EYE GAZE FOR UNDERSTANDING CONVERSATIONAL SPEECH

Anna Prokofieva[1,2] , Dilek Hakkani-Tür[1], Malcolm Slaney[1]

Microsoft Research[1], Columbia University[2]

ABSTRACT

Eye gaze is a useful indication of attention and, as such, can be a valuable feature to improve spoken-language understanding in human-computer interaction. Based on the hypothesis that users look at a link before selecting it, we investigate the use of novel eye-gaze features to improve link click event prediction. Our data comprises users performing a variety of online tasks such as form filling and web browsing, and we show significant performance improvement by incorporating the use of gaze features. In addition, our analysis shows that there is much user-specific variation in gaze, so we are also looking to improve the modeling of gaze by user- and task-specific adaptation.

Index Terms— eye gaze, SLU

1. INTRODUCTION

With the proliferation of computerized devices, humans have more chances to interact with screens both large and small. Speech-directed interaction is in the mainstream, with the computer providing spoken output. However, we find it important to also investigate situations where the output is visual rather than spoken. This design allows the system to provide a greater quantity of information to the user all at the same time, giving them more choice and increasing the probability of returning a result that the user deems acceptable. The issue of interpreting what is presented on the screen arises in these situations, however, since on-screen object identification is necessary in order to understand what the user is saying [2], [7]. We seek to solve this problem by investigating a variety of gaze-based features by themselves and when combined with lexical features in our current experiments.

We approach this task by collecting data in a Wizard of Oz experiment. By posing a variety of web-browsing tasks to our test subjects, we cover a number of display configurations that allow us to understand a user's eye-gaze behavior. Other possible tasks that have been examined include object selection or form filling. However, we wanted to focus on tasks that users may perform more often in their daily lives – such as buying plane tickets, or finding a restaurant – as those would be the scenarios in which multi-modal systems would be used. We looked at a variety of features, both having to do with the user's gaze behavior, and with the details of the screen display. We combined these gaze features with lexical features and we found that a combination of both produced the best results, as gaze-based features seemed to complement the lexical features when the user's utterances did not contain explicit mentions of the desired link.

This paper is organized in the following manner: a review of previous work in both measuring eye gaze and using it as input is found in section 2; section 3 contains a description of the data used in our experiments; a list of the features we used follows in section 4; section 5 presents the results of our classification experiments; and in section 6, we provide some analysis of trends we discovered in the data during the course of our work.

2. PREVIOUS WORK

Our research is novel in that we investigate free-form web-browsing tasks with eye-gaze data. Most previous work focused on collecting gaze data from users doing passive question-answering based on the contents of the display or performing very narrowly-defined tasks that might not generalize to general human-computer interaction.

Previous work looked at three directions: one group of researchers has focused on measuring gaze in a set of very specific object selection tasks; another has attempted to use gaze as an indicator of the user's interest or attention; and the third has delved into using gaze as a way to manipulate the screen (in an eye-gaze-as-mouse scenario).

In the first camp, there is a variety of work that attempts to measure eye gaze as it relates to the users' utterances. Prasov et al. [11] asked users to answer simple questions about a scene displayed on a screen (eg. "Is there a bed in the room?") and measured the users' gaze throughout. They then used different measures incorporating eye gaze fixation points in order to predict which object the users were talking about. Another set of experiments asked users to describe scenes depicted on-screen were carried out by Griffin (in [4] and [5]); the users had to say a prescribed statement about the locations of the objects displayed. Kaur et al. [8] also wanted to investigate the relationship between deictic speech and gaze positioning, and had users fixate on a highlighted object on the screen while telling the system to move it to another designated spot on the screen. This work is just one example of many conducted in the put-that-there paradigm, as originated by Bolt [1]. In all instances however, the domains were very limited. Only one scene was tested in Prasov et al. [11], and despite using images of

different objects in their work, Griffin and Bock [5] always had them positioned in the same way on the screen and the users had to say the exact same statement about these images. Zhang et al [16] also had the users speak a prescribed type of utterance in each trial condition.

While these previous studies focused primarily on measuring the gaze, others have tried to incorporate it as an input. Tan et al. [15] conducted a form-filling experiment wherein they compared a variety of input modalities. In this case, eye gaze was used to select which field was to be filled in; however, they did not measure success in selecting these fields but rather the users' satisfaction with using a variety of modalities. It does stand to be noted, however, that they found that users enjoyed using the gaze and speech input combination more than any other – which is a motivation to continue research in this area.

