What's in a Gaze? The Role of Eye-Gaze in Reference Resolution in Multimodal Conversational Interfaces

Zahar Prasov and Joyce Y. Chai
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824, USA
{prasovza, jchai}@cse.msu.edu

ABSTRACT

Multimodal conversational interfaces allow users to carry a dialog with a graphical display using speech to accomplish a particular task. Motivated by previous psycholinguistic findings, we examine how eye-gaze contributes to reference resolution in such a setting. Specifically, we present an integrated probabilistic framework that combines speech and eye-gaze for reference resolution. We further examine the relationship between eye-gaze and increased domain modeling with corresponding linguistic processing. Our empirical results show that the incorporation of eye-gaze significantly improves reference resolution performance. This improvement is most dramatic when a simple domain model is used. Our results also show that minimal domain modeling combined with eye-gaze significantly outperforms complex domain modeling without eye-gaze, which indicates that eye-gaze can be used to potentially compensate a lack of domain modeling for reference resolution.

Keywords

Eye-gaze, Multimodal Conversational Interfaces, Reference Resolution

INTRODUCTION

Multimodal conversational interfaces allow users to carry a dialog with a graphical display using speech to perform a particular task or a group of related tasks. This method of interaction deviates from more traditional direct manipulation and WIMP (windows, icons, menus, and pointing devices) interfaces. Conversational interfaces are more flexible and natural to use. They allow for a better use of human communication skills, a larger diversity of users and tasks, and a faster task completion rate for visual-spatial tasks. Yet, conversational interfaces have the disadvantage that spoken language input can be noisy and ambiguous.

We believe that exploiting user eye-gaze can help alleviate the problem with language processing in conversational interfaces. There has been a significant body of work in the field of psycholinguistics studying the link between eye-gaze and speech [10] [11] [15] [25]. Eye-gaze has been shown to be a window to the mind. That is, the eye fixates on symbols that are being cognitively processed. The direction of gaze carries information about the focus of a person's attention [15].

Eye-tracking technology is being incorporated into speech-driven computer interfaces more and more frequently [16] [23] [24]. However, it remains unclear exactly to what degree eye-gaze is capable of helping automated language processing. In this work, we investigate what information can be obtained from eye-gaze as it relates to reference resolution in multimodal conversational interfaces. Reference resolution is the process of identifying application specific entities that are referred to by linguistic expressions. For example, identifying a picture on the interface that is referenced by the expression "the painting of a waterfall".

When a referring expression is uttered by a user, according to Conversation Implicature [8], users will not intentionally make ambiguous expressions to confuse the system. The reason that the system fails to correctly resolve some references is because the system lacks the adequate domain models or linguistic processing capability that a human may have. More referring expressions can be resolved with more sophisticated domain modeling and corresponding linguistic processing. However, sufficiently representing domain knowledge and processing human language has been a very difficult problem. Moreover, some domain information is very difficult to encode during design time. This is especially true for dynamic information that can change as the user manipulates the interface. For example, an object—presented via the user interface—that is perceived to be upside-

IUI'08, January 13-16, 2008, Maspalomas, Canary Islands, Spain

down can be rotated by the user to no longer have this property. The rotation operation is likely to change the language used to describe this object. While encoding the orientation information of each object is relatively simple, encoding this potential change in language is difficult. This domain information is often necessary to disambiguate spoken references. Thus our investigation considers whether the use of eye-gaze can compensate for some of these limitations. We hypothesize that the use of eye-gaze has the capacity to eliminate the need to encode some complex domain information without sacrificing reference resolution performance.

To investigate the role of eye-gaze in reference resolution in multimodal conversational interfaces we proceed with two primary research objectives. First, we construct a general probabilistic framework that combines speech and eye-gaze information for robust reference resolution. This framework is capable of independently encoding linguistic, dialog, domain, and eye-gaze information and combining these various information sources to resolve references to objects on the interface. Second, we attempt to validate our hypothesis that utilizing eye-gaze in reference resolution can compensate for a lack of sophisticated domain modeling. Specifically, we compare the amount of information provided by utilizing eye-gaze relative to that provided by increasing levels of encoded domain complexity (along with linguistic processing capable of handling this new domain information). To pursue these objectives, we use transcribed speech rather than automatically recognized speech data as a first step in our investigation. This will allow us focus on studying the relationship between eye-gaze and linguistic processing without additional confounding variables. The findings from this initial investigation will help us better understand the role of eye-gaze in spoken language understanding and establish an upper bound performance for our next step of processing recognized speech.

Our empirical results indicate that incorporating eye-gaze into our probabilistic framework improves reference resolution performance. This improvement is most dramatic when a simple domain model is used. Our results also show that minimal domain modeling combined with eye-gaze significantly outperforms complex domain modeling without eye-gaze.

In the following sections we discuss our investigation in more detail. We first describe some motivating work related to this research. Next, we described the domain we considered and the data we collected to conduct this investigation. We then describe our integrated reference resolution framework. We proceed to discuss the experiments we conducted using the integrated framework to process our collected data. Finally, we conclude with some implications of this investigation and provide some potential research directions for the future.

