Eye Gaze for Attention Prediction in Multimodal Human-Machine Conversation

Zahar Prasov and **Joyce Y. Chai** and **Hogyeong Jeong**
Department of Computer Science
Michigan State University
East Lansing, MI 48823, USA
prasovza@msu.edu and jchai@cse.msu.edu and jeonghog@msu.edu

Abstract

In a conversational system, determining a user's focus of attention is crucial to the success of the system. Motivated by previous psycholinguistic findings, we are currently examining how eye gaze contributes to automated identification of user attention during human-machine conversation. As part of this effort, we investigate the contributing roles of various features that are extracted from eye gaze and the visual interface. More precisely, we conduct a data-driven evaluation of these features and propose a novel evaluation metric for performing such an investigation. The empirical results indicate that gaze fixation intensity serves an integral role in attention prediction. Fixations to objects are fairly evenly distributed between the start of a reference and 1500 milliseconds prior. When combined with some visual features (e.g., the amount of visual occlusion of an object), fixation intensity can become even more reliable in predicting user attention. This paper describes this empirical investigation of features and discusses the further implication of attention prediction based on eye gaze for language understanding in multimodal conversational interfaces.

Introduction

Previous studies have shown that eye gaze is one of the reliable indicators of what a person is "thinking" about (Henderson & Ferreira 2004). The direction of gaze carries information about the focus of the user's attention (Just & Carpenter 1976). In human language processing tasks specifically, eye gaze is tightly linked to cognitive processing. The perceived visual context influences spoken word recognition and mediates syntactic processing (Tanenhous *et al.* 1995). Additionally, directly before speaking a word, the eyes move to the mentioned object (Griffin & Bock 2000). Not only is eye gaze highly reliable, it is also an implicit, subconscious reflex of speech. The user does not need to make a conscious decision; the eye automatically moves towards the relevant object, without the user even being aware.

Motivated by these psycholinguistic findings, we are currently investigating the role of eye gaze in human-machine conversation, in particular for spoken language understanding. As a first step in our investigation, we are currently examining how eye gaze contributes to automated identification of user attention during human-machine conversation. This investigation differs from previous work in two aspects. First, previous studies examine the use of eye gaze as an active mode of input that controls the navigation of the interface. This work has shown that gaze fixation intensity is an important feature for predicting user attention (Qvarfordt & Zhai 2005). The work reported here addresses a different scenario, where speech is the main mode of interaction, while eye gaze is a naturally occurring byproduct. Second, unlike previous investigation focusing on the role of eye gaze in language production (Meyer & Levelt 1998; Griffin 2001), our work is conducted in a conversational setting that involves interaction between a user and a machine. These unique settings, which have received less attention, apply to a range of realistic and important problems that involve speech communication between a user and a graphical display. This paper investigates the role of eye gaze in this important setting. In particular, we address the following questions:

- Question 1: How are pertinent eye gaze fixations temporally distributed relative to spoken utterances?
- Question 2: What effect do different representations of gaze fixation intensity have on performance of the user attention prediction task?
- Question 3: Can auxiliary visual features further enhance the reliability of the prediction based on gaze fixation intensity? If so, what effect do different representations of these auxiliary features have on performance of the user attention prediction task?
- Question 4: What are appropriate evaluation metrics for measuring different features' reliability for the attention prediction task?

To gain an overall understanding of this problem, we conducted a user study to collect eye gaze data from subjects interacting with a multimodal conversation system. We investigated the role of key gaze features (e.g., fixation intensity) and auxiliary visual features (e.g., object size and visual occlusion) in attention prediction based on a Bayesian Logistic Regression model. Our results indicate that fixation intensity is the foremost important indicator of user attention in this setting. The visual occlusion feature, which takes into consideration physically overlapping objects on the display, can be used to modify the fixation intensity measurement for more reliable attention prediction. Although intuitively the size of the object on the graphical display and the frequency of fixation for an object can affect the measurement of fixation intensity, these two features have shown no significant effect in attention prediction. We believe these findings will enable better recognition of user attention during human-machine conversation.

This will, in turn, have further implications for constructing artificial assistants designed to interact with humans. Improving an artificial assistant's understanding of which objects are the focus of conversation will allow these systems to make fewer unexpected and erroneous responses. This will speed up the dialog and improve task completion rate. Additionally, users will have more trust that the system is making correct interpretations of their speech and thus will be more apt to continue using such artificial assistants.

