Predicting User Attention using Eye Gaze in Conversational Interfaces

Zahar Prasov
Department of Computer Science
Michigan State University
East Lansing, MI 48823
prasovza@msu.edu

Joyce Chai
Department of Computer Science
Michigan State University
East Lansing, MI 48823
jchai@cse.msu.edu

Abstract

In a conversational system, determining the 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 conversation. In particular, we are developing techniques that can predict an objects' activation in a given time frame based on user eye gaze behavior. The predicated object activation can be combined with speech input to better identify user attention. This paper describes our on-going effort on this topic.

1 Introduction

Previous studies have shown that eye gaze is one of the reliable indicators of what a person is "thinking about" (Henderson, 2004). The direction of gaze carries information about the focus of the user's attention (Just, 1976). In human language processing tasks specifically, eye gaze is tightly linked to processing. The perceived visual context influences spoken word recognition and mediates syntactic processing (Tanenhous, 1995). Additionally, directly before speaking a word, the eyes move to the mentioned object (Griffin, 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. This research is conducted in a conversational system where users can look at a graphical interface and converse with the system through speech. This is different from traditional spoken dialog systems in that the system receives eye gaze information in addition to speech utterances. It is also different from multimodal dialog systems in that, instead of being proactively provided by a user (as a pen-based gesture), eye gaze in the proposed work is a subconscious input which naturally occurs with speech utterances. Since eye movements are executed automatically and involuntarily, this unique setting is fundamental to understanding any human speech communication with graphical interfaces.

As a first step in our investigation, we are currently examining how eye gaze contributes to automated identification of user attention during conversation. In particular we are developing techniques that can predict an object's activation in a given time frame based on user eye gaze behavior. The distribution of object activation can be used to disambiguate speech recognition results and improve input interpretation. This paper reports our work on this topic—primarily addressing the object activation problem—and the current results.

2 Related Work

There have been several attempts to use eye gaze to facilitate interaction in human-machine communication.

Kaur, et. al. (2003) explores the temporal alignment between eye gaze and speech during a simple on-screen object movement task. Users move designated objects, in a potentially ambiguous setting

—to new locations by using speech and gaze as a pointing mechanism. The results have shown that the eye fixation that most likely identifies the object to be moved occur 630 ms (on average) before the onset of the commanding utterance.

In the iTourist project, Qvarfordt et. al. (2005a) 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 (IScore) 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 this project, object activation was determined by a quadratic combination of eye gaze intensity along with a linear combination of auxiliary features extracted from eye gaze data. The influence weights of the factors were empirically determined. The IScore calculation is shown in Figure 1.

$$IScore=\alpha_i(1+\alpha(1-\alpha_i))=\alpha_i+\alpha_i\alpha-\alpha_{i^2}\alpha$$

where

$$\alpha=c_i\frac{c_f\alpha_f+c_c\alpha_c+c_s\alpha_s+c_a\alpha_a}{c_f+c_c+c_s+c_a}=\frac{\sum_{x=1}^{N}c_i c_x\alpha_x}{\sum_{x=1}^{N}c_x}$$

α_i = Absolute Fixation Intensity
α_f = Fixation Frequency
α_c = Categorical Relationship
α_s = Object Size
α_a = IScore of the previous active object
c_i, c_f, c_c, c_s, c_a = empirically determined constants

Figure 1. IScore calculation

In this paper we investigate various features associated with eye gaze that contribute to attention prediction. In particular, we extend the model developed in Qvarfordt (2005a) and present an approach that automatically learns the weights associated with different features for the prediction.

3 Identifying Object Activation

The object activation problem is addressed by formulating it as a binary classification problem for each object that appears on the interface at a specific time period. Here, the class label for each object is true if the user is, in fact, attending to the object and false otherwise.

This model can be a regression model (So, 1993) rather than a classification model. Such a model can be used to create an N-best ranking of activated objects during a given time window rather than dividing the objects into an activated and a non-activated set.

In this paper we describe an extension of the IScore function that we use to create our N-best ranking. Our goal is twofold. Our first aim is to examine the features that are useful for prediction of object activation. Our second aim is to learn the optimal weight for each feature in our ranking function. The activation ranking function described in this paper is based on the likelihood of object activation.