RELATED WORK

This investigation differs from previous work in several aspects. First, unlike previous investigations focusing on the role of eye-gaze in language production [9] [20], our work is conducted in a conversational setting that involves interaction between a user and a machine. Second, previous studies examine the use of eye-gaze as an active mode of input that controls the navigation of the interface [12] [16] [24] [23]. The work reported here addresses a different scenario, where speech is the main mode of interaction, while eye-gaze is a naturally occurring byproduct. Third, rather than simply using eye-gaze to construct an attentional salience model of various objects appearing on the visual interface [21] [22], we combine speech and eye-gaze into a single framework used for interpreting object references. Fourth, most multimodal reference resolution algorithms focus on combining speech and deictic gestures [4] [5] [13] [14] as opposed to speech and eye-gaze.

Role of Eye-Gaze in Language Production

The direction of eye-gaze has been shown to carry information about the focus of a person's attention [15]. Even though there may be some dissociation between *covert* visual attention and *overt* visual attention, which is expressed in the form of eye movements, behavioral neurophysiological studies have shown a tight coupling between the two [7] [11]. Consequently, eye-gaze has been continually used as a window to the mind in psycholinguistic language production tasks. Meyer et. al. [20] studied eye movements in an object naming task. It was shown that people consistently fixated objects prior to naming them. Griffin [10] showed that when multiple objects were being named in a single utterance, speech about one object was being produced while the next object was fixated and lexically processed.

Our work differs from these studies in several ways. First, our goal is to improve automated interpretation of utterances rather than to study the reasons that such utterances are made. Second, our domain contains a richer visual environment. Third, utterances produced in our setting are conversational in nature.

Role of Eye-Gaze in Human-Machine Interaction

The chief use of eye-gaze in human-machine interaction has been as a pointing mechanism in direct manipulation interfaces [12] or as an auxiliary modality to speech during multimodal communication [3] [16] [24]. These studies have shown that people do not like manipulating a user interface with their eye-gaze. Instead, in our investigation, we use eye-gaze as a naturally occurring byproduct of spoken interaction with a visual interface.

Our work is most similar to Campana, et. al. [3]. This work proposes integrating eye tracking information into the reference resolution component of a simulated robot intended for deployment on the Space Shuttle and/or International Space Station. Unfortunately, several issues such as how eye movement is used to determine fixated objects and how eye movement is in-

tegrated with speech are not addressed. Additionally, eye movement information is used only when speech-only reference resolution is unable to uniquely identify the referenced object(s). Instead, we focus on incorporating eye movement into reference resolution without system knowledge of whether speech alone is sufficient to resolve a particular referring expression.

Attentional Salience in Conversational Interfaces

Attentional salience modeling has been an important aspect of spoken language understanding. We have previously conducted studies using eye-gaze for attentional salience model building [21] [22]. In [22] eye-gaze has been used to indicate the salience of objects and this salience has shown to be useful for improving automated speech recognition. Our current investigation combines eye-gaze based salience models with speech to improve reference resolution in multimodal conversational interfaces.

Speech & Gesture Reference Resolution

Previous work on multimodal reference resolution has focused on fusing speech and deictic gesture information to identify intended referents [4] [5] [13] [14]. Like deictic gestures, eye-gaze provide information about a user's focus of attention. The difference is that eye fixations are produced subconsciously, there are far more eye fixations than gestures during interaction, and eye-gaze data is much more noisy than deictic gesture data. This makes eye-gaze more challenging to process than deictic gestures. Additionally, different cognitive factors govern the relationship between speech and gesture vs. speech and eye-gaze. Thus, speech & gesture reference resolution should not be reused directly and new algorithms need to be constructed for fusing speech and eye-gaze.

DATA COLLECTION

We use a simplified multimodal conversational interface to collect synchronized speech and eye-gaze data. Users are asked to communicate with a graphical interface depicting a static bedroom scene, as shown in Figure 1. A total of 28 objects (e.g. a bed, two chairs, several lamps, etc.) are explicitly modeled in the scene. Communication with the interface involves engaging in each of the 14 tasks supported by this interface. These tasks are designed to elicit both open-ended utterances (e.g., the system asks users to describe the room) and more restricted utterances (e.g., the system asks the user whether he/she likes the bed) that are commonly supported in conversational systems. The dialog exhibits many features of conversation, but is not truly interactive. Each of the tasks are related to a cohesive theme—the layout of the bedroom scene. However, all of the dialog is initiated by the system. This is done in order to constrain the dialog to the cohesive theme, which allows us to collect relevant data.

The Eyelink II head-mounted eye tracker sampled at 250 Hz is used to track user eye movements. The eye

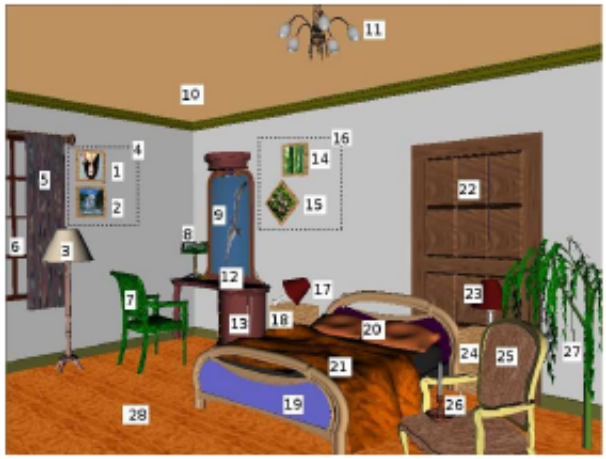

Figure 1. The bedroom scene for user studies. Here, each object is shown with a unique ID that is hidden from the users.

tracker automatically saves the screen coordinates and time-stamps of each gaze point. The collected raw gaze data is extremely noisy. The data is processed to eliminate irrelevant gaze points and smoothed to identify fixations. Irrelevant gaze points occur either when a user looks off-screen or when a user makes saccadic eye movements. Vision studies have shown that no cognitive processing occurs during saccadic eye movements; this is know as "saccadic suppression" [19]. The data is smoothed to compensate for the frequently jerky nature of eye movements. This is done by averaging the locations of five consecutive gaze points (occurring in the span of 20ms[1]). If none of these gaze points significantly deviate from one another, they are deemed to correspond to a fixation. Otherwise, they are considered to correspond to a saccade[2]. Given the domain model specifying the location of each object on the screen, these coordinates can be used to determine which objects are fixated at a particular time.