Related Work

Eye gaze has been extensively studied in psycholinguistic comprehension tasks. Some psycholinguistic studies have focused on the role of eye gaze in language production tasks. Meyer et. al. (1998) studied eye movements in an object naming task. It was shown that people fixated objects prior to naming them. Objects that are difficult to name were fixated for a longer period of time than those that are easy to name. The Griffin (2001) study 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.

In addition to psycholinguistic eye gaze studies, there have been several attempts to use eye gaze to facilitate interaction in human-machine communication. Most of these attempts focus on either using eye gaze to directly manipulate an interface via pointing or using eye gaze as a disambiguation tool for multimodal interfaces. For example Jacob (2000) explores the use of eye gaze as a substitute for pointing in a virtual environment. This study shows that interaction using eye gaze is faster than pointing, but causes a decline in memory recollection of spatial information. Kaur, et. al. (2003) explores the temporal alignment between eye gaze and speech during a simple on-screen object movement task, which combines gaze—as a pointing mechanism—with speech. These results have shown that the eye fixation that most likely identifies the object to be moved occur, on average, 630 milliseconds (ms) before the onset of the commanding utterance. In the iTourist project, Qvarfordt et. al. (2005) attempt to take a step toward using eye gaze as an integrated channel in a multimodal system. They attempt to determine object activation as a user views a map interface designed to facilitate a trip planning task. As people gaze at objects on the screen, an "arousal" score is calculated for each object. Once this score reaches a predefined threshold, the object becomes activated and the system provides information about this object to the user. In each of these scenarios eye gaze is knowingly used by a participant as an active mode of input.

A Data-driven Approach

To examine the questions raised earlier, we employed a data-driven approach that is based on the logistic regression model. In this section, we first describe the important features that are considered in this investigation and then give a brief introduction to the logistic regression model that employs these features to perform the attention prediction task.

Relevant Features

Fixation Intensity Gaze fixation has been shown to be closely tied to user attention. Fixation intensity can be crudely defined as the length of a fixation upon an object in a visual scene. Generally, long fixations signify that a user is paying attention to this object. This increase in likelihood of attention increases the likelihood that this object will be referenced in the user's speech. Here, we describe four different representations of the fixation intensity measure. The reason to consider these variations is to evaluate their potentially different impact and reliability for prediction of user attention. We identified the following variations:

- Absolute Fixation Intensity (AFI): AFI is the amount of time spent fixating on an object during a particular time window W. The time window considered in this study ranges from onset of a spoken reference to 1500 ms prior to the onset. Objects that are fixated for a long period of time are considered to be more likely to be activated than those fixated for a short period of time.
- Relative Fixation Intensity (RFI): Given a time window W, RFI is the ratio between the AFI of a candidate object and the maximal AFI in W. An object may have a low AFI, but still have a high RFI if the maximal AFI in W is relatively low.
- Weighted Absolute Fixation Intensity (WAFI): Previous language production studies have shown that the mean object fixation occurs between 630 (Kaur *et al.* 2003) and 932 (Griffin & Bock 2000) ms before it is referenced, depending on the task and domain. Fixations made to objects occurring near the mean are more indicative of user attention than other fixations. The WAFI measure takes this factor into consideration by weighting the fixation durations based on a skew-normal distribution.
- Weighted Relative Fixation Intensity (WRFI): WRFI is the ratio between WAFI of an object and the maximal WAFI in a particular W.

While fixation intensity is likely to be the most important factor, we hypothesize that other auxiliary features can also contribute to the eye gaze behavior and thus the prediction of attention. We discuss those features next.

Object Size When users interact with a graphic display, the size of the object on the display can potentially affect user eye gaze behavior, and thus affect the prediction of attention. For example, it is difficult to fixate on small objects for long periods of time. People instinctively make small jerky eye movements. Large objects are unaffected by these movements because these movements are unlikely to escape the object boundary. Thus our hypothesis is that small objects with a lower fixation intensity are still likely to be the attended objects (i.e., be activated by the eye gaze). To take this effect into consideration, we use the object size to represent the area of a candidate object relative to a baseline object (e.g., the largest object in the scene). The values for this feature are computed by finding the area of the smallest rectangle that fully encompasses a particular object. The object size feature is normalized by taking a ratio of a candidate object to a baseline object (the largest object in the visual scene).