3.1 Logistic Regression

Logistic regression uses a well-known objective function to determine the likelihood that a given data instance contains a particular class label. The logistic regression model assumes that the log-ratio of the positive class to the negative class can be expressed as a linear combination of features. The result of this assumption is shown in Figure 2. Here, y refers to the class label, $\vec{x}$ refers to the feature vector, and $\vec{w}$ refers to the influence weight constants.

$$\frac{p(y=true|\vec{x})}{p(y=false|\vec{x})}=e^{\vec{x}\cdot\vec{w}+c}$$
$$p(y=true|\vec{x})+p(y=false|\vec{x})=1$$

Figure 2. Logistic Regression Model

We are interested in ranking data instances based on their likelihood of activation (y = true). This can be done by solving for $p(y=true|\vec{x})$ as shown in Figure 3.

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

Figure 3. Likelihood of Activation

3.2 Standard Feature Set

It is important to consider the features that can be useful for this regression task. While fixation intensity is likely to be the most important factor, we hypothesize that the other listed factors will have a strong contribution towards our model's accuracy. Our regression model currently uses the following features:

- Absolute Fixation Intensity (AFI): the amount of time spent fixating on an object during a particular time window W. In order to normalize this feature to maintain a range of values between 0 and 1, AFI is divided by W. 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): AFI of a candidate object in time window W relative to the candidate object with the maximal AFI in W. This feature inherently ranges between 0 and 1. Given that our goal is to rank the candidate objects, we want to consider the fixation intensities only of those objects appearing in W. A user may look away from the screen during a portion of this time window while clearly signaling the activation of an object. Thus, this object may have a low AFI. RFI compensates for such a situation.
- Fixation Frequency: the number of times an object was fixated in W. 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. This feature constitutes a discrete value ranging from 0 to infinity (though 10 was the highest value encountered in practice). When a user looks back and forth toward an object, it seems likely that the user is interested in this object.
- Object Size: the area of a candidate object relative to a baseline object (the smallest object in the scene). This feature ranges from 1 to infinity (capping out at about 100 in practice). Each object is specified by a list of (x, y) scene coordinates. An object's area is represented by the bounding box considering only the minimal and maximal x and y coordinates. Size is considered to be a useful feature because 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, it would seem that a lower fixation intensity is necessary to activate a small object than a large object.
- Object "Frontness": the number of objects appearing in W that obstruct a candidate object from the users viewpoint. This value ranges between 0 and infinity (no more than 10 in practice). We hypothesize that when users simultaneously look at two objects, they are more interested in the objects appearing in front. Thus, the likelihood of activation of objects appearing behind other objects should be discounted.

The categorical relationship (between the previous activated object and the current candidate object) feature was ignored because categories have not been assigned to the objects in our domain.

Figure 4 shows the ranking function for our model applied to the standard feature set. Here, y refers to the class label, $\vec{x}$ refers to the feature vector, α_x refers to a particular feature in $\vec{x}$, and c_x refers to the influence weight of this feature.

$$p(y=true|\vec{x}) = \frac{1}{1+\exp(\sum_{x=0}^{N} \alpha_x c_x)}$$

Figure 4. Logistic Regression Applied to Standard Feature Set

3.3 Extended Feature Set

The IScore feature set can be incorporated into the logistic regression framework by viewing each combination of features as a single complex feature. The IScore can be decomposed as shown in Figure 5a.

Each feature α_x (except α_i) can be split into two features as shown in Figure 5b. Thus, if there are originally N auxiliary features to fixation intensity, the extended data set will contain 2N+1 features (the fixation intensity feature plus double the auxiliary features). Figure 5c. shows the resulting

ranking function after the extended feature set is applied to our model.

Using this feature set with logistic regression will find the optimal values for the constants. One thing to note is that we do not ensure that

$$c_x^{'} = -c_x^{"}$$

for all x.

$$IScore = \alpha_i + \alpha_i \frac{\sum_{x=1}^{N} c_i c_x \alpha_x}{\sum_{x=1}^{N} c_x} - (\alpha_i)^2 \frac{\sum_{x=1}^{N} c_i c_x \alpha_x}{\sum_{x=1}^{N} c_x}$$

$$IScore = \sum_{x=1}^{N} \left(\alpha_i + \alpha_i \alpha_x \frac{c_i c_x}{\sum_{x=1}^{N} c_x} - (\alpha_i)^2 \alpha_x \frac{c_i c_x}{\sum_{x=1}^{N} c_x} \right)$$

$$\text{let } c_x^{'} = -c_x^{"} = \frac{c_i c_x}{\sum_{x=1}^{n} c_x}$$

$$IScore = \sum_{x=1}^{N} (\alpha_i + c_x^{'} \alpha_i \alpha_x + c_x^{''} (\alpha_i)^2 \alpha_x)$$

