PROBABILISTIC FEATURES FOR CONNECTING EYE GAZE TO SPOKEN LANGUAGE UNDERSTANDING

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

Microsoft Research, Mountain View, CA
a.prokofieva2@gmail.com, malcolm@ieee.org, dilek@ieee.org,

ABSTRACT

Many users obtain content from a screen and want to make requests of a system based on items that they have seen. Eye-gaze information is a valuable signal in speech recognition and spoken-language understanding (SLU) because it provides context for a user's next utterance—what the user says next is probably conditioned on what they have seen. This paper investigates three types of features for connecting eye-gaze information to an SLU system: lexical, and two types of eye-gaze features. These features help us to understand which object (i.e. a link) that a user is referring to on a screen. We show a 17% absolute performance improvement in the referenced-object F-score by adding eye-gaze features to conventional methods based on a lexical comparison of the spoken utterance and the text on the screen.

Index Terms— Spoken language understanding, referring expression resolution, eye gaze, heat maps, classification.

1. INTRODUCTION

Many of the scenarios in which we are interested combine a screen with voice input. Screens of all sizes are great ways to present a lot of information to a user. And speech is a natural and high-bandwidth input signal. Yet speech recognition remains a challenging problem, especially in the natural (noisy) environments where we often want to communicate with our devices. Thus, attention is the key. What we are attending to is probably a good clue about what we might say next. Were you just looking at the Italian restaurant listing, or the Indian? We use the eyes as an important cue for better recognizing and understanding speech.

A multimodal conversational interactive system depends on automatic speech recognition (ASR) and spoken language understanding (SLU), both of which are ambiguous and prone to error. This paper does not address the ASR issues, but improves the connection between what was recognized and what the user's intentions are, based on what he/she is viewing. Better gaze-directed SLU will help make up for imperfect ASR performance.

[1]Now at Columbia University, New York, NY
[2]Now at Google Research, Mountain View, CA

In this paper we describe an improved way to characterize eye-gaze information, using a probabilistic model characterized by a heat map, in order to select information on a screen. A similar approach, but with less accuracy due to the inherent ambiguities, also applies to face-pose information [1]. Eye-gaze information gives a significant improvement in spoken-language understanding (SLU).

2. PREVIOUS WORK

Most previous work on using information from user's eye gaze to understand and interpret what they are saying focused on tasks that involve passive question-answering based on the contents of the display or performing narrowly-defined tasks that might not generalize to general human-computer interaction. One group of researchers has focused on measuring gaze in a set of very specific object-selection tasks for situated understanding [2, 3, 4, 5, 6], however in many of these the domains and scenarios are very limited. Another group has attempted to use gaze as an indicator of the user's interest or attention, and tried to incorporate gaze information as an input for form filling [7] and referring expression resolution [8, 9]. Misu et al. [10] have come the closest to building a successful speech 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. The third group of researchers has delved into using gaze as a way to manipulate the screen (in an eye-gaze-as-mouse scenario) [7].

Previous work on heat maps has been aimed at developing new visualization techniques [11], conducting qualitative assessments in web search [12] or developing new interaction techniques [13]. We on the other hand hope to use the heat map as another input feature that probabilistically reflects what the user may have seen on screen. In web-browsing tasks such as the ones found in our data, we assume that the space of possible utterances from the user is constrained by what the task is and what is on the screen. By modeling

the probability of seeing any particular link text on the screen with a heat map, we therefore hope to predict what the user has said—that is, to which link box their utterance is referring.

Our work is novel in comparison to these previous studies, in that we investigate the control of free-form web-browsing tasks, where the screen contents change as users browse, with eye-gaze data and spoken-languate understanding. We introduce new, probabilistically motivated eye gaze features, as a novel contribution to our previous work [14, 15, 16, 17] and investigate their contribution to other lexical and gaze-related features.

3. DATA

We collected our data by having 27 users complete 8 different web-based tasks that involved issuing spoken commands about textual content shown on 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.