Note that it is not always possible to *uniquely* identify exactly which object is fixated at a given moment because several objects may overlap. This problem occurs because the Eyelink II guarantees accuracy to only 0.5° of visual angle (which is actually quite good for an eye tracker). Thus, when an eye fixation occurs on the border between two objects it is unclear which object is intended to be fixated. When viewing a 3D scene rather than a 2D projection (as we have described) another problem becomes evident. It stems from the fact that humans are capable of attending to objects at various depths along the same gaze direction. Eye trackers, however, provide only the gaze direction, but not gaze depth information. This is an important issue to keep in

[1]A saccade takes at least 20ms to occur.

[2]This threshold is measured in terms of visual angle. Typically eye tracking systems consider eye movements over 1° to be saccades. This threshold can be converted into a pixel value by knowing the screen resolution and the distance between the eye and the screen.

S_1	What is your favorite piece of furniture?
U_1	I would say my favorite piece of furniture in the room would be **the bed**
U_2	It looks like **the sheet** is made of leopard skin
U_3	And I like **the cabinets** around **the bed**
U_4	Yeah I think **that's** is my favorite
S_2	What is your least favorite piece of furniture?
U_5	**The brown chair** toward the front with **the candlestick** on **it**
U_6	**It** seems pretty boring

Figure 2. A conversational fragment demonstrating interaction with different types of referring expressions

mind when designing a multimodal interface that supports eye tracking.

Concurrently with eye tracking, speech data is recorded and time-stamped using the Audacity toolkit. Each referring expression in the speech utterance is subsequently manually annotated with the correct references to either a single object or a set of objects. The resulting data set is used to conduct our investigation. This data, collected from 6 users, consists of 319 utterances containing 371 referring expressions.

MULTIMODAL REFERENCE RESOLUTION

Reference resolution in a Multimodal Conversational Interface is the process of identifying the intended objects given a user's spoken utterance. Figure 2 shows a sample dialog in our conversational interface with the referring expressions to concrete objects marked in bold. Each S_i represents a system utterance and each U_i represents a user utterance. In this scenario, referring expressions tend to be definite noun phrases, such as "the bed", or pronouns, such as "it" or "that". Note that some pronouns may have no antecedent (e.g. "it" in utterance U_2) or an abstract antecedent [1], such as "democracy". Some expressions, (i.e. *evoking* references) initially refer to a new concept or object. Conversely, *anaphoric* references subsequently refer back to these objects. Anaphoric referring expressions are the most common type of *endophoric* referring expressions—referring to an entity in the dialog. Expressions that are not endophoric, must be *exophoric*—referring to an entity in the extralinguistic environment. Evoking references are typically exophoric.

A robust reference resolution algorithm should be capable of handling referring expressions that belong to a multitude of grammatical categories, as described above. We focus on resolving the most common referring expressions: definite noun phrases, which can be either exophoric or endophoric, and anaphoric pronouns whose antecedents are entities in the extralinguistic environment. These referring expressions compose a grave majority of the references found in our collected dialogs.

We have constructed an integrated probabilistic framework for fusing speech and eye-gaze for the purpose of conducting reference resolution. An *integrated* framework achieves two important objectives: (1) a single framework is capable of combining linguistic, dialog, domain, and eye-gaze information to resolve references and (2) a single framework is capable of resolving both exophoric and endophoric references. In fact, the system does not need to know whether an expression is exophoric or endophoric to reliably resolve it. This framework contains parameters that can be determined empirically or through machine learning approaches. The following section describes the framework.

Integrated Reference Resolution Framework

Given an utterance with n referring expressions r_1, r_2, ..., r_n, the reference resolution problem can be defined as the process of selecting a sequence of sets of objects $O_1, O_2, \ldots, O_n$ that best matches the referring expressions in the given utterance. Directly determining the best match is very difficult and potentially intractable.

To make this problem tractable we simplify it in two ways. First, we make the assumption that the occurrence of one referring expression is independent of another. Thus, the matching between O_j and r_j can be done individually for each expression r_j. Next, we construct each object set O_j using a two phase process. In the first phase, the likelihood $P(o_i|r_j)$ that r_j refers to object o_i is determined for each object in O, which is the set of all possible objects in the domain. In the second phase, O_j is constructed by combining the top k ranked o_i's, where k depends on whether referring expression r_j is singular or plural.

Previous studies have shown that user referring behavior during multimodal conversation does not occur randomly, but rather follows certain linguistic and cognitive principles [5] [17]. Different referring expressions signal the cognitive status of the intended objects. In our investigation, we model the user's cognitive status by considering two sources of potential referents: Visual History (VH) and Dialog History (DH). Each source reflects a different cognitive status, which is often associated with different types of referring expressions. Visual History represents the visible display information available on the time scale of the currently processed spoken utterance. It is usually associated with exophoric references. Dialog History represents the stack of previously resolved spoken utterances in the current conversation. It is often associated with anaphoric references.