Figure 5a. IScore Reformulation

$$\alpha_i$$

$$\text{for each feature } \begin{Bmatrix} \beta_x^{'} = \alpha_i \alpha_x \\ \beta_x^{"} = (\alpha_i)^2 \alpha_x \end{Bmatrix}$$

Figure 5b. Feature Extension

$$p(y = true \mid \vec{x}) = \frac{1}{1 + \exp(\alpha_i) \exp(\sum_{x=1}^{N} (c_x^{'} \beta_x^{'} + c_x^{"} \beta_x^{"}))}$$

Figure 5c. Logistic Regression Applied to Extended Feature Set

Figure 5. Extended Feature Set

The extended feature set may outperform the standard feature set because it centers around fixation intensity. All other features are considered secondary. The standard feature set assumes each feature has equivalent predictive power. We do not expect this to be the case.

3.4 Model Training and Testing

Our goal is to build a computational model that can rank objects of interest in a given time frame based on their likelihood of activation (they are the focus of attention). Thus, the goal of the training phase is to build a model that can determine the probability that an object is activated. The goal of the testing phase is to rank candidate objects for activation according to the model.

The logistic regression framework consists of an objective function that allows us to estimate the likelihood that an object is activated. In the training phase we learn the influence weights of our various features that maximize this function's correspondence with our data (or equivalently, minimize deviation from our data).

In our framework, training data consists of instances with features pertaining to a particular object's activation as well as a binary class label denoting whether the object is activated or idle. Test data consists of frames of the same type of data instances. Each data instance in a frame relates to a single candidate object for activation. These instances are ranked according to our model for each frame supplied in our test data.

4 Data Collection

4.1 User Study

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

4.1.1 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 expected fashion, while other objects are out of place. Many objects visually overlap (are in front or behind) 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 par-

ticular objects to open-ended questions about collections of objects.

The image used in this experiment is a 2-dimensional snapshot of a 3-dimensional virtual bedroom. This scene is shown in Figure 6. The rational behind using this image lies in the fact that the scene contains a conglomeration of 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. Here, activation refers to object(s) that a user attends to during an utterance. No visual feedback is given to the user about which object is activated.

Figure 6. Bedroom Scene

4.1.2 Equipment

User eye gaze is recorded using the Eye Link II eye tracker sampled at 250 Hz. Eye fixations were garnered using only pupil reflection (rather than pupil and corneal reflection). Eye gaze is deemed to be a fixation when five consecutive gaze points appear within a threshold distance of each other. All regions of interest circumscribing a fixation point are considered to be fixated.

User speech is recorded using a noise-canceling microphone. The Audacity toolkit along with a human annotator is used to timestamp each word in the resulting speech file.

4.1.3 Procedure

SR Experiment builder is used to construct the flow of the experiment. Our experiment involves 14 trials (one per question) that are randomly split into 3 sessions. Each session is preceded by eye tracker calibration. Re-calibration of the eye tracker at least every 10 minutes ensures accuracy.

A particular trial involves displaying the scene to the user, asking the user a question about the scene, and recording their response as well as eye gaze. The user decides when he or she is finished with the trial by pressing any key on the keyboard. Each trial is followed by drift correction that forces the user to look at the center of the screen. This has two purposes. First, it acts as a mini-calibration that compensates for the drift between the eye and the crosshairs on the eye tracker. Second, it ensures that all trials begin with the user looking at the same point in the image. This eliminates any user bias of looking at a particular object before speech onset.

4.2 Data Corpus

The collected eye gaze data consists of a list fixation, each of which is time-stamped and labeled with a set of regions of interest. Speech data is manually transcribed and time-stamped using the freeware Audacity tool. Each reference to an object or multiple objects is manually labeled with the IDs of the objects; these IDs correspond to the IDs of the regions of interest found in the eye gaze data.

A frame of training/testing data is derived from each spoken reference. For each frame, features mentioned in section 3.2 are extracted from the eye gaze data and labeled using the id of the referenced object. As seen in Table 1, a single frame may consist of multiple data instances.