Visual Occlusion In a graphical display, it is likely that objects overlap with one another. The visual occlusion of an object represents how much of this object is obstructed (by other objects) from the user's viewpoint. We hypothesize that when user eye gaze happens to simultaneously fixate on two overlapping objects, the user is likely to be more interested in the object appearing in front. Objects in the back are less likely to be attended. This aspect can be considered on either a fixation level or a time window level.

When considering visual occlusion on a fixation level, this feature can be used to clarify ambiguous fixations. Ambiguous fixations are those for which a single unique object cannot be determined. Fixations are partially disambiguated by removing all objects that are visually occluded by another object within the duration of this fixation. Unfortunately, this does not completely disambiguate fixations because a fixation may cover the area of two non-overlapping objects if these two objects are very close together.

When considering visual occlusion on a time window level—as a group of consecutive fixations in a particular time window—we calculate the amount of visual occlusion for each object during a particular time window. To take this aspect into consideration, we can use the following two variations of this feature:

- Absolute Visual Occlusion: the area of an object that is obstructed from vision by objects that were fixated during time window W.
- Relative Visual Occlusion: the percentage of an object that is obstructed from vision. This value is equivalent to the ratio between Absolute Visual Occlusion and Size for a particular object.

It is important to note the difference between representing visual occlusion on a fixation level vs. a time window level. First, two objects may be considered visually occluded during the entire time window, but not during any fixation in W. This occurs, when two overlapping objects are fixated during W, but never at the same time. Notably, if an object is visually occluded during any fixation in W, its Absolute Visual Occlusion and Relative Visual Occlusion measures must have non-zero values. The other important difference is how this feature is used in our model. When considering visual occlusion on a time window level, this feature is one of the features used to create models for attention prediction. However, when considering visual occlusion on a fixation level, it is not directly used in model creation. Instead, it is used to preprocess the data, disambiguating indistinct fixations supplied to the logistic regression framework.

Fixation Frequency Fixation frequency, which represents the number of times an object is fixated in W, may be an important feature to consider (2005). For example; if a user looks at an object, then looks away from it, and then looks back; the fixation frequency for this object will be 2. When a user looks back and forth toward an object, it is likely that the user is interested in this object.

Logistic Regression Model

Approximately half of all eye gaze fixations made during interaction with a multimodal conversation system are irrelevant to the attention prediction task. Additionally, even when a fixation is determined to be relevant it can encompass multiple objects. This makes predicting object activation quite difficult. However, several features exist to help this process. We applied the logistic regression model to examine how these features contribute to automated prediction of attention.

We formulate the attention prediction task as an *object activation* problem. This task involves identifying for each object on the graphic display, whether it is activated or not. An object is considered activated if it is, indeed, the focus of attention. Fixation intensity and the auxiliary visual features for a particular object during a specific time frame comprise the feature set for this problem, while a boolean value specifying if this object is deemed to be the focus of attention by a human annotator comprises the class label.

The binary decision of whether an object is activated or not seems to be too coarse to reflect the usefulness of our features. Therefore, instead of attempting to determine whether an object is activated or not, we determine the likelihood that an object is activated. This method allows us to make a more fine-grained evaluation because an object may have many more possible ranking scores than boolean values.

To serve our purpose, we chose the logistic regression model in our investigation because this model can be used combine an unlimited number of continuous numerical features for predicting object activation and because it directly computes the probability of an object to be activated—this value can be used to rank object of activations. This approach computes a model that best describes the data while minimizing assumptions made about how the data is generated (maximizing entropy). It can be used to objectively determine the reliability of various features for the object activation task. Features that consistently cause the model to achieve higher performance can be considered more reliable.

Logistic regression uses a well-known objective function to determine the likelihood that a given data instance belongs to a particular class. This model assumes that the log-ratio of the positive class to the negative class can be expressed as a linear combination of features as in the following equation:

$$\log\left(\frac{p(y=true|\overrightarrow{x})}{p(y=false|\overrightarrow{x})}\right) = \overrightarrow{x}\overrightarrow{w} + c$$

where the following constraint holds:

$$p(y=true|\overrightarrow{x}) + p(y=false|\overrightarrow{x}) = 1$$

Here, y refers to the class label (in our case, activated or not activated), $\overrightarrow{x}$ refers to the feature vector, and $\overrightarrow{w}$ and c are parameters to be learned from the data. $\overrightarrow{w}$ refers to the weights associated with features.