We can use Equation 1 to determine the probability of a particular object being referenced given a particular referring expression. Here, we introduce a hidden variable T, which represents the cognitive status model of the intended object (either VH or DH):

$$P(o_i|r_j) = \sum_{T \in \{VH, DH\}} P(o_i|T, r_j)P(T|r_j) \quad (1)$$
$$= P(o_i|VH, r_j)P(VH|r_j) + P(o_i|DH, r_j)P(DH|r_j)$$

Equation 1 first evaluates the likelihood of each cognitive status associated with the given referring expression. Then, given the cognitive status and the referring expression, it determines the probability that a particular object is referenced. Linguistic and domain information can be extracted for each referring expression r_j while dialog and eye-gaze information is encoded into the DH and VH, respectively. Thus, Equation 1 allows us to achieve the objectives of integrating linguistic, dialog, domain, and eye-gaze information into a single probabilistic framework that is capable of resolving both exophoric and endophoric references. Next we show the specifics of how this information is integrated into the framework.

Likelihood of Cognitive Status

In Equation 1, $P(VH|r_j)$ and $P(DH|r_j)$ represent the likelihood of the cognitive status given a referring expression. These quantities are determined based on the grammatical category of the referring expression (pronoun, definite noun, etc.). For example, if r_j is a definite noun phrase, it is likely referring to an element from the visual history. This is because definite noun phrases are often made as evoking references, mentioning objects that do not appear in the dialog history. Those definite noun phrase references that are subsequent references to already mentioned objects are unlikely to have recently occurred in the dialog history because the more recent mentions tend to be pronouns. Thus, in this situation $P(VH|r_j)$ will be high and $P(DH|r_j)$ will be relatively low.

In the case expression r_j is a pronoun, we consider the VH and DH equally important in resolving pronominal referring expressions. This is because we expect the intended object to be recently present in the VH as well as the DH. Given that in our domain referent objects rarely go out of view, users are likely to gaze at these objects, thereby putting them in the recent visual history. Thus, we set

$$P(VH|r_j) = P(DH|r_j) = 0.5$$

If r_j is not a pronoun, it is typically a definite noun phrase given our domain. In this case the DH is unlikely to contain useful information. Thus we set

$$P(VH|r_j) = 1.0$$
$$P(DH|r_j) = 0.0$$

This means that these phrases take into account only the currently processed utterance.

Although we empirically set these parameters, it is possible to directly learn them from the data. It is important to note that small deviations in these probabilities have a minuscule effect on our reference resolution results.

Likelihood of Objects

Given the cognitive status, the likelihood that an object is referred to can be represented by the following equation:

$$P(o_i|T, r_j) = \frac{AS(o_i, T)^{\alpha(T)} \times Compat(o_i, r_j)^{1-\alpha(T)}}{\sum_i AS(o_i, T)^{\alpha(T)} \times Compat(o_i, r_j)^{1-\alpha(T)}} \quad (2)$$

where

- AS: Attentional salience score of a particular object. This score estimates the salience of object o_i given a cognitive status. For example, if the modeled cognitive status is VH, the salience of the object can be determined based on eye fixations that occur at the time the user utters referring expression r_j along with a combination of the object's visual properties. If the modeled cognitive status is DH, the score is based on how recently in the dialog o_i has been mentioned.
- $Compat$: Semantic compatibility score. This score represents to what degree that a particular object is semantically compatible with the referring expression. Factors such as the object's semantic type, color, orientation, and spatial location should be considered. Here, linguistic and domain information are brought together. For example, every color that is contained in the domain model must be contained in the system's lexicon and must be identified as a color when spoken by a user.
- $\alpha(T)$: Importance weight of Attentional Salience relative to Compatibility. This weight depends on whether the cognitive status T is modeled as the visual history or dialog history.

The details of these functions are explained next.

Attentional Salience with Cognitive Status Modeled as VH

When the cognitive status is modeled as $T = VH$, $AS(o_i, VH)$ represents an object's visual salience. There are many different ways to model the attentional salience of objects on the graphic display. For example, visual interface features, such as object centrality and size, can be extracted using image processing [18]. In continuously changing domains temporal features such as prominence and recency of an object appearing on the visual interface [2] can be used.

Instead of using visual features, we compute the attentional salience score with the use of eye-gaze. One method of doing this is by using our previously developed attention prediction algorithm [21]. In this work

the attentional salience score of an object is represented by the Relative Fixation Intensity of this object. That is,

$$AS(o_i, VH) = RFI(o_i) \in [0..1]$$

We have previously shown that the attentional salience of an object can be reliably modeled by the Relative Fixation Intensity (RFI) feature obtained from collected eye-gaze data. Fixation Intensity represents the amount of time that an object is fixated in a given a time window W. The RFI of an object is defined as the ratio between the amount of time spent fixating a candidate object during W and the amount of time spent fixating the maximally long fixated object during W. The intuition here is that objects that are fixated for a long period of time are considered to be more salient than those fixated for a short period of time. However, long period of time should be defined relative to other fixated objects during the same time period. A user may look at an object for a short period of time relative to W, but if this is the only object fixated during this time, it is likely to be salient. In our previous study [21], a good time range for W was empirically found be [-1500..0] ms relative to the onset of a spoken reference. This range conforms to prior psycholinguistic studies that show that on average eye fixations occur 900 ms (with sizable variance) prior to the onset of a referring expreession [9] [10].