Misu et al. [10] have come the closest to building a successful system utilizing gaze as input. However, they deployed this in the automobile domain where the users were querying the system about landmarks as they drove past them. They had to approximate eye gaze with face pose and they also found that they were unable to actually capture accurate face pose data due to the lighting, nor to calculate useful features from what they *were* able to capture due to the fast-moving nature of the vehicle.

In the body of work on using eye gaze as a mouse or as a way to explicitly select objects or fields on the screen, we find work that uses eye gaze to control the screen through panning and zooming or alternatively, through object selection [9]. In the latter case, Qvarfordt and colleagues set up a map-based task that used eye gaze as a trigger for activating the information display [12], [13]. They viewed gaze as a rich modality that can help model the user's attention and focus [8], [13]. We follow in this vein – after all, a user must look at something on the screen before they can react to it. However, we are more interested in using eye gaze to supplement spoken commands in order to improve the spoken language understanding component of the system, rather than to control it. We attempt to do so by using gaze features in conjunction with lexical features to determine what objects or links the user wishes to select on the screen.

This previous work informed our feature selection process. We extend our experiments beyond the limited domains that gaze has traditionally been measured in, expanding to a series of different web-based interaction tasks that people might more naturally partake in.

3. DATA

We collected our data by having 27 users complete 8 different web-based tasks that involved issuing spoken commands to a large-scale monitor. The tasks included activities such as buying a pair of shoes online and registering a boat at the DMV website (as described in Table 1). In general, each task consisted of several different components, which included browsing, object selection and form filling. Figure 1 shows the experimental set-up for the wizard of Oz data collection paradigm. A wizard had access to both the user's speech and could view where the user was looking in real time as captured by eye tracking hardware and overlaid on the contents of the screen, which allowed the wizard to perform all of the necessary actions. We recorded the speech commands, along with eye gaze fixation data and screen contents. We transcribed the inputs and aligned them with the click actions on the screen.

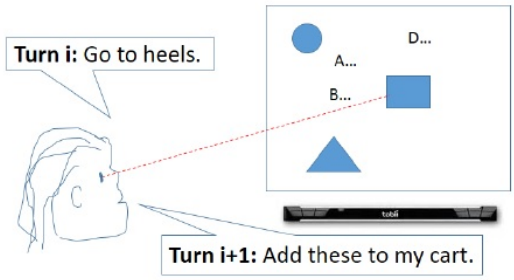

Figure 1. The experimental set-up.

Task	*Task Description*
1	Buy a pair of shoes online
2	Find a sushi restaurant
3	Write a review for a restaurant
4	Buy movie tickets online
5	Contact/book caterer for a wedding
6	Look for movie ratings on IMDB
7	Register a boat at the DMV website
8	Buy flight tickets

Table 1. Descriptions of the tasks.

4. EXPERIMENTS

We asked a machine-learning system to classify each link box on the page in a binary fashion as being the one that the user was referring to or not. To perform the classification, we used *icsiboost*, an implementation of a learning approach that creates a weighted sum of the values of a number of weak classifiers (in this case, one-level decision trees, or 'stumps') [3]. In each case, we held out one user's data and trained the classifier on the rest; we used the held out user's data for testing.

4.1. Features

We used the following lexical features in our experiments.

- *L1*: cosine similarity between term vectors of the link text and the user's utterance

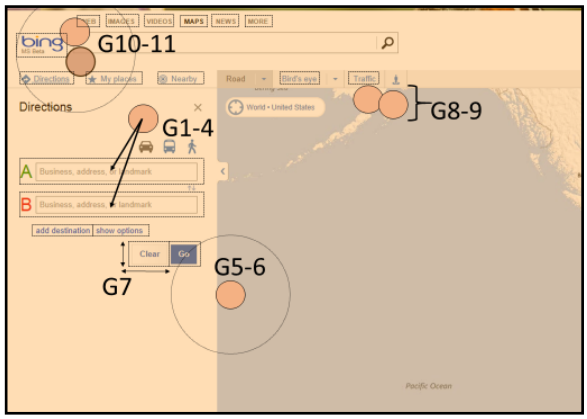

Figure 2. The gaze features. Each circle represents a user's fixation point on the page. This screen is similar to what the wizard saw during the data collection process.

