Eye Gaze for Spoken Language Understanding in Multi-Modal Conversational Interactions

Dilek Hakkani-Tür
Microsoft
Mountain View, CA, USA
dilek@ieee.org

Malcolm Slaney
Microsoft
Mountain View, CA, USA
malcolm@ieee.org

Asli Celikyilmaz
Microsoft
Mountain View, CA, USA
asli@ieee.org

Larry Heck
Microsoft
Mountain View, CA, USA
larry.heck@ieee.org

ABSTRACT

When humans converse with each other, they naturally amalgamate information from multiple modalities (i.e., speech, gestures, speech prosody, facial expressions, and eye gaze). This paper focuses on eye gaze and its combination with speech. We develop a model that resolves references to visual (screen) elements in a conversational web browsing system. The system detects eye gaze, recognizes speech, and then interprets the user's browsing intent (e.g., click on a specific element) through a combination of spoken language understanding and eye gaze tracking. We experiment with multi-turn interactions collected in a wizard-of-Oz scenario where users are asked to perform several web-browsing tasks. We compare several gaze features and evaluate their effectiveness when combined with speech-based lexical features. The resulting multi-modal system not only increases user intent (turn) accuracy by 17%, but also resolves the referring expression ambiguity commonly observed in dialog systems with a 10% increase in F-measure.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Language Parsing and Understanding—*speech and gaze understanding*

Keywords

spoken language understanding, reference resolution, eye gaze

1. INTRODUCTION

Humans naturally use information from multiple modalities when interacting with each other, motivating the integration of these signals when interpreting user requests to a conversational-interaction system. In this work, we focus on the use of two modalities for spoken language understanding: voice and eye gaze. Users see information on a display and use their voice to control a conversational interaction, such as finding information about a topic or performing a transaction.

In such systems, there are two types of conversational turns: generic and referential. A referential turn refers to an item on display, while a generic turn does not. First we must detect that the user has a generic request, independent of the display contents. For example, the first turn of an interaction (such as a user asking *show me movies nearby*) is usually not referring to an item on display. In the second type of user turn the user's utterance refer to an item on display—understanding involves resolution of that item. In this paper, we focus on this second step. We wish to identify an item to which the user is referring in a spoken utterance, using the location of their eye gaze on a display.

In our previous work on using multi-modal information for spoken-language understanding, we integrated hand-pointing gestures with spoken utterances for click intent detection [7]. Furthermore, in simulated experiments, we analyzed the accuracy of click intent detection with respect to the distance between the targeted item and the pointing position. We also performed a detailed analysis of lexical features for this task for a variety of display designs, interaction domains and user devices [2]. Other related studies include work by Misu *et al.* [9], who proposed an in-car spoken dialog system that integrates multi-modal inputs of speech, geo-location, gaze (as estimated from face direction) and dialog history to answer drivers' queries about their surroundings. Due to the in-car nature of the dialog application, what the user sees changes over time. Cooke *et al.* [3] also investigated integration of eye gaze with spoken utterances, with a focus on dynamic model-based adaptation methods for noise-robust automatic speech recognition (ASR). For evaluation, they use a dataset recorded for a 'put that there' task [1], where the user tells a receiver to position a colored shape on a displayed map. Along similar lines to our work, Prasov and Chai [10] studied contribution of eye gaze for reference resolution. Their work targets situated understanding, where users refer to objects in context, and eye gaze compensates for the lack of domain modeling. Similarly, Kennington *et al.* [8] looked into interpolation of lexical, eye gaze and pointing models for this task, where the semantic space is predefined as the set of puzzle-pieces. A special feature of their modeling framework is incremental understanding, allowing for understanding of user's input incrementally during each user turn.

ICMI'14, November 12–16, 2014, Istanbul, Turkey.
Copyright 2014 ACM 978-1-4503-2885-2/14/11 ...$15.00.
http://dx.doi.org/10.1145/2663204.2663277 .

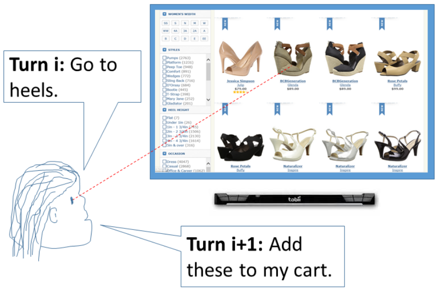

Figure 1: Setting for gaze supported interactions with a conversational system.