Thus, the likelihood for a particular object to be activated ($y = true$) given a set of features $\overrightarrow{x}$ can be expressed as follows:

$$p(y = true|\overrightarrow{x}) = \frac{1}{1 + e^{-\overrightarrow{x}\overrightarrow{w}-c}} \qquad (1)$$

This likelihood will allow us to rank object activation.

Based on the general model of logistic regression, in our current investigation, we use the following features discussed earlier: fixation intensity, visual occlusion, object size, and fixation frequency.

Features are combined in the logistic regression framework as show in Equation (1) (So 1993; Genkin, Lewis, & Madigan 2004).

Activation Model Training The Bayesian Logistic Regression Toolkit (Genkin, Lewis, & Madigan 2004) provided by Rutgers University is used to create computational models that rank objects of interest in a given time window W based on their likelihood of activation. In the training phase the system automatically learns the influence weights of our various features that maximize this function's correspondence with our data (or equivalently, minimizes deviation from our data). Given an unknown data instance, the resulting model provides the likelihood that this data instance belongs to the *activated* class.

Data Collection

User Study

We have conducted user studies to collect data involving user speech and eye gaze behavior. In these studies, users interact with a graphical display to describe a scene and answer questions about the scene in a conversational manner. The Eyelink II head-mounted eye tracker sampled at 250 Hz is used to track gaze fixations.

Experimental Design A simplified conversational interface is used to collect speech and gaze data. Users view a static scene of a room containing objects such as a door, a bed, desks, chairs, etc. Some of the objects in the room are arranged in a typical fashion, while other objects are out of place. Many objects visually overlap other objects. Users are asked to answer a series of 14 questions about various objects in the scene. These questions range from factual questions about particular objects to open-ended questions about collections of objects. The following are three examples of such questions:

- Is there a bed in this room?
- What do you dislike the most in this room?
- How would you like to change the furniture in the room?

The scene used in this experiment is a 2-dimensional snapshot of a 3-dimensional virtual bedroom. This scene is shown in Figure 1. The rationale behind using this scene lies in the fact that it contains many distinct objects, most of which users are familiar with. Each object is defined as a Region of Interest and forms a candidate for activation during each user utterance. No visual feedback is given to the user about which object is activated.

Data Corpus

The collected raw gaze data is extremely noisy. The raw data consists of the screen coordinates of each gaze point sampled at every four milliseconds. As can be seen in Figure 1(a), this data is not very useful for identifying fixated objects. The raw gaze data can be processed to eliminate invalid and saccadic gaze points, leaving only pertinent eye fixations. Invalid fixations occur when subjects look off the screen. Saccadic gaze points occur during ballistic eye movements between fixations. Vision studies have shown that no visual processing occurs during saccades (i.e., saccadic suppression) (Matin 1974). In addition, to removing invalid gaze points, the data is smoothed by aggregating short, consecutive fixations. It is well known that eyes do not stay still, but rather make small frequent jerky movements. In order to best determine fixation locations, five consecutive gaze locations are averaged together to identify fixations. The processed eye gaze data can be seen in Figure 1(b).

The collected eye gaze data consists of a list fixations, each of which is time-stamped and labeled with a set of interest regions. Speech data is manually transcribed and timestamped using the Audacity toolkit. Each referring expression in the speech utterance is manually annotated with the correct references to either a single object (region of interest found in the eye gaze data) or multiple objects.

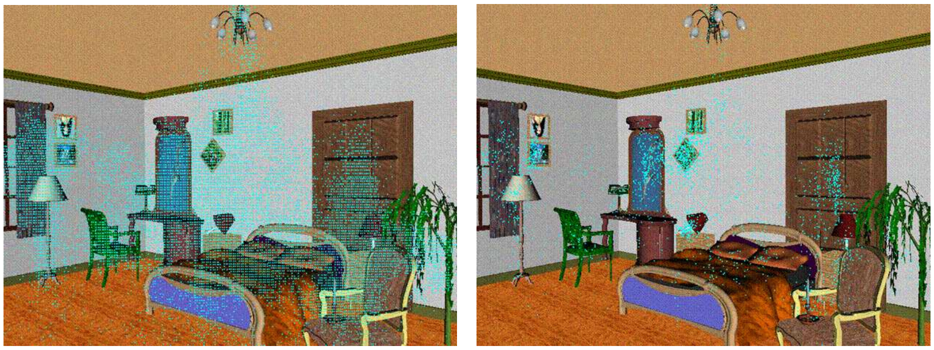
The collected data are further processed and segmented into a list of frames. Here a frame denotes a list of data instances occurring in the same time window W. Currently, we have set W to [0..1500] ms prior to the onset of an utterance referring to an on-screen interest area. However, other time windows are possible. Note that a single frame may consist of multiple data instances. In total, we have collected 449 frames containing 1586 data instances.