Before each session, subjects were presented a task description and asked to perform the task naturally, using multiple modalities. We captured user's real-time eye-gaze data using a Tobii REX. When performing the tasks, users were seated at slightly more than arm's distance from a 24-inch display. We used the standard Tobii calibration process. This system provides eye-gaze information at approximately 30 Hz.

A wizard had access to both the user's speech and could view where the user was looking in real time as captured by the eye-tracking hardware and overlaid on the contents of the screen as a shaded circle. This setup allowed the wizard to perform all of the necessary actions to satisfy the user's request, such as clicking on a link, or filling in web forms. At each turn, we recorded the user's spoken utterances along with eye-gaze fixation data, the list of candidate links on display, as well as the complete contents of the web page. We also recorded each wizard action, time synched with the user's actions. We transcribed the speech inputs and aligned them with the click actions on the screen to create our experimental data.

4. EXPERIMENTS

In this work we look at three types of features to help us connect eye gaze to user's intent: lexical, eye-gaze heuristics, and eye-gaze heat maps. This section also describes our classifier approach and our evaluation method.

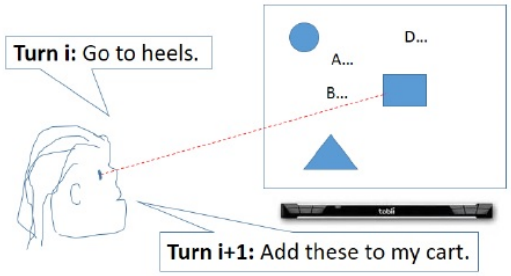

Fig. 1. The experimental set-up.

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 user tasks.

4.1. Lexical Features

To capture references to screen contents using user's spoken input, we extract a set of features that compute lexical similarity between the text associated with each candidate link, $l_k(t)$, and the user utterance, $s(t)$ (referred to as lexical features). These features include cosine similarity between term vectors of $l_k(t)$ and $s(t)$, number of characters in the longest common subsequence of $l_k(t)$ and $s(t)$, and a binary feature that indicates if the link text was included in user's utterance or not, and if so, the length of the link text. We compute similarity features both in terms of word and character counts for robustness to possible speech recognition and tokenization errors.

4.2. Heuristic Gaze Features

In our previous work [15, 16], to investigate the contribution of a user's eye-gaze information for the spoken link-detection task, we computed a set of features using the user's eye-gaze fixation points. These included eye-gaze features that represent the distance from the bounding box of the candidate link to the fixation point at the beginning and end of the utterance, to the distance to the closest fixation point during the utterance and during the 2 second window before the user's utterance starts, the size of the link's bounding box, how frequently the user looked at the link box, etc. More details about the set of features can be found in the previous publications.

4.3. Heat map Features

A more elegant solution is to represent gaze fixation points on the screen as a heat map, approximating the probability that a given point on the screen was seen. We compare this probabilistic approach to our previous heuristic approach, where we used various distance-based and attentional features. Our hope was that a more elegant probabilistic approach would do as well, or better, than a heuristic approach.

Our eyes alternate between ballistic movements known as saccades, and short stationary times known as fixations. During the times when the eyes are fixated we can read text, and then the eyes saccade to another point on the screen to absorb some new information. Given information about the time and (screen) location of each fixation point, we want to estimate whether an item on the screen is the likely target of a speech request.

A heat map is a probabilistic representation of where on the screen a user looks [11]. Even with perfect data about where the fovea is pointed, there is a small region around this location where a user can read and understand text. The size of this region depends on the user's needs (skimming versus concentrating), the accuracy of the sensor, and the characteristics of the text on the screen (size and font).

