, \mathcal{C}) = \Pr(k|w_i, \mathcal{C})$.

This probability is represented as

$$\Pr(k|w_i,\mathcal{C}) = \frac{\Pr(w_i|k) \cdot \Pr(k|\mathcal{C})}{\sum_{j=1}^{N} \{\Pr(w_i|k_j) \cdot \Pr(k_j|\mathcal{C})\}}$$

where N is the number of objects.

These assumptions yield

$$\Pr(\mathcal{P}_t|k,\mathcal{C}) = \sum_{i=1}^{n} \frac{\Pr(w_i|k) \cdot \Pr(T(w_i)|\mathcal{P}_t)}{\sum_{j=1}^{N} \{\Pr(w_i|k_j) \cdot \Pr(k_j|\mathcal{C})\}} \quad (11)$$

where $\Pr(T(w_i)|\mathcal{P}_t)$, the probability of the time of word w_i given the time of the pointing gesture, is obtained from the following Gaussian time distribution for pointing.

$$\Pr(T(w_i)|\mathcal{P}_t) = \frac{\beta}{\sqrt{2\pi}\sigma_{P_t}} e^{-\frac{(T(w_i)-PTime)^2}{2\sigma_{P_t}^2}} \quad (12)$$

where β is a normalizing constant, *PTime* is the time of the gesture, and σ_{P_t} is the standard deviation of the Gaussian density function, which is currently set to 650 msec (based on our corpus).

As in our previous work (Makalic et al., 2008), we estimate $\Pr(w_i|k)$ using the Leacock and Chodorow (1998) WordNet similarity metric. This metric also yields a match probability between most objects and generic words like "object, thing, here, there", enabling us to handle requests such as "Get that *thing* over *there*".

3.2 Calculating the probability of a referring expression

As mentioned in Section 2, the intrinsic features previously considered in *Scusi?* are lexical item, colour and size (Makalic et al., 2008). Pointing affects referring expressions in that people may point instead of generating complex descriptions, they may employ demonstrative determiners together with generic terms such as "thing" (especially when they are unfamiliar with the name of an object), and they may use demonstrative pronouns. The first two behaviours were exhibited in our user study (Section 4), but none of our trial participants used demonstrative pronouns.

To incorporate pointing into the calculation of $\Pr(u|k,\mathcal{C})$, we add determiners to *Scusi?*'s formalism for intrinsic features, which yields

$$\Pr(u|k,\mathcal{C}) = \Pr(u_{\text{lex}}, u_{\text{det}}, u_{\text{colr}}, u_{\text{size}}|k,\mathcal{C})$$

After adding weights for the intrinsic features (inspired by (Dale and Reiter, 1995)), and making some simplifying assumptions, we obtain

$$\Pr(u|k,\mathcal{C}) = \quad (13)$$
$$\Pr(u_{\text{lex}}|k,\mathcal{C})^{w_{lex}} \cdot \Pr(u_{\text{det}}|k,\mathcal{C})^{w_{det}} \cdot$$
$$\Pr(u_{\text{colr}}|k)^{w_{colr}} \cdot \Pr(u_{\text{size}}|u_{\text{lex}},k)^{w_{size}}$$

The estimation of $\Pr(u_{\text{lex}}|k,\mathcal{C})$, $\Pr(u_{\text{colr}}|k)$ and $\Pr(u_{\text{size}}|u_{\text{lex}},k)$ is described in (Makalic et al., 2008). Here we focus on $\Pr(u_{\text{det}}|k,\mathcal{C})$.

3.2.1 Estimating $\Pr(u_{\text{det}}|k,\mathcal{C})$

$\Pr(u_{\text{det}}|k,\mathcal{C})$ is estimated as follows.

$$\Pr(u_{\text{det}}|k,\mathcal{C}) = \frac{\Pr(k|u_{\text{det}},\mathcal{C}) \cdot \Pr(u_{\text{det}}|\mathcal{C})}{\Pr(k|\mathcal{C})} \quad (14)$$

$$= \frac{\Pr(k|u_{\text{det}},\mathcal{C})}{\Pr(k|\mathcal{C})} \left[\begin{array}{l} \Pr(u_{\text{det}}|\mathcal{P},\mathcal{C}) \cdot \Pr(\mathcal{P}|\mathcal{C}) + \\ \Pr(u_{\text{det}}|\neg\mathcal{P},\mathcal{C}) \cdot \Pr(\neg\mathcal{P}|\mathcal{C}) \end{array} \right]$$

where det = {def_article, indef_article, demonstr_this, demonstr_that}; $\Pr(\mathcal{P}|\mathcal{C})$ and $\Pr(\neg\mathcal{P}|\mathcal{C})$ are returned by the gesture system; $\Pr(u_{\text{det}}|\mathcal{P},\mathcal{C})$ and $\Pr(u_{\text{det}}|\neg\mathcal{P},\mathcal{C})$ are obtained from our corpus; and for now we assume that $\Pr(k|u_{\text{det}},\mathcal{C}) = \Pr(k|\mathcal{C})$.[5] This yields

$$\begin{aligned}\Pr(u_{\text{det}}|k,\mathcal{C}) = \Pr(u_{\text{det}}|\mathcal{P},\mathcal{C}) \cdot \Pr(\mathcal{P}|\mathcal{C}) + \\ \Pr(u_{\text{det}}|\neg\mathcal{P},\mathcal{C}) \cdot \Pr(\neg\mathcal{P}|\mathcal{C})\end{aligned} \quad (15)$$