For each frame, all features are extracted from the eye gaze data and labeled using the id of the referenced object. The fixation intensity, fixation frequency, and visual occlusion are calculated within a particular time window from the gaze data log. These features along with visual occlusion are calculated relative to W and all of the objects that are fixated during W. This procedure can be more easily understood with the following example:

Imagine that the *dresser* object is referenced at time 6050 ms. This means that time window W is set to [4550..6050]. During this time, imagine that the user fixates *dresser* throughout most of W, looks away, and fixates it again. During W, the user also looks at *bed*, *bed cabinet* and *photo frame*. The four resulting data instances for this frame are shown in Table 1.

Empirical Results

In this section, we present the empirical results that address the four questions raised in the Introduction Section. Since



(a) Raw fixation on the display (b) Smoothed fixation on the display

Figure 1: The 3D room scene for user studies and eye fixations on the interface

Object	*dresser*	*photo frame*	*bed*	*bed cabinet*
AFI	0.6020	0.0953	0.2807	0.1967
RFI	1.0000	0.1584	0.4662	0.3267
Relative Visual Occlusion	0.0000	0.0000	0.0000	0.1500
Size	0.4312	0.0214	1.0000	0.0249
Frequency	2	1	1	1
Class Label	TRUE	FALSE	FALSE	FALSE

Table 1: Sample Data Frame with Four Instances

we need to use some metrics to evaluate the performance on attention prediction based on different features, we first explain how to apply the activation model to our dataset and present a novel evaluation metric for the object activation problem (addressesing the fourth question). We then discuss the answers to the first three questions in turn.

Application of Activation Model

Here we discuss how to apply an activation model to rank objects of interest in a given time window (represented as a frame of unclassified data instances). As we have already mentioned, an activation model can provide the probability that a particular data instance belongs to the *activated* class. Given a data frame associated with a particular time window, the model is applied to each instance in the frame. The data instances are then ranked in descending order based on their likelihood of activation as determined by the model. Note that our data collection scheme guarantees that each data instance in a particular data frame must correspond to a unique object of interest. Thus, the result is a ranked list of objects.

Q4: Evaluation Metrics

To evaluate the impact of different features on attention prediction, we borrowed the evaluation metrics used in the Information Retrieval (IR) and Question Answering (QA) fields. In these fields, three key metrics have been widely used to assess system performance are Precision, Recall, and Mean Reciprocal Ranking (MRR). In IR, Precision measures the percentage of retrieved relevant documents out of the total number of retrieved documents and Recall measures the percentage of retrieved relevant document out of the total number of relevant documents. In QA, MRR measures the average reciprocal rankings of the first correct answers to a set of questions. For example, given a question, if the rank of the first correct answer retrieved is N, then the reciprocal ranking is $1/N$. MRR measures the average performance across a set of questions in terms of their reciprocal rankings.

Given these metrics, we examined whether they could be applied to our problem of attention prediction. In the context of attention prediction, the *document retrieved* or *answer retrieved* should be replaced by *objects activated*. For example, the precision would become the percentage of correctly identified activated objects (i.e., those objects are indeed the attended objects) out of the total number of activated objects. The reciprocal ranking measurement would become the reciprocal ranking of the first correctly activated object.

Since the result from the logistical regression model is the likelihood for an object to be activated, it is difficult to precisely determine the number of objects that are activated based on the likelihood. Presumably, we can set up a threshold and consider all the objects with the likelihood above that threshold activated. However, it will be difficult to determine such a threshold. Nevertheless, the likelihood of activation can lead to the ranking of the objects that are likely to be activated. Thus, the desired evaluation metric for this investigation should determine how well our model ranks the objects in terms of their possible activation. For these reasons, we decided that the precision and recall metrics are not suitable for our problem and MRR seems more appropriate.