Symbol	Desctiption
t	turn id
$s(t)$	sentence at turn t
$l_k(t)$	candidate link k at turn t
$L(t)$	set of links at turn t
$d[\cdot]$	distance between two points on the display
$P_{start(t)}$	fixation point at the beginning of the utterance
$P_{end(t)}$	fixation point at the end of the utterance
$md[\cdot]$	closest fixation point during the utterance
$f(t)$	features extracted at turn t
$\hat{l}(t)$	predicted link at turn t

Table 1: The notation used in the paper.

In this work, to simulate various forms of visual displays and conversational interaction tasks, we chose a conversational search and browse task that allows users to navigate web pages that include various forms of hyperlinks, buttons, text boxes, etc. Figures 1 and 2 show the system setting, user utterances and system actions in a sample interaction. In such a conversational system, users may explicitly refer to items on web pages (i.e., user utterance includes the full or partial text of a link) or make implicit referrals (i.e., *select* or use position of items on the page, *select the top one*). Lexical features that measure the textual similarity between user utterances and displayed items are expected to be useful for resolving explicit referrals, but are not satisfactory for the implicit referrals. Gaze features help resolve both cases, especially with erroneous ASR, where user utterances are not always correctly recognized.

In the following sections, we first describe our approach for modeling the resolution of referring expressions using spoken utterances and user's eye gaze. Then, in Section 3, we describe the experimental setup and the data collection. In the experiments section, we first analyze the collected data in terms of different types of referrals. We then present experiments where we analyze the contribution of lexical and eye gaze features. Results from the experiments where we use both types of features show that both modalities are useful for resolution of referring expressions, and their contribution is complementary.

2. MODELING FOR RESOLUTION OF REFERRING EXPRESSIONS

We frame the resolution of referring expressions as a binary classification problem, where all links on the page displayed to the user $l_k(t) \in L(t)$ $(k = 1, ..., |L(t)|)$, paired with user's utterance $s(t)$ at turn t, and their eye gaze fixation points form the examples to be classified. We consider examples that include the link targeted by the user as positive examples, and all the rest of the candidates are negative examples.

User: I wanna shop for women's shoes at <website>.com
System: [opens up <website>.com, women's shoes page]
User: Go to heels
System: [follows "heels" link]
User: Add these shoes to my cart

Figure 2: An example multi-modal conversational interaction, with a transcription of user utterances, system actions and display contents.

We then 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.

To provide robustness to possible errors in tokenization and ASR, we compute similarity features both at the word and character levels.

To capture the information from user's eye gaze, we first compute fixation points where a user's eye gaze lands, and the start and end time of that fixation. Our eyes process information during short fixation times when the eye is not moving. The eyes reorient during quick ballistic movements known as saccades, but we are not processing information during these times. We need to identify the fixation points to know what has been read. We use an algorithm by Salvucci [4] to identify each fixation point from our eye-gaze data. For each recorded eye-gaze location, we look for a set of points extending over at least 100ms that are clustered together. A cluster is defined by a Manhattan distance of less than 40 pixels. Thus with an average sampling rate of 30Hz, we need at least three points in close proximity to determine that there is a fixation point. In this work, we use the centroid of this cluster as the fixation point, and we suggest that the subject is referring to items in close proximity to this point.

We then extract eye gaze features that represent distance $d[\cdot]$ from the surrounding box of candidate link, $l_k(t)$, to:

- the fixation point at the beginning of the utterance $start(t)$, $P_{start(t)}$: $d[P_{start(t)}, l_k(t)]$
- the fixation point at the end of the utterance $end(t)$, $P_{end(t)}$: $d[P_{end(t)}, l_k(t)]$
- closest eye gaze fixation point during the utterance, $md[start(t), end(t)]$:
  $$md[start(t), end(t)] = min_{x \in (start(t), end(t))} d[P_x, l_k(t)]$$
- closest eye gaze fixation point during the 2 second window before the start of the user's utterance,
  $$md[start(t) - 2, start(t)]$$

Figure 3 depicts the computation of the distance. Note that, during the saccades, the distance to any candidate link is set to a large number.