4 Evaluation

To obtain a corpus, we conducted a user study whereby we set up a room with labeled objects (Figure 1), and asked trial participants to request 12 selected items from *DORIS* (the room included 33 items in total, including distractors, and one of the authors pretended to be *DORIS*). The objects were selected and laid out in the room to reflect a variety of conditions, e.g., common and rare objects (e.g., vacuum tube); unique, non-unique and similar objects (e.g., white cups); and objects placed near each other and far from each other.

We divided our corpus of requests into two parts: with and without pointing. *Scusi?*'s performance was tested on input obtained from the ASR and on textual input (perfect ASR). We considered two scenarios for each sub-corpus: Pointing, where our pointing mechanism was activated on the basis of a simulated pointing gesture,[6] and No-Pointing, where no pointing gesture was detected. This was done in order to test two hypotheses: (1) when people point, pointing information improves interpretation performance; and (2) when they do not point, even perfect pointing has little effect on interpretation performance.

Scusi? was set to generate at most 300 sub-interpretations in total (including texts, parse trees, UCGs and ICGs) for each spoken request, and at most 200 sub-interpretations for each textual request. On average, *Scusi?* takes 10 seconds to go from texts to ICGs. An interpretation was deemed successful if it correctly represented the speaker's intention, which was encoded in one or more *Gold ICG*s. These ICGs were manually constructed on the basis of the requested objects and the participants' utterances. Multiple Gold ICGs were allowed if there were several suitable actions and objects.

4.1 The Corpus

19 people participated in the trial, generating a total of 276 requests, of which 136 involved pointing gestures (3 participants were asked to repeat the experiment after it became clear that they were refraining from pointing, as they erroneously assumed they were not allowed to gesture). We filtered out 64 requests, which included concepts our system cannot yet handle, specifically "the end of the table", projective modifiers (e.g., "behind/left"), ordinals ("first/second"), references to groups of things (e.g., "six blue pens"), and zero- and one-anaphora. This yielded 212 requests, of which 105 involved pointing gestures.

In addition, the software we used has the following limitations: the gesture recognition system (Li and Jarvis, 2009) requires users to hold a gesture for 2 seconds, and the ASR system is speaker dependent and cannot recognize certain words (e.g., "mug", "bowl" and "pen"). To circumvent these problems, each pointing gesture was manually encoded into a time-stamped vector; and one of the authors read slightly sanitized versions of participants' utterances into the ASR: "can you", "please" and "DORIS" were omitted; long prepositional phrases were shortened (e.g., "the thing with wires *sticking out of it*"); and words that were problematic for the ASR were replaced (e.g., "pencil" was used instead of "pen").

There was some difference in the length of requests with and without pointing, but it wasn't as pronounced as reported in (Johnston et al., 2002): requests with/without pointing had 5.84/6.27 words on average. ASR performance was worse for the requests that had pointing, with the top ASR interpretation being correct for only 46% of these requests, compared to 57.5% for the requests without pointing. This difference may be attributed to the ASR having trouble with sentence constructs associated with pointing. Overall

[5] In the future, we will incorporate distance from the user to refine the probabilities of determiners.

[6] At present, we assume accurate pointing and gesture detection, and precise information regarding the position of objects. In the near future, we will study the sensitivity of our mechanism to pointing inaccuracies, and to errors in gesture detection and scene analysis.

	% Gold ICGs in top 1	% Gold ICGs in top 3	Avg adj rank (rank)	% Not found	Avg adj rank (rank) 20	% Not found 20
Sub-corpus without pointing						
Text, *Scusi?*-NoPointing	89.7	93.5	4.39 (0.78)	0.9	1.18 (0.13)	4.7
Text, *Scusi?*-Pointing	86.9	87.9	3.28 (1.89)	0.9	**0.39** (0.35)	4.7
ASR, *Scusi?*-NoPointing	81.3	85.0	4.67 (0.83)	7.5	1.24 (0.17)	12.1
ASR, *Scusi?*-Pointing	79.4	81.3	5.00 (2.62)	5.6	**0.46** (0.40)	12.1
Sub-corpus with pointing						
Text, *Scusi?*-NoPointing	84.8	89.5	3.54 (0.59)	4.8	1.48 (0.20)	9.5
Text, *Scusi?*-Pointing	82.9	86.7	4.19 (1.63)	1.9	**0.41** (0.29)	7.6
ASR, *Scusi?*-NoPointing	76.2	82.9	7.93 (0.95)	10.5	1.79 (0.27)	15.2
ASR, *Scusi?*-Pointing	73.3	81.0	8.65 (2.76)	8.6	**0.68** (0.40)	14.3

Table 1: *Scusi?*'s interpretation performance

the ASR returned the correct interpretation, at any rank, for 88% of the requests.