We hypothesize the overall probability that a word has been read and processed by a subject is a function of the distance from the word to each fixation point, the time spent fixated at each point, and a scalar parameter that characterizes the size of the Gaussian associated with each fixation point. Thus around each fixation point, there is a small region where it is possible that the user has read words on the screen. We model this region with a two-dimensional Gaussian probability "bump," which is weighted by the duration of the fixation. Thus, the probability, p, that a user sees a pixel at x, y is equal to

$$p(x, y, t) = \sum_i N(|x_i - x, y_i - y|, \sigma) d(i) w(t - t_i) \quad (1)$$

where x_i, y_i is the position of fixation point i, $d(i)$ is the duration of the fixation at time t_i, $w()$ is a temporal weighting function, and $N(r, \sigma)$ is a normal probability function centered at a radius r of zero, and with a standard deviation of σ,

Finally, in this work we are interested in whether a word might lead to a speech utterance. This is going to be a function of the amount of time *before* the user starts speaking. In this work we hypothesize a linear decay function, so $w(t)$ in the expression above is equal to

$$w(t) = \begin{cases} (T_d - t)/T_d & \text{if } 0 \geq t \geq T_d \\ 0 & \text{elsewhere.} \end{cases} \quad (2)$$

Since we are interested in gaze fixation patterns over time, we took three time windows (at 2s, 4s, and 8s before the utterance start) and applied the linear weighting function (2 to the Gaussian values to account for the decreasing importance of a fixation point the further away from the utterance start it is. Figure 2 shows the time slices for a set of fixation points as represented by a heat map.

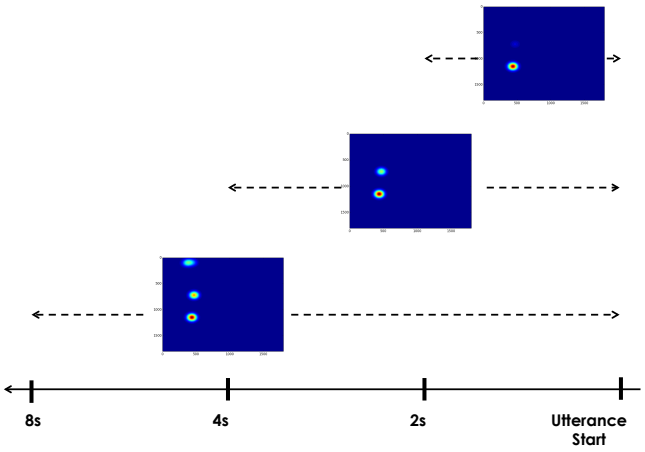

Fig. 2. We captured a user's eye gaze using heat maps calculated over three different windows before the start of each utterance. These three plots show the evolution of the heat map over time. Each heatmap includes all the gaze fixation points recorded in that time window, with those occurring closest to the start of the utterance being more heavily weighed. Words included by wider (temporal) windows are more likely to include the referenced object, but also lead to higher false-alarm rates.

Given the heat map, we still need to evaluate the probability that a link was seen by summarizing the values of the heat map within the bounding box of the link on the screen. We did this by both finding the average value of the heat map within the bounding box, as well as the maximum. We also found it useful to include a feature based on the absolute size of the link's bounding box. This seems to capture how likely the link is to capture the user's attention, perhaps similar to a prior probablity.

4.4. Classifier Approach

We frame the resolution of referring expressions as a binary classification task (i.e. we have two classes, positive and negative). For each link l_i^t displayed on the page to the user at turn t, we compute a set of features f_i^t and estimate $P(positive|f_i^t)$. We use icsiboost [18] for classification. During training, only the links referred by the user are assigned the positive class, and all the rest of the candidate links are assigned the negative class. At runtime, we find the links l_j^k at turn k, that have $P(positive|f_j^k) > \theta$, where θ is the posterior probability threshold, and return them as estimated intended links.

	Recall	Precision	F-score
lexical [15]	70.4%	52.0%	55.7%
gaze heuristics [15]	47.7%	23.5%	26.3%
lexical+gaze [15]	70.7%	63.6%	65.2%
lexical+extended gaze [16]	73.1%	71.4%	71.9%
heat map	42.2%	37.4%	38.8%
heat map+lexical+gaze [16]	74.0%	72.1%	72.7%

Table 2. Comparing lexical and heat map features, including references to previous work.