- *L2*: number of characters in the longest common subsequence of the link text and the user's utterance
- *L3* and *L4*: a binary feature that indicates if the link text was included in the user's utterance or not, and if so, the length of the link text

See our original work for details on the computation of the lexical features [6].

We used four gaze features in our original work:

- *G1*: minimum distance to the link box at utterance start
- *G2*: minimum distance to the link box at utterance end
- *G3*: minimum distance to the link box during the time between the utterance start and the utterance end
- *G4*: minimum distance to the link box from any gaze fixation point during the 2 second window before the user's utterance starts

Our previous work [6] was a preliminary study that investigated if eye gaze features could help, and as such, only examined a few gaze features. The finding was that gaze features complement lexical features, but fall short when used in isolation. In this current work, we are focusing on improving them. As such, we explored a variety of new gaze features that might affect where the user's gaze would fall on the screen:

- *G5*: whether the link was within a certain radius of the fixation point at -1, -2, -3 s before utterance start (3cm)
- *G6*: whether the link was within a certain radius of the fixation point at the start of the utterance (3cm)
- *G7*: size of the link box
- *G8*: how frequently the user looked at the link box during the utterance
- *G9*: for how long in total the user looked at the link box during the utterance
- *G10*: how frequently the link box was within a certain radius (3cm) of the fixation point during the utterance
- *G11*: for how long in total the link box was within a radius (3cm) of any fixation point during the utterance

Figure 2 shows the computation of the gaze features. Our new features attempt to take into account attentional features, both in terms of the size of the link box (as bigger boxes might be deemed more important by the web page designer and are more likely to be looked at) and in terms of fixation frequency (with the theory being that users would be more likely to look back at the link they desired than at a link they didn't want to select).

5. RESULTS

The results were computed by the classifier on a turn-by-turn basis. Turn level f-measure was calculated and then averaged out over all of the turns for each experiment. The results with the original gaze features were promising and we used them as the baseline for this current experiment. Table 2 shows the precision, recall and f-score for the original gaze features from [6]; the combination of features G3 and G4 only; the original gaze features alongside all of the new gaze features; and the combination of the top 3 gaze features overall. Row 2 of Table 2 makes clear that features G3 and G4 were responsible for the performance of the original gaze features. More importantly, the new gaze features (G5-11) increase the f-score by 12.7% absolute.

	Precision	Recall	F-score
Orig. Gaze [6]	0.236	0.444	0.256
G3 and G4	0.236	0.477	0.263
Orig. + new gaze	0.374	0.422	0.388
Top 3	0.365	0.413	0.380

Table 2. Results based on gaze-based features only.

	Precision	Recall	F-score
Lex. [6]	0.704	0.520	0.557
Lex + Orig. [6]	0.640	0.707	0.656
Lex+ G3+G4	0.690	0.626	0.641
Lex+Orig.+new gaze	0.714	0.731	0.719
Lex+Top 3	0.707	0.723	0.712

Table 3. Results based on gaze and lexical features.

Table 3 shows the results of adding the gaze features to the lexical-feature baseline (in row 1). Most importantly, it shows that the new gaze features do significantly improve upon our baseline results. That is, the addition of eye gaze features provides great benefit to a lexically based system and leads to an improvement of 16% absolute in f-score (0.719 with all of the features combined versus 0.557 with only lexical features).

We achieved the best results with a combination of all of the old features and all of the new features. But the top 3 features (G3: minimum distance to link box during the utterance, G4: minimum distance to link box 2s before the utterance until the utterance start, and G7: link box size) performed almost as well as the rest of the features combined. These top 3 features were selected using leave-one-out analysis and were found to contribute the most to the final performance gains. Two of these top features were those mentioned previously (G3 and G4), and they accounted for most of the performance in the original results. The link box size feature seems to be successful due to the fact that in browsing tasks, users are most likely to look at (and eventually choose) the most salient items, i.e. those that are bigger. In future work, we hope to investigate other display-based features that deal with saliency.

6. DISCUSSION

One issue we have not addressed heretofore is the timing relationship between eye gaze and the user's speech. Having access to the coordinates of all of the fixation points on a given screen, the problem was to specify to the classifier which fixation points to attend to. We selected a 2s time window for the features looking at the minimum distance from a fixation point to a link box because of previous work in speech planning [14].