We chose to use the Attentional Salience model as defined here for three reasons. First, we believe that models based on eye-tracking data are more reliable than models based on image processing. Second, the model based on the RFI feature is the simplest eye-gaze based model. Although, auxiliary gaze-based features (e.g. fixation frequency) and visual interface-based features (e.g. object occlusion and object size) can potentially improve the AS model, [21] shows that consideration of these features provide only a marginal improvement in the reliability of the attention prediction model given the domain described in the Data Collection section. Finally, the RFI-based model can generalize to any visual domain and makes for a good baseline.

Attentional Salience with Cognitive Status Modeled as DH

When cognitive status is modeled as $T = DH$, an object's salience within the dialog is represented by $AS(o_i, DH)$. In our investigation, this score indicates how recently (on a dialog history time scale) a particular object has been referred to. An object's salience within a dialog decays exponentially over time as shown in Equation 3

$$AS(o_i, DH) = \frac{(1/2)^{n_i}}{Z} \in [0..1] \quad (3)$$

where, n_i is the number of referring expressions that have been uttered since the last time object o_i has been referred to and Z is the normalization factor $\sum_i (1/2)^{n_i}$.

Referring Expression		Object				
		o_{18}	o_{19}	o_{21}	o_{24}	$\cdots$
U_1	the bed	0	1	0	0	0
U_2	the sheet	0	1/2	1	0	0
U_3	the cabinets	1	1/4	1/2	1	0
U_3	the bed	1/2	1	1/4	1/2	0
U_4	that's	1/4	1	1/8	1/4	0

Table 1. Example of dialog history table

The dialog history information is represented as a table of such salience scores. Initially each object's score is 0. When an expression referring to an object is uttered, its score is set to 1. At the same time, the scores of each other object are divided by 2. For example, the first four utterances of the dialog in Figure 2 would result in the scores shown in Table 1. Here, the first two columns represent a time step during which a referring expression is uttered. The remaining columns represent all of the objects that could possibly be mentioned. The object IDs correspond to those shown in Figure 1. Each cell represents the corresponding object's salience score within the dialog. In this example, after U_2 is uttered and its only referring expression is resolved to object o_{21}, this object's score is set to 1. The salience scores of all other objects are divided by 2. Thus, when U_3 is uttered,

$$AS(o_{21}, DH) = \frac{1}{1 + 1/2} = 2/3$$

making o_{21} the most salient object. Utterance U_3 contains two referring expression which are processed in the order of being spoken. Given that the referring expression "the cabinets" is correctly resolved to the two cabinets o_{18} and o_{24}, the salience scores of each of these two objects are set to 1 and once again all other scores are discounted.

Semantic Compatibility

The semantic compatibility score $Compat(o_i, r_j)$ represents to what degree object o_i is semantically compatible with referring expression r_j. The compatibility score is defined in a similar manner as in our previous work [5] [6]:

$$Compat(o_i, r_j) = Sem(o_i, r_j) \times \prod_k Attr_k(o_i, r_r) \quad (4)$$

In this equation:

- $Sem(o_i, r_j)$ captures the coarse semantic type compatibility between o_i and r_j. It indicates that the semantic type of a potential referent should correspond to the semantic type of the expression used to refer to it. Consequently, in our investigation, $Sem(o_i, r_j) = 0$ if the semantic types of o_i and r_j are different and $Sem(o_i, r_j) = 1$ if they are the same or either one is unknown.

- $Attr_k(o_i, r_r)$ captures the object-specific attribute compatibility (indicated by the subscript k) between o_i and r_j. It indicates that the expected features of a potential referent should correspond to the features associated with the expression used to refer to it. For example, in the referring expression "the brown chair", the color feature is *brown* and therefore, an object can only be a possible referent if the color of that object is *brown*. Thus, we define $Attr_k(o_i, r_r) = 0$ if both o_i and r_j have the feature k and the values of this feature are not equal. Otherwise, $Attr_k(o_i, r_r) = 1$.

Attentional Salience vs. Semantic Compatibility

In Equation 2, $\alpha(T)$ represents the importance tradeoff between attentional salience and semantic compatibility. A high value of $\alpha(T)$ indicates that the AS score is more important for reference resolution than the semantic compatibility score. A low value indicates that the opposite is true. For different kinds of interfaces $\alpha(T)$ may change. For example, if an interface is composed of many very small objects or many overlapping objects, eye-gaze becomes less reliable. In this case, $\alpha(T)$ should be fairly low. In our investigation, we wanted both quantities to have equal importance. However, semantic compatibility should win out in case of a tie. A tie can occur if all elements in T are semantically incompatible with referring expression r_j and no compatible elements are salient. Thus, we set $\alpha(T) = 0.49$ for both VH and DH.

Selection of Referents

Once the likelihoods for each candidate object are computed according to Equation 1 we can proceed to selecting the most probable referenced-object set O_j given a referring expression r_j. First, objects are ranked according to the probabilistic framework. Next, the number of objects being referenced is considered. Each reference can be singular or plural. If it is plural, the exact number of objects being referenced may be known (e.g. "these two pictures") or unknown (e.g. "these pictures"). If r_j is singular, the highest ranked object is selected. If more than one object has the highest score, there is an ambiguity and reference resolution fails. If r_j is plural and the number of referenced objects is known to be k, then the top k objects are selected. If r_j is plural, but the number of referenced objects is unknown, then all of the objects that have the top score are selected.