4.5. Method of evaluation

For evaluation, we compare the links chosen according to the posterior probabilities with the links intended by the user, and then compute turn-level F-measure, which is the harmonic mean of recall and precision. Recall checks if the intended link was returned amongst the estimated intended links, and precision checks the ratio of correctly returned links. In our experiments, icsiboost was run for 100 iterations, and the posterior probability threshold was chosen to be 0.5.

5. RESULTS

As can be seen from Table 2, the combination of the heat map features with lexical features significantly improves the performance. This comes specifically from a gain in precision, as the gaze feature helps to narrow down the field of potential link candidates to which the user could be referring.

If we compare the performance of the heuristic gaze features and the heat map, we find that the heat map features improve the overall classifier performance by a greater amount, leading to an 17% absolute change in f-score over the use of lexical features alone (from 55.7% to 72.7%).

6. DISCUSSION

When using a heat map, there are both temporal and spatial parameters that we can vary. For one, we can vary the size of the Gaussian. We hypothosized that a radius of 65 pixels on the screen would capture the necessary words, based on our estimates of the sensor noise and the fovea size. We compared those results to experiments where the Gaussian was $1/2$ and 2 times that size, and found that our original choice of radius gave us a better performance on our SLU task. Fitting a curve through those results suggest that the optimal Gaussian radius is actually 61.6 pixels (see Figure 3.)

It is also possible to explore a variety of time horizons. We used all three horizons (2, 4, and 8s) in this work and let the classifier pick the best. We didn't see much difference in our SLU performance when computing only one of them.

One last parameter to vary is the shape of the weighting function. A linear weighting function is the easiest to implement. However, data analysis (carried out in our previous work) showed that the elapsed time between fixation on the eventually selected link and the beginning of the user utterance was on average 3.4s, with a standard deviation of 5.7s and a distribution shown in Figure 4. If we were to fit a curve to this distribution and use it as a weighting function, thus more heavily weighting fixations that occur near the average fixation time, we would expect to see improved results.

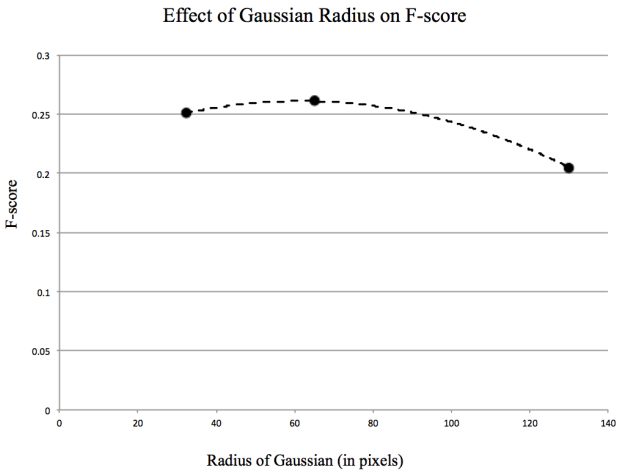

Fig. 3. SLU performance as a function of Gaussian radius when computing the heat map.

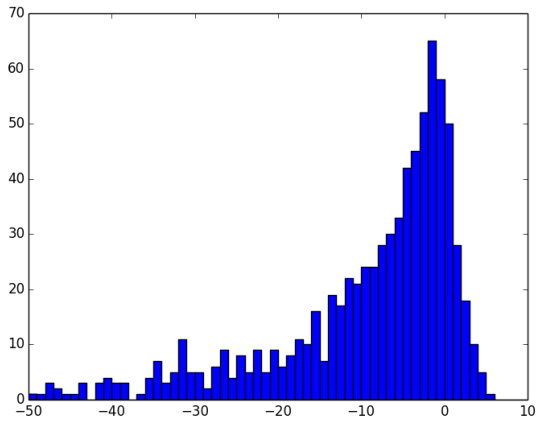

Fig. 4. Frequency of time intervals between eye gaze and referring utterances.

7. 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.