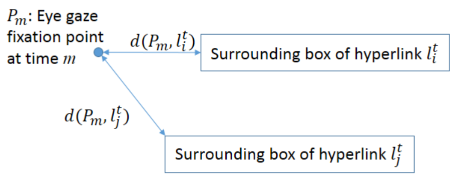
P_m: Eye gaze fixation point at time m
$d(P_m, l_i^t)$
Surrounding box of hyperlink l_i^t
$d(P_m, l_j^t)$
Surrounding box of hyperlink l_j^t

Figure 3: Computation of distance between hyperlinks displayed at turn t and eye gaze fixation points.

During runtime, at each turn t, we parse the displayed pages for the set of candidate links $L(t)$ and select the candidate $l_k(t)$ that has the highest probability of being the positive class example given the user's utterance and the associated set of features $f(t)$:

$$\hat{l}(t) = argmax_{l_k(t) \in L(t)} P(positive|l_k(t), f(t))$$

In experiments, we compare the contribution of each type of feature, as well as their combination. Furthermore, to see the effect of the quality of speech transcription, we experiment with clean (i.e. manual) and noisy (i.e., automatic) speech transcription conditions.

3. DATA COLLECTION AND EXPERIMENT SET UP

We collected real-time eye-gaze data using a Tobii REX. 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.

Before each session, subjects were presented a task description and asked to perform the task naturally, using multiple modalities. The tasks included browsing for information, such as finding a nearby restaurant, as well as transactions, such as buying flight tickets. The wizard heard the spoken utterances of the subjects and was shown the same display contents as the subject. However, on wizard's display, the area where the user's eye gaze was pointing to (as captured by Tobii) was shaded with a circle. The wizard took 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, the list of candidate links on display as well as the complete contents of the web page. We also recorded each wizard action, time synchronized with the user's actions.

4. EXPERIMENTS

4.1 Data Sets

In our experiments we collected data from 27 speakers, each performing 9 tasks. User turns included requests to follow links on display, system commands such as *scroll down*, utterances possibly not addressed to the machine (such as, *'oh', 'no'*), and form filling requests. The tasks result in an average of 12.2 user turns. The data set includes 2,965 turns, 581 of which aim to follow a link on the screen. There are in total 175,113 candidate links on the web pages visited by the users (an average of 301.4 per each click turn).

We use a state-of-the-art large vocabulary ASR system in our experiments [5]. The acoustic models incorporate the latest advances in context-dependent deep neural networks (DNN) for estimating senone likelihoods. The language model (LM) is a general-purpose backoff 4-gram model with a vocabulary of about 400K words. This generic LM (GLM) was trained on a wide variety of sources ranging from transcribed speech from deployed ASR applications, such as voice search, to text from a diverse set of web resources. The GLM was not tailored or adapted to the tasks of our study.

As described earlier, users can refer to items on display in several different ways. Table 2 shows the types of utterances observed in our data set and their relative frequencies.

4.2 Results

We performed 27-fold classification experiments, where we take out the data set of one user at each fold, to use it as the test set, and use all the remaining examples as the training set. We use icsiboost [6] for classification. Table 3 shows the turn accuracy and F-measure results from these experiments. Turn accuracy refers to the percentage of turns where only a single link was returned by the classifier and that was the link targeted by the user. For many examples, two or more links were assigned the same probability by the classifier, but only one was the one targeted by the user. These cases are considered as system errors for turn accuracy. However, reducing the set of candidates from many examples to a few, where the system can ask a clarification question to the user (instead of asking the user to repeat the request) may result in a better conversational interaction experience. Hence, we also report F-measure results, macro-averaged over turns in this table.

As seen in Table 3, lexical features result in higher turn accuracy and F-measure than gaze features alone. This could be explained by the fact that majority of utterances in our data set include exact match or explicit referrals, which can easily be resolved, especially since usually (but not always) there is a single link with the matching text. Also, many pages are densely populated with links, and many links are in close proximity to user's eye-gaze fixations. Hence, while the set of candidates may be reduced by gaze features, it may be difficult to identify the correct one. However, the gaze features are also useful, and the combination of the two types of features results in the best performance in terms of both measures. The turn accuracy (and F-measure) of 42.7% (55.6%) with just lexical features improves to 59.7% (65.6%) when we include gaze features in the experiments.

When erroneous ASR transcriptions are used instead of the clean/manual transcriptions, as will be the case with a real application, the performance with lexical features degrades significantly: turn accuracy drops from 42.7% to 32.9%, while performance with gaze features stays the same. The ASR word error rate on this data set is 44.8%. The performance numbers with the combined set of features is again better than the performance with individual set of features.