EXPERIMENTAL RESULTS

As we have already mentioned, a major goal of this work is to investigate whether or not the use of eye-tracking data can compensate for an incomplete domain model without sacrificing reference resolution performance. We categorize four different levels of domain models. These models vary in complexity such that a complex model completely entails a simpler model as shown in Figure 3. In addition to these four levels, we consider an *empty* domain model that contains no domain information or linguistic processing.

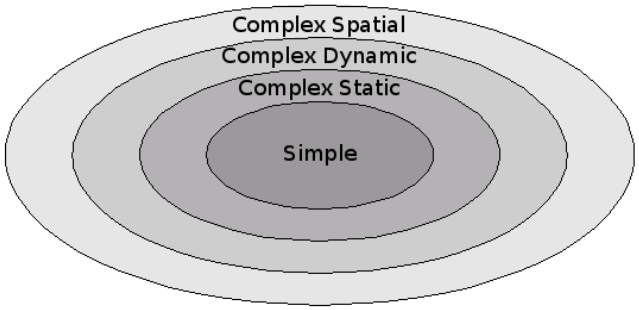

Figure 3. Domain Hierarchy

- Empty (⊘): This model makes no domain or linguistic information available. When it is employed, referents are determined purely based on eye-gaze.
- Simple (*Simp*): Domain and linguistic information is limited to a coarsely defined semantic type of each object (e.g. *lamp*, *chair*, *painting*, etc.).
- Complex Static (*CStat*): Various object attributes are encoded. These attributes satisfy the requirement of being inherent to a particular object. More precisely, these attributes cannot be changed through interaction with the interface (e.g. *floor* lamp, painting of a *waterfall*).
- Complex Dynamic (*CDyn*): Object attributes that can potentially change are encoded (e.g. *green* chair, *crooked* lamp). For example, a user may say "paint the chair red" when talking about the green chair. After this operation, the chair is no longer green.
- Complex Spatial (*CSpat*): Spatial relationships between objects are encoded (e.g. the chair *next to* the bed, the one *above* the waterfall). Given that a user can move around in the interface, spatial relationships are clearly dynamic attributes.

The domain model variation is incorporated into our probabilistic framework via the semantic compatibility function shown in Equation 4. The Simple domain encodes information necessary to make coarse semantic type compatibility judgments $Sem(o_i, r_j)$ while each of the more complex domains add more attributes to be considered by $Attr_k(o_i, r_r)$. It is important to note that an increasing level of linguistic processing capacity is associated with an increasing level of domain model complexity. For example, if the domain model contains the notion of a painting of a waterfall (o_2 in Figure 1), the token *waterfall* must be present in the system's lexicon and the system must be able to determine that this token is an attribute of object o_2.

The *Simp* domain model has the advantage of being completely independent of the specific tasks that can be performed by the conversational interface. That is, this domain model can be constructed by simply knowing the semantic type of the objects present in the interface.

The more complex domain models attempt to encode information that uniquely distinguishes objects from one another. This require knowledge about the similarity and difference between these objects, the kinds of references people tend to make to various objects, as well as information about the visual layout of these objects. This information is highly dependent on the design of the interface as well as the task being performed.

Here, we aim to compare the reference resolution performance on the speech & eye-gaze algorithm to a speech-only algorithm. In this comparison the inclusion or exclusion of eye-gaze is the only variable. As we have shown, eye-gaze can be encoded into our reference resolution framework as $AS(o_i, VH)$. Eye-gaze information can be removed by changing the $AS(o_i, VH)$ score to be uniformly distributed over all objects o_j. Next, we proceed to analyze the impact of eye-gaze on reference resolution, we use the 371 collected multimodal inputs described in the Data Collection section. Note that in our current work, most of the parameters were empirically determined. The results presented here should not be viewed as a formal evaluation of the approach, but rather an analysis of the role of eye-gaze in compensating for insufficient domain modeling. We are in the process of collecting more data and will do a formal evaluation in the near future.

Figure 4 shows the performance comparison of the reference resolution algorithm that incorporates eye-gaze data vs. the speech-only algorithm. The comparison is performed using each of the aforementioned domain complexities in terms of reference resolution accuracy. This chart shows that the speech & eye-gaze algorithm performs significantly better than the speech-only algorithm for each of the domain variations. The best performance of approximately 73% is reached when eye-gaze is combined with the most complex level of domain (*CSpat*).

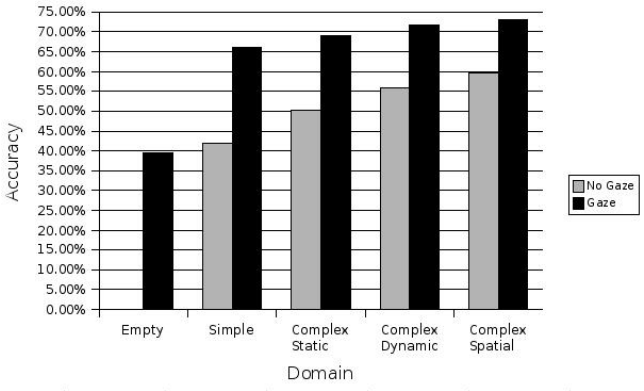

Figure 4. Overall Accuracy

Table 2 shows a clearer picture of the effect of eye-gaze in multimodal reference resolution for each level of domain complexity. References are divided into pronouns and definite noun phrases. In our data set, all referring expressions to concrete objects are comprised of these two grammatical categories. Here, we investigate whether the effect of eye-gaze varies depending on the grammatical category of the referring expression. Of the 371 total inputs (Table 2a), 84 (22.64%) are pronominal references (Table 2b) and 287 (77.36%) are definite nominal references (Table 2c). Each category of expression exhibits a significant improvement in reference resolution accuracy when eye-gaze is used. Additionally, they both exhibit diminishing returns with increasing domain complexity. That is, the improvement gained by using eye-gaze is larger when the domain is simpler. This effect is slightly more evident in the case of pronominal noun phrases.