Even with the MRR measurement, there is still a problem. MRR in QA is concerned with the reciprocal ranking of the first correct answer; but, here in attention prediction, multiple objects could be simultaneously attended to (i.e., acti-

Object	*dresser*	*lamp*	*bed lamp*	*bed*
Class Label	FALSE	TRUE	TRUE	FALSE
Rank	1	2	3	4

Table 2: Sample Test Data Frame with Four Ranked Instances

vated) in a given frame. We need to consider the reciprocal ranking for each of these objects. Therefore, we extended the traditional MRR to a *normalized* MRR (NMRR) which takes all attended objects into consideration. The normalized MRR is defined in Equation (2)

$$NMRR = \frac{MRR}{UpperBoundMRR} \quad (2)$$

where the upper bound MRR represents the MRR of the best possible ranking.

For example, suppose the logistic regression model ranks the likelihood of activation in a frame of four objects as shown in Table 2. Among these objects, only *lamp* and *bed lamp* are referenced in user speech within this frame. In this case,

$$NMRR = \frac{1/2 * (1/2 + 1/3)}{1/2 * (1 + 1/2)} = 0.556$$

Here, the numerator represents the mean reciprocal ranking of the predicted activations in this frame. The denominator represents the MRR of the best possible ranking, which in this case would rank lamp and bed lamp as the top two ranked objects.

With the evaluation metrics defined, next we report empirical results that address the remaining three questions.

Q1: Temporal Distribution of Gaze Fixations

Previous psycholinguistic studies have shown that eye gaze fixations to an object in a visual scene occur, on average, between 630 to 932 ms before the onset of a spoken reference in a language production task. Pertinent eye fixations—that is, those to objects that will be referenced—can range anywhere from 1500 ms before onset up until onset of a spoken reference. Knowing the range and the mean does not provide sufficient information about the nature of pertinent eye fixations in our complex scenes. We conducted an investigation to obtain a more accurate picture of the temporal distribution of eye gaze fixations relative to the onset of a spoken reference in a multimodal conversational domain.

Figure 2 shows a histogram reflecting the percentage of pertinent fixations. Here the X-axis represents the starting point of a fixation within time window W, occurring prior to the onset of a reference to an object. This value ranges from 1500 ms before onset until precisely at onset. Fixations are clumped together into interval bins of 100 ms. The Y-axis represents the proportion of fixations that contain an object matching the referenced object. This proportion was calculated with the following procedure:

1. Each fixation is classified as pertinent or irrelevant. Irrelevant fixations are those that do not contain an object that is reference within time window W. Note that a several objects may be fixated during a single fixation. Also, note that this classification occurs relative to a particular spoken reference. Thus, a particular fixation can be classified as pertinent for one reference, but irrelevant for another.
2. Each fixation is classified into a bin of length 100 ms. The bins represent the amount of time that passes between the start of an eye fixation and an object reference. For example, the bin labeled 200 contains all fixations starting between 200 and 300 ms prior to an object reference.
3. To calculate the percentage, the number of pertinent fixations in each bin is divided by the total number of fixations in this bin.

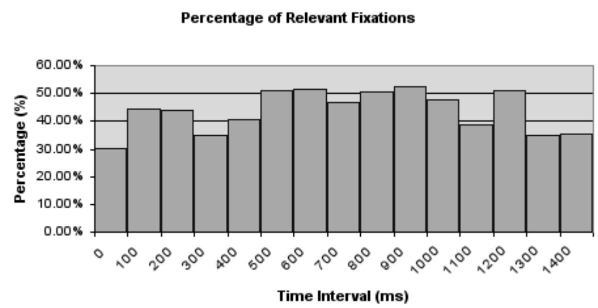


Figure 2: Proportion of Pertinent Eye Fixations Divided into 100 ms Interval Bins

It is important to determine which time periods are most likely to generate a pertinent fixation. In order to determine this, we found the mean (μ) time weighted by the likelihood of a pertinent fixation appearing during this time bin. Assuming that the data represents a skew-normal distribution, we also found the standard deviation (σ^2), and skewness (γ_1). We obtained the following results: $\mu = 758$, $\sigma^2 = 466$, and $\gamma_1 = -1.22$.