Previous research has put forth numbers such as 900ms before utterance onset as the time the users fixate on a subject before they start talking about it [5]. However, this was in tasks where the user was told to produce sentences about the specific objects pictured. It is expected that given a browsing task where the users have to choose between objects, the users may take more time before starting an utterance. Kaur et al. [8] on the other hand found that users in their study fixated on a object a mere 630ms before starting an utterance mentioning that object (with a range of 150ms-1200ms) but once again, the task was to specifically focus on the highlighted object and say a prescribed phrase. In free-form tasks such as the ones found in our data, there is much more freedom for the user to take their time. As such, we wanted to conduct an analysis on when the users were looking at the link that was eventually selected.

We found that in fact the time between when the users fixate on the desired link and the start of their utterance user

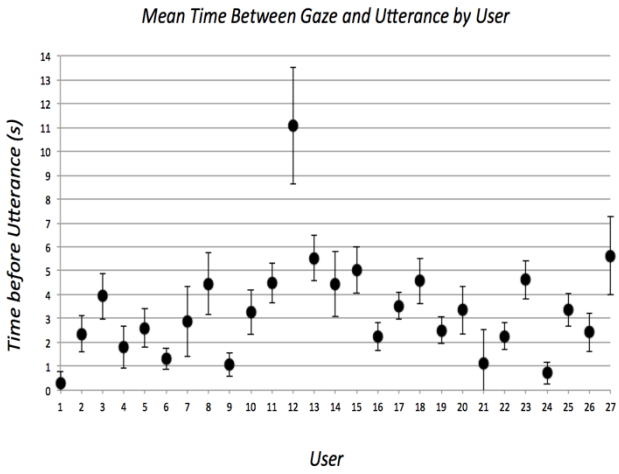

Figure 3. Mean duration between closest gaze fixation point and the desired link box subdivided by user.

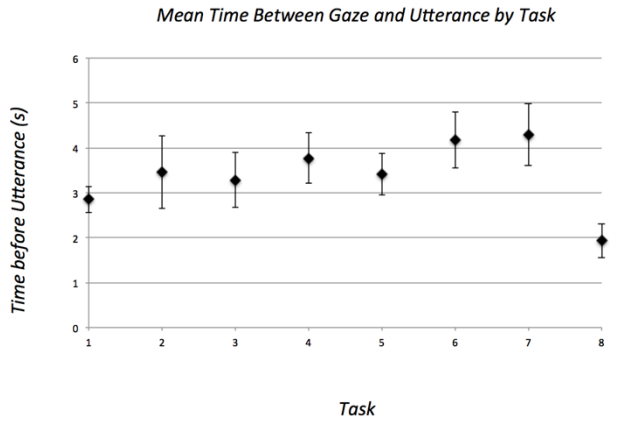

Figure 4. Mean duration between closest gaze fixation point and the desired link box subdivided by task.

varied greatly between users (see Figure 3)[1]. However, the mean fixation on a link happened on average 3.4s before the started speaking (with a standard deviation of 5.5s). This is a much bigger gap between fixation and utterance than has been found in other work; however, we believe that the free-form browsing nature of our tasks explains this difference.

In future work, we wish to explore user-dependent modeling as that may help us calibrate the gaze features and obtain more accurate results. We also found that the type of task tended to affect the gaze data as well. Although most of the differences were not statistically significant, Figure 4 shows that task 8 (buying flight tickets online) involved users fixating on a link very shortly before they spoke to

[1] User 23 seems to be an outlier – having examined the data, we found that this particular individual tended to go off on tangents during the task, which meant that they took longer overall to complete as well.

select it, as compared to the other tasks. This might have to do with the users' familiarity with this type of task (one would expect more people to have bought flight tickets than to have registered a boat at the DMV website, for example), or the fact that it was a form-filling task (wherein the fields were close to each other and the sequence in which the user fills them is dictated by their location on the screen) rather than browsing tasks (where the user has more choice about what to select and where to look).

7. CONCLUSION

Eye gaze can provide important and useful input to spoken language understanding systems. Our experiment shows that a small number of gaze features can produce great performance gains, and more specifically, the inclusion of attentional features in addition to distance-based gaze features leads to improvements of 16% absolute change in f-score over a simple lexical feature baseline. Future work should investigate user-adaptation, since our analysis has shown that different users have different eye gaze fixation patterns with respect to their speech. It would also be interesting to compare the eye gaze fixation patterns in different tasks in a more systematic manner. Another avenue of exploration would be to expand the lexical features beyond the four listed here and also to incorporate the type of phrase (explicit mention of link versus an implicit 'Click that') into the feature set.