	Domain Type	Without Gaze	With Gaze	Improvement
(a) Total	⊘	0	147	—
	Simp	156	245	57.05%
	CStat	186	256	37.63%
	CDym	207	266	28.50%
	CSpat	221	271	22.62%
(b) Pronoun	⊘	0	30	—
	Simp	21	43	104.76%
	CStat	26	44	69.23%
	CDym	33	47	42.42%
	CSpat	33	47	42.42%
(c) Definite	⊘	0	117	—
	Simp	135	202	49.63%
	CStat	160	212	32.50%
	CDym	174	219	26.86%
	CSpat	188	224	19.15%

Table 2. Improvement

Interestingly, the addition of eye-gaze information to the simple domain has a significantly larger affect on reference resolution performance than the addition of domain information. In Table 2a, compare the top row (*Simp*) with gaze to the bottom row (CSpat) without gaze. The former outperforms the latter 245 to 221, for a 57.05% vs. 41.61% increase over the baseline case (*Simp* without gaze). This result implies that gaze-data provides more useful information toward the reference resolution task than a fairly complex domain model.

Our next objective is to determine the coverage of eye-gaze relative to domain complexity as it pertains to the reference resolution task. Here, *coverage* refers to the percentage of references that require a complex domain definition to be resolved by the speech-only algorithm that can instead be resolved by resorting to eye-gaze with the *Simp* domain model. Table 3 displays this comparison. Here references are subdivided by according to the minimal domain information necessary to resolve the particular referent. For example, the second row of Table 3a shows the number of references that cannot be resolved by the *Simp* domain, but can be resolved by adding static attribute information to the domain model.

	Domain Type	Resolved References	Coverage by *Simp* Domain with Eye-Gaze	
(a) Total	*Simp*	156	154	98.72%
	CStat–Simp	30	19	63.33%
	CDyn–CStat	21	9	42.86%
	CSpat–CDyn	14	6	42.86%
(b) Pronoun	*Simp*	21	20	95.24%
	CStat–Simp	5	5	100.00%
	CDyn–CStat	7	2	28.57%
	CSpat–CDyn	0	0	—
(c) Definite	*Simp*	135	134	99.26%
	CStat–Simp	25	14	56.00%
	CDyn–CStat	14	7	50.00%
	CSpat–CDyn	14	6	42.86%

Table 3. Coverage

Several implications can be made from these results. The first is that eye-gaze can be used to resolve a significant number of references that require a complex domain when eye-gaze is not used. Eye-gaze is most beneficial when little domain information is available. Additionally, eye-gaze does not completely eliminate the need for domain modeling. As shown in table 2, eye-gaze alone can be used to resolve only about 40% (147 out of 371) references. Rather some of the errors that are caused by having an insufficient domain model are alleviated when eye-gaze is used. The second thing to note is that the *Simp* domain with eye-gaze information can resolve almost all (98.72%) of the references that can be resolved with the same domain, but without eye-gaze information. While this result is not very surprising, one may have expected that the coverage would be 100% in this case, but it is apparent that eye-gaze can introduce a small amount of noise into the reference resolution process. The Attention Prediction model is not perfect and eye-gaze is well known to be noisy.

DISCUSSION

Given that the best performing reference resolution algorithm achieves about 73% accuracy, it is important to consider the sources that cause the errors in the remaining 27% of referring expressions. Typically, errors occur when neither domain information nor gaze is capable of the resolving a referring expression alone. Several factors can cause errors in each of the modalities.

First, the current framework is only capable of handling anaphoric pronouns (those that refer back to an entity in the dialog), but not cataphoric pronouns (those that refer to an entity in the forthcoming speech). Additionally, the language processing component has some limitations. For example, the phrase "the chair to the left of the bed" can be interpreted, but "the bed with a chair to the left" cannot. This example demonstrates the difficulty of linguistic processing and provides more motivation for using eye-gaze to compensate for such deficiencies.

Second, as has been noted, eye-gaze data is very noisy and eye-gaze is far from a perfect predictor of user focus of attention. Some of this noise comes from inaccurate eye tracker readings. Additionally, errors in visual attention prediction can arise from an insufficiently accurate temporal mapping between speech and eye-gaze. For example, a user may look at an object and then wait a significant amount of time before referring to it. Thus, this object's salience will be far lower than intended by the user. Alternatively, a user's overt attention (eye-gaze) may be disassociated from the user's covert attention. This may happen because the user has become very familiar with the visual interface and no longer needs to fixate an object to be able to refer to it. Further investigation is necessary to pinpoint exactly what causes these eye-gaze errors and how much reference resolution suffers because of them.

Our current investigation has focused on processing transcribed linguistic expressions. When real-time speech is used, an important issue that needs further investigation is the effect of eye-gaze when automated speech recognition is considered. Speech recognition errors cause even more difficulty for linguistically processing domain information. For example, even if there exists a sophisticated spatial model of the objects appear on the interface, a failure in linguistic processing can cause these domain models to become useless. Eye-gaze is likely to compensate for some of these failures in the same way as it compensates for insufficient domain modeling. We will investigate this in our future work.