The fixation intensity, fixation frequency, and object frontness are calculated within a particular time-window from the gaze data log. This window ends when a reference to an area of interest is uttered and begins W milliseconds before the utterance. Currently we have set time window W to -1500 ms. However, other time windows are possible. The fixation intensity, fixation frequency, and object frontness 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 was 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 lamp},

and {photo frame}. The 4 resulting data instances for this frame are shown in Table 1.

Object	{dresser}	{photo frame}	{bed}	{bed lamp}
Absolute Fixation Intensity	0.602	0.0953	0.2807	0.1967
Relative Fixation Intensity	1.0000	0.1584	0.4662	0.3267
Fixation Frequency	2	1	2	1
Size	21.6155	1.0738	50.1286	1.2481
Frontness	2	0	0	0
Class Label	TRUE	FALSE	FALSE	FALSE

Table 1. Sample Data Frame with 4 instances

4.3 Activation Models

In total, we have collected 84 frames containing 244 data instances. This data can easily be converted to the extended feature set. All results reported in this paper make use of this data.

The resulting data frames are randomly divided into five sets used in a five-fold cross validation. 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 section 5.

The Bayesian Logistic Regression Toolkit (Genkin, 2004) provided by Rutgers University was used to create all logistic regression models presented in this paper. The resulting logistic regression models are used to rank the data instances in each test data frame.

5 Results

The evaluation was conducted by computing the Mean Reciprocal Rank (MRR) of each frame as ranked by the algorithm relative to the correct ranking. Given that a single test data frame may contain multiple positive instances (corresponding to multiple activated objects) the MRR was normalized by the upper bound (highest achievable) MRR. Figure 7 shows the Normalized MRR calculation of the sample frame appearing in Table 2.

Our logistic regression models with both the standard and extended feature sets are compared to a baseline of ranking the test frame instances based on the absolute fixation intensity. Note that the baseline ranking would be exactly the same if relative fixation intensity was used because each frame is evaluated independently of other frames. Given two objects, the one with the higher AFI is guaranteed to have a higher RFI.

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

Table 2. Sample Test Data Frame with 4 ranked instances

The MRR for this frame is

$$\text{Normalized MRR} = \frac{\text{MRR}}{\text{Upper Bound MRR}} = \frac{\frac{1}{2}\left(\frac{1}{2}+\frac{1}{3}\right)}{\frac{1}{2}\left(1+\frac{1}{2}\right)} = \frac{0.417}{0.75} = 0.556$$

Figure 7. Normalized MRR Calculation for Sample Test Data Frame

The result comparing our models with the standard as well as the extended feature set is shown in Table 3.

	Logistic Regression Feature Set	
	Standard	Extended
Absolute Fixation Intensity (AFI) Only	0.582	
Relative Fixation Intensity (RFI) Only	0.776	
RFI + Size	0.634	0.736
RFI + Frontness	0.762	0.770
RFI + Frequency	0.637	0.743
ALL	0.642	0.775
	Baseline	
Rank by AFI (or equivalently RFI)	0.769	

Table 3. MRR evaluation

Table 3 also shows that the Extended Feature set outperforms the Standard feature set regardless of which individual features are used. This finding may mean that the Extended Feature set better represents the data in the object activation detection problem or that logistic regression models employing this feature set are less sensitive to noisy data.

The overall results show that fixation intensity alone is a good indicator of object activation. This is the feature used to construct our baseline of

ranking the test frame instances. Using only the Relative Fixation Intensity in our logistic regression framework achieves slightly better results than this baseline. Most auxiliary features do not appear to aid the object activation detection. Modifying these features along with the time window in which they are collected may lead to better results. Thus, there is still much potential for improvement.

5.1 Feature Set Evaluation

Logistic regression with the extended feature set seems to perform better than logistic regression with the regular feature set regardless of the features used. Clearly, if only the fixation intensity feature is used, the algorithms are exactly the same. We cannot conclude that the extended feature set is necessarily better than the regular feature set, we can only conclude that it is less sensitive to noisy data. Given our results it appears that the size and frequency features consist of a large amount of noise and even the frontness feature is somewhat noisy. These features cause logistic regression with the extended feature set to perform only slightly worse than using only the fixation intensity. However, these features cause logistic regression with the regular feature set to perform significantly worse than fixation intensity alone.

5.2 Individual Feature Evaluation

It appears that the auxiliary features do not improve object activation detection performance. Models combining RFI with either Size or Frequency alone perform worse than models using all listed features (with the exception of AFI, which is considered a redundant feature when RFI is used). The frontness feature does not appear to help or hinder model performance.