8. REFERENCES

[1] Malcolm Slaney, Andreas Stolcke, and Dilek Hakkani-Tür, "The relation of eye gaze and face pose: Potential impact on speech recognition," in *ACM International Conference on Multimodal Interactions (ICMI)*, November 2014.

[2] Z. Prasov, J.Y. Chai, and H. Jeong, "Eye gaze for attention prediction in multimodal human-machine conversation," in *AAAI Spring Symposium: Interaction Challenges for Intelligent Assistants*, 2007, pp. 102–110.

[3] Z.M. Griffin, "Gaze durations during speech reflect word selection and phonological encoding," in *Cognition 82*, 2001, pp. B1–B14.

[4] Z.M. Griffin and K. Bock, "What the eyes say about speaking," in *Psychological science 11,*, 2000, number 4, pp. 274–279.

[5] R.A. Bolt, "Put-that-there: Voice and gesture at the graphics interface," in *Proceedings of SIGGraph 1980*, 1980, vol. 14, pp. 262–270.

[6] Q. Zhang, K. Go, A. Imamiya, and X. Mao, "Designing a robust speech and gaze multimodal system for diverse users," in *IEEE International Conference on Information Reuse and Integration (IRI)*, 2003, pp. 354–361.

[7] Y.K. Tan, N. Sherkat, and T. Allen, "Eye gaze and speech for data entry: a comparison of different data entry methods," in *Proceedings of International Conference on Multimedia and Expo ICME'03*. 2003, vol. 1, pp. I–41, IEEE.

[8] Z. Prasov and J.Y. Chai, "What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces," in *ACM Proceedings of the 13th international conference on Intelligent user interfaces*, 2008, pp. 20–29.

[9] C. Kennington, S. Kousidis, and D. Schlangen, "Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information," in *Proceedings of SigDial*, 2013.

[10] T. Misu, A. Raux, I. Lane, J. Devassy, and R. Gupta, "Situated multi-modal dialog system in vehicles," in *Proceedings of the 6th ACM workshop on Eye gaze in intelligent human machine interaction*, 2013, pp. 25–28.

[11] O. Špakov and D. Miniotas, "Visualization of eye gaze data using heat maps," in *Electronics and electrical engineering*, 2007, vol. 2, pp. 55–58.

[12] Susan T. Dumais, Georg Buscher, and Edward Cutrell, "Individual differences in gaze patterns for web search," in *Proceedings of the Third Symposium on Information Interaction in Context*, New York, NY, USA, 2010, IIiX '10, pp. 185–194, ACM.

[13] Sophie Stellmach, Sebastian Stober, Andreas Nürnberger, and Raimund Dachselt, "Designing gaze-supported multimodal interactions for the exploration of large image collections," in *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications*, New York, NY, USA, 2011, NGCA '11, pp. 1:1–1:8, ACM.

[14] L. Heck, D. Hakkani-Tür, M. Chinthakunta, G. Tur, R. Iyer, P. Parthasarathy, L. Stifelman, A. Fidler, and E. Shriberg, "Multimodal conversational search and browse," in *Proceedings of IEEE Workshop on Speech, Language and Audio in Multimedia*, 2013.

[15] D. Hakkani-Tür, M. Slaney, A. Celikyilmaz, and L. Heck, "Eye gaze for spoken language understanding in multi-modal conversational interactions," in *Proceedings of ACM ICMI*, 2014.

[16] A. Prokofieva, D. Hakkani-Tür, and M. Slaney, "Eye gaze for spoken language understanding in multi-modal conversational interactions," in *Proceedings of IEEE SLT Workshop*, 2014.

[17] A. Celikyilmaz, Z. Feizollahi, D. Hakkani-Tür, and R. Sarikaya, "Resolving referring expressions in conversational dialogs for natural user interfaces," in *Proceedings of EMNLP*, 2014.

[18] Benoit Favre, Dilek Hakkani-Tür, and Sebastien Cuendet, "Icsiboost," http://code.google.come/p/icsiboost, 2007.