4.2 Results

Table 1 summarizes our results. Column 1 displays the test condition (sub-corpus with/without pointing, text/ASR, and with/without *Scusi?*'s pointing mechanism). Columns 2-3 show the percentage of utterances that had Gold ICGs whose probability was among the top 1 and top 3, e.g., in the sub-corpus with pointing, when *Scusi?*-Pointing was run on text, 82.9% of the utterances had Gold ICGs with the highest probability (top 1). The average *adjusted rank (AR)* and average *rank* of the Gold ICG appear in Column 4. The rank of an ICG I is its position in a list sorted in descending order of probability (starting from position 0), such that all equiprobable ICGs are deemed to have the same position. The adjusted rank of an ICG I is the mean of the positions of all ICGs that have the same probability as I. For example, if we have 4 equiprobable ICGs in positions 0-3, each has a rank of 0, but an adjusted rank of $\frac{r_{\text{best}}+r_{\text{worst}}}{2} = 1.5$. Column 5 shows the percentage of utterances that didn't yield a Gold ICG. Column 6 shows the average AR for interpretations with AR < 20 (and their average rank), and Column 7 shows the percentage of utterances that had AR ≥ 20 or were not found. We distinguish between Gold ICGs with ARs 0 to 19 and total Gold ICGs that were found, because a dialogue manager is likely to inspect the promising options, i.e., those with AR $< K$ (we assume $K = 20$). In addition, there is normally a trade-off between the number of Not Found Gold ICGs and average AR. ICGs that are not found by one approach but are found by another approach typically have a high (bad) rank when they are eventually found (Zukerman et al., 2008). Thus, an approach that fails to find such "difficult" ICGs usually yields a lower average AR than an approach that finds these ICGs. Capping the ARs of the found Gold ICGs at 20 clarifies the trade-off between average AR and Not Found.

Our results show that, as expected, the main role of pointing is in referent disambiguation. This is evident from the significant reduction in average AR-20 (Column 6) for the pointing and no-pointing sub-corpora, under the text/ASR input conditions. All the differences are statistically significant with $p < 0.01$.[7] Nonetheless, the improvements in average AR-20 obtained by artificially introduced pointing in the no-pointing sub-corpus are smaller for both text and ASR than the improvements obtained with actual pointing. We posit that this smaller impact is due to the fact that utterances without pointing are more descriptive than those with pointing, hence benefitting less from the disambiguating effect of pointing.

The Pointing condition has a seemingly adverse effect on the number of interpretations with top ranks (Columns 2-3). This is explained by the fact

[7]The differences were calculated using a paired t-test for all the Gold ICGs that were found in both configurations.

that all equiprobable interpretations have the same rank, which happens more often under the No-Pointing condition than under the Pointing condition (as pointing has a disambiguating effect).

Finally, under all conditions, the rank of the request at the 75%-ile is 0, which indicates creditable performance. The larger number of Not Found Gold ICGs for the ASR condition is expected, as the ASR failed to find 12% of the correct texts on average, performing worse for the pointing sub-corpus. The other Not Found Gold ICGs were mainly due to parsing preferences, and multiple parses for some utterances that had the word "thing" (which matched all objects).

5 Related Research

Gesture recognition systems endeavour to detect the gesture being made. Common approaches include Hidden Markov Models, e.g., (Nickel and Stiefelhagen, 2003), and Finite State Machines, e.g., (Li and Jarvis, 2009). Systems that focus on pointing also identify the target object, without recognizing the type of this object (Nickel and Stiefelhagen, 2003; Li and Jarvis, 2009).

Most of the research in gesture and speech integration focuses on pointing gestures, employing speech as the main input modality, and using semantic fusion to combine spoken input with gesture. Different approaches are used for gesture detection, e.g., vision (Stiefelhagen et al., 2004; Brooks and Breazeal, 2006) and sensor glove (Corradini et al., 2002); and for language interpretation, e.g., dedicated grammars (Stiefelhagen et al., 2004; Brooks and Breazeal, 2006) and keywords (Einstein and Christoudias, 2004). Fusion is variously implemented using heuristics based on temporal overlap (Bolt, 1980; Johnston et al., 2002), querying a gesture-sensing module when ambiguous referents are identified (Fransen et al., 2007), or unification to determine which elements can be merged (Corradini et al., 2002; Stiefelhagen et al., 2004). These are sometimes combined with search techniques coupled with penalties (Einstein and Christoudias, 2004; Brooks and Breazeal, 2006). With the exception of Bolt's system, these systems were tested on utterances that were quite short and constrained.