4.3 Discussions and Future Work

Our study shows significant connections between gaze features and the variables characterizing the lexical overlap with the targeted link. It is interesting to observe that the target focused gaze indicates the presence of the targeted link nearby. Nevertheless, we observed data and framework related issues that may have interfered with the model's accuracy and that we should investigate more as a follow-up study. These can be summarized as:

- **Text normalization issues**: We observe such issues when the text normalization becomes complex, for in-

Referral Type	Example User Utterance	Example Text Associated with Link/Image/Button	Relative Frequency
Exact Match	*how to register a vessel*	*how to register a vessel*	35.3%
Explicit	*go to sneakers and athletic shoes*	*sneakers and athletic shoes*	34.1%
Explicit Partial	*that to my cart*	*add to cart*	9.6%
Implicit, Position	*select the second one*	*yumeya sushi*	3.9%
Implicit	*select this*	*how to register a vessel*	8.6%
Multiple	*click on yumeya that's number two*	*yumeya sushi*	1.0%
Other	*find vendors*	*search* or no text associated with link	7.5%

Table 2: Examples from conversational interactions.

Experiment	MANUAL		ASR	
	TA	F	TA	F
Only Lexical Features	42.7%	55.6%	32.9%	43.2%
Only Gaze Features	18.6%	25.6%	18.6%	25.6%
All Features	59.7%	65.6%	51.8%	56.3%

Table 3: Turn accuracy (TA) and F-measure with manual and ASR transcriptions.

stance, the link text shows '*7:00 pm*', but the user speaks '*seven o'clock*'. In the future, we plan to use better text normalization as a preprocessing step to interpret the relation between the utterance and link text.

- **Display related issues**: The display screen in our framework is a typical web page, which may introduce additional ambiguities to our model. Specifically, when users refer to an item on a web page they would think that the item is a clickable object, when it is actually not.

The main conclusion we glean from our experiments is that we observe the best performance when we include both types of features. We think that it would be beneficial had we integrated the features from the two modalities, such as, adding a linking binary feature to indicate the closest link candidate with matching text. In a follow-up study, we plan to investigate such features.

Other avenues that we can investigate include:

- joint modeling of automatic detection of referrals to displayed items and resolution of referring expressions.
- improving ASR by using displayed page contents [7] and eye-gaze focused language models [11],
- extending the set of lexical features by using extended set of ASR hypotheses (such as, n-best lists and word lattices) and matching between phone sequences.

5. CONCLUSIONS

We presented a framework that exploits gaze features for spoken language understanding in human-machine dialog systems. Our model interprets the user's intent in a conversational browsing scenario (e.g., clicking on a link) using a combination of gaze and voice. We introduced several new gaze features and evaluated their efficiency together with the lexical features. The features were extracted from the user's eye-gaze pattern, spoken utterances, and the text displayed on the screen. The resulting multi-modal system not only increases user intent (turn) accuracy by 17%, but also resolves referring expression ambiguity commonly observed in dialog systems with a 10% increase in F-measure.

Acknowledgments: Authors would like to thank Rahul Rajen, Gokhan Tur, Zhaleh Feizollahi, and Ruhi Sarikaya for discussions and help with the data collection.

6. REFERENCES

[1] R. A. Bolt. "put-that-there": Voice and gesture at the graphics interface. volume 14. ACM, 1980.

[2] 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.

[3] N. Cooke, A. Shen, and M. Russell. Exploiting a 'gaze-lombard effect' to improve asr performance in acoustically noisy settings. In *Proceedings of IEEE International Conference onAcoustics, Speech and Signal Processing (ICASSP)*, 2014.

[4] J. H. D. D. Salvucci. Identifying fixations and saccades in eye-tracking protocols. In *Eye Tracking Researchand Application*, pages 71–79, 2000.

[5] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, and J. Williams. Recent advances in deep learning for speech research at microsoft. In *Proceedings of IEEE International Conference onAcoustics, Speech and Signal Processing (ICASSP)*, 2013.

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

[7] 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.

[8] 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.

[9] 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*, pages 25–28, 2013.

[10] 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*, pages 20–29, 2008.

[11] M. Slaney, R. Rajen, A. Stolcke, and P. Parthasarathy. Gaze enhanced speech recognition. In *Proceedings of IEEE International Conference onAcoustics, Speech and Signal Processing (ICASSP)*, 2014.