One potential folly of our feature set is that both the frequency and frontness features are too coarse. These are discrete valued features that have a limited range of values in practice. It may be difficult to fit a regression model to these discrete values. A possible way to improve the frontness feature is to determine the percentage of the area of the candidate object that is obstructed rather than counting the number of objects obstructing this candidate. Making frequency a continuous value is a more difficult task. The best we can do is to normalize frequency over the average or maximal frequency over an entire user session. Even this case will result in relatively few distinct frequency values making it difficult to fit a function curve to this data.

As we already mentioned, the single fixation intensity feature is outperforms almost every other configuration. This feature is the key to successful identification of object activation. Even a small improvement to this feature is likely to lead to an improvement in overall performance. Fixation intensity can potentially be improved by further using psycholinguistic knowledge to aid in identifying object activation. According to (Kaur, 2003), an object is fixated about 600 ms before it is referenced. This seems to mean that objects fixated closer to 600 ms before a reference are more likely to be activated than other objects. Fixation intensity can likely be improved by weighting it by a function that gives higher weight to objects fixated closer to 600 ms. The current algorithms give each fixation in the 1500 ms time window equal weight.

6 Discussion

We have shown that fixation intensity can be used to predict object activation. Moreover, there is much potential for improvement. Other features are yet to be explored. The current features can still be augmented to improve performance. The time window W has not yet been optimized. Additionally, the results described in this paper were evaluated only on 84 frames with 244 instances. More data needs to be collected and evaluated to obtain conclusive results. This work is currently ongoing.

This preliminary work has set up a framework that can use eye gaze for predicting object activation. Even if it is determined that logistic regression along with feature auxiliary to fixation intensity are not very useful, we have shown that using the baseline of fixation intensity can achieve fairly accurate results. This activation model, alone, can be used as a cornerstone for improving interpretation in speech and eye-gaze conversational systems.

This work can be extended to consider how 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 Automated

Speech Recognizer (ASR). The speech-based ranked lists of utterances and the gaze-based 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, 2005). We believe that this work will open new directions for using eye gaze in spoken language understanding.

7 Acknowledgements

This work is supported by an IGERT training grant and grant IIS-0535112 from National Science Foundation.

Our thanks go to Cassandra Nicole Jackson and Zubin Abraham for assisting with the design and execution of the user study, Graham Pierce and Jiye Shen for Eyelink II expertise, and to Venkata Ravi Dayana for assistance in data analysis.

References

J.Y. Chai, Z. Prasov, J. Blaim, R. Jin (2005). *Linguistic Theories in Efficient Multimodal Reference Resolution: An Empirical Investigation*. International Conference of Intelligent User Interfaces, San Diego, CA, ACM Press.

A. Genkin, D. Lewis, & D. Madigan (2004). *Large-scale Bayesian Logistic Regression for Text Categorization*. Journal of Machine Learning, submitted, 2004.

Z.M. Griffin, & K. Bock (2000). *What the eyes say about speaking*. Psychological Science, 11, 274-279.

J.M. Henderson & F. Ferreira (2004). T*he interface of language, vision, and action: Eye movements and the visual world*. New York: Taylor & Francis.

M. Johnston (1998). *Unification-based Multimodal Parsing*. ACL/COLING.

M.A. Just & P.A. Carpenter (1976). *Eye fixations and cognitive processes*. Cognitive Psychology 8, 441-480.

M. Kaur, et. Al. (2003) *Where is "it"? event synchronization in gaze-speech input systems*. Proc. Fifth International Conference on Multimodal Interfaces, 151-157.

S.L. Oviatt (1999). *Mutual disambiguation of recognition errors in a multimodal architecture*. Proc. Of the Conference on Human Factors in Computing Systems, New York.

P. Qvarfordt, S. Zhai (2005a). *Conversing with the User Based on Eye-Gaze Patterns*. Proc. CHI.

Pernilla Qvarfordt, David Beymer, and Shumin Zhai (2005b). *RealTourist – A Study of Augmenting Human-Human and Human-Computer Dialogue with Eye-Gaze Overlay*.

M.K. Tanenhous, M. Spivey-Knowlton, E. Eberhard, and J. Sedivy (1995). *Integration of Visual and Linguistic Information during Spoken Language Comprehension*. Science, 268:1632-1634.

Y. So (1993). *A Tutorial on Logistic Regression*. Proc. Eighteenth Annual SAS Users Group International Conference, New York, NY.