Our approach integrates spatial and temporal aspects of gesture into our probabilistic formalism (Zukerman et al., 2008), focusing on the effect of pointing on object salience. Other salience-based approaches are described in (Einstein and Christoudias, 2004; Huls et al., 1995). However, they are not directly comparable with our approach, as they use salience to weigh the importance of factors pertaining to gesture-speech alignment, but there is no uncertainty associated with the visual salience resulting from pointing.

Our use of a probabilistic parser enables us to handle more complex utterances than those considered by most speech-gesture systems (Section 2). At the same time, we do not yet handle speech disfluencies, which are currently handled by (Einstein and Christoudias, 2004; Stiefelhagen et al., 2004). Also, at present we do not consider the challenges pertaining to the real-time synchronization of the output of a gesture-sensing and a speech-recognition system (Stiefelhagen et al., 2004; Brooks and Breazeal, 2006).

6 Conclusion and Future Work

We have extended *Scusi?*, our spoken language interpretation system, to incorporate pointing gestures. Specifically, we have offered a formalism that takes into account relationships between aspects of gesture and spoken language to integrate information about pointing gestures into the estimation of the probability of candidate interpretations of an utterance. Our empirical evaluation shows that our formalism significantly improves interpretation accuracy.

In the future, we propose to refine our model of demonstrative determiners. We also intend to perform sensitivity analysis regarding the accuracy of the vision system, and that of the gesture recognition system. In addition, we will conduct user studies to gain insights with respect to conditions that influence the probability of pointing, e.g., type of object and its position relative to the speaker.

Acknowledgments

This research was supported in part by ARC grant DP0878195. The authors thank R. Jarvis and D. Li for their help with the gesture system.

References

Bolt, R.A. 1980. "Put-that-there": voice and gesture at the graphics interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*, pages 262–270, Seattle, Washington.

Brooks, A.G. and C. Breazeal. 2006. Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, pages 297–304, Salt Lake City, Utah.

Corradini, A., R.M. Wesson, and P.R. Cohen. 2002. A Map-Based system using speech and 3D gestures for pervasive computing. In *ICMI'02 – Proceedings of the 4th International Conference on Multimodal Interfaces*, pages 191–196, Pittsburgh, Pennsylvania.

Dale, R. and E. Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 18(2):233–263.

Einstein, J. and C.M. Christoudias. 2004. A salience-based approach to gesture-speech alignment. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 25–32, Boston, Massachusetts.

Fransen, B., V. Morariu, E. Martinson, S. Blisard, M. Marge, S. Thomas, A. Schultz, and D. Perzanowski. 2007. Using vision, acoustics, and natural language for disambiguation. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*, pages 73–80, Washington, DC.

Gildea, D. and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Huls, C., W. Claassen, and E. Bos. 1995. Automatic referent resolution of deictic and anaphoric expressions. *Computational Linguistics*, 21(1):59–79.

Johnston, M., S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. 2002. MATCH: an architecture for multimodal dialogue systems. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 376–383, Philadelphia, Pennsylvania.

Leacock, C. and M. Chodorow. 1998. Combining local context and WordNet similarity forword sense identification. In Fellbaum, C., editor, *WordNet: An Electronic Lexical Database*, pages 265–285. MIT Press.

Li, Z. and R. Jarvis. 2009. Real time hand gesture recognition using a range camera. In *Proceedings of the Australasian Conference on Robotics and Automation*, Sydney, Australia.

Makalic, E., I. Zukerman, M. Niemann, and D. Schmidt. 2008. A probabilistic model for understanding composite spoken descriptions. In *PRICAI 2008 – Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence*, pages 750–759, Hanoi, Vietnam.

Nickel, K. and R. Stiefelhagen. 2003. Pointing gesture recognition based on 3D-tracking of face, hands and head orientation. In *ICMI'03 – Proceedings of the 5th International Conference on Multimodal Interfaces*, pages 140–146, Vancouver, British Columbia.

Sowa, J.F. 1984. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, Reading, MA.

Stiefelhagen, R., C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. 2004. Natural human-robot interaction using speech, head pose and gestures. In *IROS 2004 – Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pages 2422–2427, Sendai, Japan.

Zukerman, I., E. Makalic, M. Niemann, and S. George. 2008. A probabilistic approach to the interpretation of spoken utterances. In *PRICAI 2008 – Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence*, pages 581–592, Hanoi, Vietnam.