The large standard deviation indicates that pertinent fixations are fairly evenly distributed during the 1500 ms time interval prior to a spoken reference. Nevertheless, as can be seen from Figure 2, there is a general trend that a fixation is more likely to be pertinent close to the mean rather than far from the mean. The fixation data is shown to have a negative skew. That is, the left (lower value) tail of the graph is longer. Thus, a fixation is more likely to be pertinent if its to the right of the mean—further from the spoken reference, but still within the 1500 ms time range—than to the left.

Q2: Evaluation of Fixation Intensity

To evaluate the role of fixation intensity (i.e., Question 2) and auxiliary features (i.e., Question 3) in attention prediction, we conducted a five-fold cross validation. More specifically, the collected data are randomly divided into five sets. Four of these sets are used for training, while the remaining set is used for testing. This procedure is repeated five times and the averaged results are reported in the following sections. The object activation models were created using two data sets. The first is the original data set described in

	NMRR evaluation	
Fixation Intensity Variation	Original Data Set	Preprocessed Data Set (Disambiguated Fixations)
AFI	0.661	0.752
WAFI	0.656	0.745
RFI	0.652	0.754
WRFI	0.650	0.750

Table 3: Evaluation of Fixation Intensity Weighting

Data Corpus section, while the second is a version of this same data set that is preprocessed to partially disambiguate fixations to multiple objects. Fixations in the preprocessed data set are disambiguated using the visual occlusion feature considered on a fixation level.

In this section we compare object activation models created by using each of the four variations of the fixation intensity measure. The goal here is to determine the effect of weighting fixations based on their distributions of starting times relative to a spoken reference versus treating every fixation equally. First, we discuss the methodology for creating weighted fixation intensity measures. Then we present results comparing the various object activation models and discuss their implications.

To create our two weighted fixation intensity measures (WAFI and WRFI) we use the statistics acquired about the distribution of fixations starts. More precisely, we weight each fixation by a skew-normal density function (Azzalini & Valle 1996) with mean, standard deviation, and skewness discovered while addressing Question 1.

The results of each model constructed from its corresponding variation of the fixation intensity feature are are shown in Table 3. These results clearly indicate that there is very little variation among the different representations of fixation intensity across each of the two datasets. The first thing to note is that this lack of variation is to be expected between relative and absolute versions of the same fixation intensity measurement (AFI vs. RFI and WAFI vs. WRFI). This is because the evaluation conducted here is based on mean reciprocal ranking. Given two objects in a single time frame W, the one with the higher absolute fixation intensity is guaranteed to have a higher relative fixation intensity. Thus, the object ranking remains unchanged.

The effect of weighting fixation intensity seems to decrease performance of the object activation task. This decrease is very slight and likely insignificant. Nonetheless, this result is somewhat vexing as we expected that weighting the fixation intensity would improve prediction of object activation. One possible explanation for this lack of improvement is that fixations are fairly evenly distributed during each time frame W. This makes the weight function very flat and virtually insignificant. Another possibility is that the distribution of fixation starts has multiple peaks rather than a single peak at the mean as is the assumption of the normal distribution. Thus, neither the normal distribution nor the skew-normal distribution accurately models the distribution of eye fixation starts relative to spoken object references.

Q3: Evaluation of Auxiliary Visual Features

In this section we evaluate the performance of auxilliary visual features in the object activation task. Configurations of various combinations of these features with fixation intensity are examined. Given that the effect of weighting fixation intensity is insignificant, only AFI and RFI are considered. The results are shown in Table 4 and discussed separately for each feature.

Visual Occlusion As Table 4 shows, all configurations that use the preprocessed data set, which augments the fixation intensity measurement with a fixation-level account of visual occlusion, perform significantly better than their counterparts that use the original data set. The only difference between the preprocessed and original data sets is the incorporation of the fixation-level visual occlusion feature. This clearly means that visual occlusion is a reliable feature for the object activation prediction task. However, it is also clear that the representation of visual occlusion is very important. Adding the frame-based visual occlusion feature to the logistic regression model (rows 2 and 3) has almost no effect. It may be possible that a better representation for visual occlusion remains unexplored.

On average, the effect of both absolute and visual occlusion is more significant for the original data set (especially when RFI is used). This is not surprising because the preprocessed data set partially incorporates the visual occlusion feature, so the logistic regression model does not get an added bonus for using this feature twice.

Object Size The object size feature seems to be a weak predictor of object activation. Using only the fixation intensity and object size features (row 4 of Table 4), the logistic regression model tends to achieve approximately the same performance as when the object size feature is excluded.