CONCLUSION

In this paper, we have presented an integrated approach that incorporates eye-gaze for reference resolution. The empirical results have shown that the use of eye-gaze improves interpretation performance and can compensate for some problems caused by the lack of domain modeling. These results have further implications for constructing conversational interfaces. The improvements in interpretation of object references gained via incorporating eye-gaze will allow systems to make fewer unexpected and erroneous responses. Additionally, reducing the reliance on complex domain modeling will make multimodal conversational interfaces easier to design. Eye tracking technology has improved significantly in the past decade. Integrating non-intrusive (e.g., Tobii system) and high performance eye trackers with conversational interfaces is becoming more feasible. The results from this work provide a step towards building the next generation of intelligent conversational interfaces.

ACKNOWLEDGMENTS

This work was supported by IIS-0535112 and IIS-0347548 from the National Science Foundation. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

1. D. K. Byron. Resolving pronominal reference to abstract entities. In *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*, 2002.

2. D. K. Byron, T. Mampilly, and T. Sharma, V.and Xu. Utilizing visual attention for cross-modal coreference interpretation. In *Spring Lecture Notes in Computer Science: Proceedings of CONTEXT-05*, pages 83–96, 2005.

3. E. Campana, J. Baldridge, J. Dowding, B. A. Hockey, R. W. Remington, and L. S. Stone. Using eye movements to determine referents in a spoken dialogue system. In *Proceedings of Perceptive User Interfaces*, 2001.

4. J. Chai, P. Hong, M. Zhou, and Z Prasov. Optimization in multimodal interpretation. In *Proceedings of 42nd Annual Meeting of Association for Computational Linguistics (ACL)*, pages 1–8, 2004.

5. J. Chai, Z. Prasov, and S. Qu. Cognitive principles in robust multimodal interpretation. *Journal of Artificial Intelligence Research*, 27:55–83, 2006.

6. J. Y. Chai, Z. Prasov, J. Blaim, and R. Jin. Linguistic theories in efficient multimodal reference resolution: An empirical investigation. In *ACM International Conference of Intelligent User Interfaces (IUI05)*. ACM Press, 2005.

7. J. M. Findlay. Eye scanning and visual search. In J. M. Henderson and F. Ferreira, editors, *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press, 2004.

8. H. P. Grice. Speech acts. In *Logic and conversation*, pages 41–58. New York: Academic Press., 1975.

9. Z. M. Griffin. Gaze durations during speech reflect word selection and phonological encoding. In *Cognition*, volume 82, pages B1–B14, 2001.

10. Z. M. Griffin and K. Bock. What the eyes say about speaking. In *Psychological Science*, volume 11, pages 274–279, 2000.

11. J.M. Henderson and F. Ferreira. In *The interface of language, vision, and action: Eye movements and the visual world*. Taylor & Francis, 2004.

12. R. J. K. Jacob. Eye tracking in advanced interface design. *In W. Barfield and T. Furness, editors, Advanced Interface Design and Virtual Environments*, pages 258–288, 1995.

13. M. Johnston. Unification-based multimodal parsing. In *Proceedings of ACL/COLING'98*, 1998.

14. M. Johnston and S Bangalore. Finite-state multimodal parsing and understanding. In *In Proceedings of COLING00*, 2000.

15. M. A. Just and P. A. Carpenter. Eye fixations and cognitive processes. In *Cognitive Psychology*, volume 8, pages 441–480, 1976.

16. M. Kaur, M. Tremaine, N. Huang, J. Wilder, Z. Gacovski, F. Flippo, and C. S. Mantravadi. Where is "it"? event synchronization in gaze-speech input systems. In *Proceedings of Fifth International Conference on Multimodal Interfaces*, pages 151–157. ACM Press, 2003.

17. A. Kehler. Cognitive status and form of reference in multimodal human-computer interaction. In *In Proceedings of AAAI00*, 2000.

18. J. Kelleher and J. van Genabith. Visual salience and reference resolution in simulated 3-d environments. *Artificial Intelligence Review*, 21(3), 2004.

19. E. Matin. Saccadic suppression: a review and an analysis. In *Psychological Bulletin*, volume 81, pages 899–917, 1974.

20. A. S. Meyer and W. J. M. Levelt. Viewing and naming objects: Eye movements during noun phrase production. In *Cognition*, volume 66, pages B25–B33, 1998.

21. Z. Prasov, J. Chai, and H. Jeong. Eye gaze for attention prediction in multimodal human-machine conversation. Technical report, In Proceedings of the AAAI Spring Symposium on Interaction Challenges for Intelligent Assistants, 2007.

22. S. Qu and J. Chai. An exploration of eye gaze in spoken language processing for multimodal conversational interfaces. In *In Proceedings of the Conference of the North America Chapter of the Association of Computational Linguistics (NAACL)*, 2007.

23. P. Qvarfordt, D. Beymer, and S. Zhai. Realtourist - a study of augmenting human-human and human-computer dialogue with eye-gaze overlay. In *INTERACT 2005, LNCS 3585*, pages 767–780, 2005.

24. P. Qvarfordt and S. Zhai. Conversing with the user based on eye-gaze patterns. In *Proc. Of the Conference on Human Factors in Computing Systems*. ACM, 2005.

25. M. K. Tanenhous, M. Spivey-Knowlton, E. Eberhard, and J. Sedivy. Integration of visual and linguistic information during spoken language comprehension. In *Science*, volume 268, pages 1632–1634, 1995.