8. ACKNOWLEDGEMENTS

We would like to thank Rahul Rajan for his help in collecting the data that we used in this experiment. We are also grateful for discussion we had with both Asli Celikyilmaz and Larry Heck.

9. REFERENCES

[1] Bolt, Richard A. ""Put-that-there": Voice and gesture at the graphics interface." Vol. 14, no. 3. ACM, 1980.

[2] Celikyilmaz, Asli, Zhaleh Feizollahi, Dilek Hakkani-Tür, and Ruhi Sarikaya. "Resolving Referring Expressions in Conversational Dialogs for Natural User Interfaces," EMNLP 2014, Doha, Qatar, October 2014.

[3] Favre, B., D. Hakkani-Tür, and S. Cuendet, "Icsiboost," http://code.google.come/p/icsiboost, 2007.

[4] Griffin, Zenzi M. "Gaze durations during speech reflect word selection and phonological encoding." *Cognition* 82, no. 1: B1-B14, 2001.

[5] Griffin, Zenzi M., and Kathryn Bock. "What the eyes say about speaking." *Psychological science* 11, no. 4: 274-279, 2000.

[6] Hakkani-Tür, Dilek, Asli Celikyilmaz, Malcolm Slaney and Larry Heck. "Eye gaze for spoken language understanding in multi-modal conversational interactions." ICMI 2014.

[7] Heck, Larry P., Dilek Hakkani-Tür, Madhu Chinthakunta, Gökhan Tür, Rukmini Iyer, Partha Parthasarathy, Lisa Stifelman, Elizabeth Shriberg, and Ashley Fidler. "Multi-Modal Conversational Search and Browse." In *SLAM@ INTERSPEECH*, pp. 96-101. 2013.

[8] Kaur, Manpreet, Marilyn Tremaine, Ning Huang, Joseph Wilder, Zoran Gacovski, Frans Flippo, and Chandra Sekhar Mantravadi. "Where is it? Event synchronization in gaze-speech input systems." In *Proceedings of the 5th international conference on Multimodal interfaces*, pp. 151-158. ACM, 2003.

[9] Latif, Hemin Omer, Nasser Sherkat, and Ahmad Lotfi. "Teleoperation through eye gaze (TeleGaze): a multimodal approach." In *Robotics and Biomimetics (ROBIO), 2009 IEEE International Conference on*, pp. 711-716. IEEE, 2009.

[10] Misu, Teruhisa, Antoine Raux, Ian Lane, Joan Devassy, and Rakesh Gupta. "Situated multi-modal dialog system in vehicles." In *Proceedings of the 6th workshop on Eye gaze in intelligent human machine interaction: gaze in multimodal interaction*, pp. 25-28. ACM, 2013.

[11] Prasov, Zahar, Joyce Y. Chai, and Hogyeong Jeong. "Eye Gaze for Attention Prediction in Multimodal Human-Machine Conversation." In *AAAI Spring Symposium: Interaction Challenges for Intelligent Assistants*, pp. 102-110. 2007.

[12] Qvarfordt, Pernilla, and Shumin Zhai. "Conversing with the user based on eye-gaze patterns." In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 221-230. ACM, 2005.

[13] Qvarfordt, Pernilla, David Beymer, and Shumin Zhai. "Realtourist–a study of augmenting human-human and human-computer dialogue with eye-gaze overlay." In *Human-Computer Interaction-INTERACT 2005*, pp. 767-780. Springer Berlin Heidelberg, 2005.

[14] Slaney, Malcolm, Rahul Rajan, Andreas Stolcke, and Partha Parthasarathy. "Gaze-enhanced speech recognition." In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 3236-3240. IEEE, 2014.

[15] Tan, Yeow Kee, Nasser Sherkat, and Tony Allen. "Eye gaze and speech for data entry: a comparison of different

data entry methods." In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 1, pp. I-41. IEEE, 2003.

[16] Zhang, Qiaohui, K. Go, A. Imamiya, and Xiaoyang Mao. "Designing a robust speech and gaze multimodal system for diverse users." In *Information Reuse and Integration, 2003. IRI 2003. IEEE International Conference on*, pp. 354-361. IEEE, 2003.