This result is quite unexpected. As we have already mentioned, human eye gaze is very jittery. Our expectation is that small objects can have a low fixation intensity and still be activated. Thus, in our model small objects should need a lower fixation intensity to be considered as activated than do large objects. Our results do not support this general trend. A possible explanation is that this trend should only be apparent when using a visual interface with a mixture of large and small objects. In our interface, most of the objects are fairly large. For large object, jittery eye movements do not alter fixation intensity because the eye jitter does not cause fixations to occur outside of an object's interest area boundary. Even when some objects are smaller than others, they are not sufficiently small to be affected by eye jitter. Thus, it is likely that the size feature should only be considered when comparing fixation intensities of sufficiently small objects to larger counterparts. At this point, it is unclear how small is sufficiently small.

Fixation Frequency Fixation frequency is a weak predictor of object activation. Incorporating the fixation frequency feature into the Bayesian Logistic Regression framework creates models that tend to achieve worse performance than

Row	Features	Original Data Set		Preprocessed Data Set (Disambiguated Fixations)	
		AFI	RFI	AFI	RFI
1	Fixation Intensity Alone	0.661	0.652	0.752	0.754
2	Fixation Intensity + Absolute Visual Occlusion	0.667	0.666	0.758	0.754
3	Fixation Intensity + Relative Visual Occlusion	0.667	0.669	0.762	0.751
4	Fixation Intensity + Size	0.656	0.657	0.763	0.759
5	Fixation Intensity + Fixation Frequency	0.653	0.644	0.743	0.756
6	All features (Absolute Visual Occlusion)	0.662	0.663	0.768	0.760
7	All features (Relative Visual Occlusion)	0.660	0.669	0.764	0.768

Table 4: Evaluation of Auxiliary Features

when this feature is left out. At best, models using fixation frequency achieve a comparable performance to those not using it. According to Qvarfodt (Qvarfordt & Zhai 2005), fixation frequency is an important feature to consider because one way of signifying interest in objects is looking back and forth between two or more objects. In this case, each of these objects would have a fairly low fixation intensity as time is spent across multiple objects, but each of the objects should be considered activated. In our user studies, however, we did not find this user behavior. This behavior is likely to be specific to the map-based route planning domain where users often need to look back and forth between their starting and destination location.

Conclusion

We have shown that fixations that are pertinent in the object activation problem are fairly evenly distributed between the onset of a spoken reference and 1500 ms prior. Fixation intensity can be used to predict object activation. Weighting fixations based on a skew-normal distribution does not improve performance on the object activation task. However, preprocessing our fixation data by including the fixation-level visual occlusion feature considerably improves reliability of the fixation intensity feature. Moreover, since performance is so sensitive to feature representation, there is much potential for improvement. We have also presented the NMRR evaluation metric that can be used to evaluate the quality of a ranked list.

This work can be extended to combine our activation model with spoken language processing to improve interpretation. This question can be addressed by constructing an N-best list of spoken input with an Speech Recognizer (ASR). The speech-based ranked lists of utterances and the gaze-base ranked lists of activations can be used to mutually disambiguate (Oviatt 1999) each other in order to more accurately determine the object(s) of interest given an utterance and a graphical display. This knowledge can be used to plan dialog moves (e.g. detect topic shifts, detect low-confidence interpretations, determine the need for confirmation and clarification sub-dialogs, etc.) as well as to perform multimodal reference resolution (Chai *et al.* 2005). We believe that this work will open new directions for using eye gaze in spoken language understanding.

References

Azzalini, A., and Valle, A. D. 1996. The multivariate skew-normal distribution. In *Biometrika*, volume 83, 715–726.

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

Genkin, A.; Lewis, D.; and Madigan, D. 2004. Large scale bayesian logistic regression for text categorization. In *Journal of Machine Learning, submitted*.

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

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

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

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

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

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

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

Oviatt, S. 1999. Mutual disambiguation of recognition errors in a multimodal architecture. In *Proc. Of the Conference on Human Factors in Computing Systems*. ACM.

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

Sibert, L., and Jacob, R. 2000. Evaluation of eye gaze interaction. In *Proceedings of CHI'00*, 281–288.

So, Y. 1993. A tutorial on logistic regression. In *Proc. Eighteenth Annual SAS Users Group International Conference*.

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