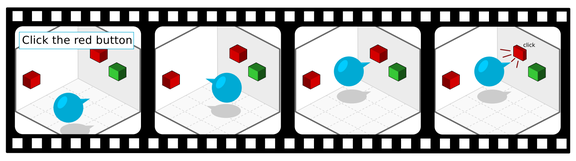
The Impact of Listener Gaze on Predicting Reference Resolution

Nikolina Koleva[1], Martín Villalba[2], Maria Staudte[1], Alexander Koller[2]

[1]*Embodied Spoken Interaction Group, Saarland University (Saarbrücken, Germany)*
[2]*Department of Linguistics, University of Potsdam (Potsdam, Germany)*

Can listener gaze information improve accuracy on RE resolution prediction?
Is listener gaze especially beneficial in cluttered scenes?

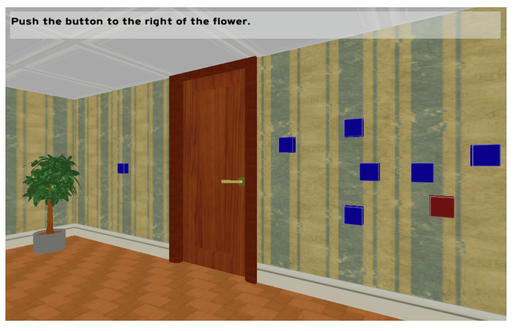
Methodology



① The GIVE Corpus

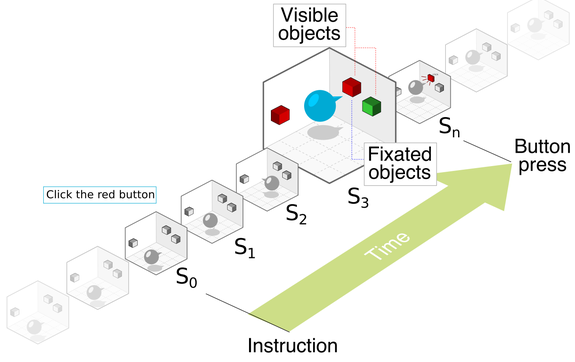
A corpora of natural language generation systems helping users perform a task in a virtual 3D world

Users have to solve a puzzle in a 3D world. They can interact with objects in the world (e.g. click on buttons) and move freely in the environment. NLG systems guide users by generating instructions, including referring expressions for objects in the environment.



Grounding problem: Systems have to predict the (mis) understandings of a referent, and also prevent mistakes by providing corrective feedback.

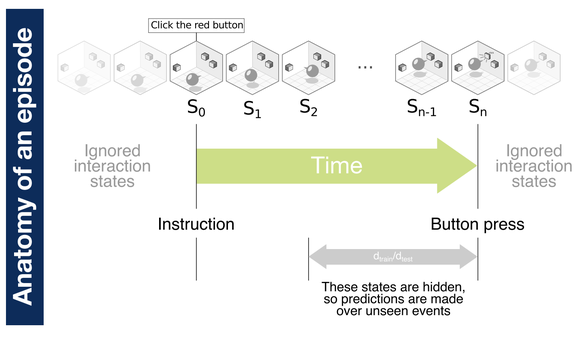
The complete interaction corpora comprises over 2500 games and more than 340 hs. of recorded interactions. 75 games also provide eye-tracking data in 8:06 hs. of interactions.



② Episode segmentation

Extracting behavioral information from recordings of unstructured interactions

Given this corpora of recorded interactions, we define an **episode** as a typically short sequence of recorded behavior states, beginning with a manipulation instruction and ending with a button press by the player, with no further utterances in between. This definition was based on (Koller et al., EMNLP 2013), extending their P_{obs} model with information about fixated objects. This extended model is named P_{Eobs}.



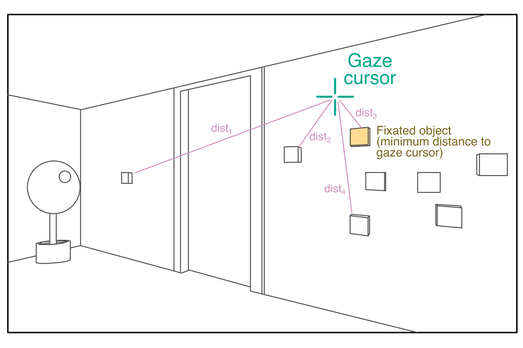
We extracted 761 episodes containing recorded eye-tracking data, amounting to 47 min. 58 sec. of recorded interactions. Average length per episode is 3.78 sec. (σ = 3.03 sec.).

Feature Matrix	Obj_0	Obj_1	Obj_2	Obj_3	...	Obj_n
VisualSalience	0.0	0.3	0.5	0.0	...	0.0
InRoom	0.0	1.0	1.0	0.0	...	0.0
LookedAt	0.0	0.5	0.7	0.0	...	0.0
LongestSequence	0.0	0.3	0.4	0.0	...	0.0
LinearDistance	0.9	0.1	0.1	0.9	...	0.7
InvSquaredDistance	0.0	9.7	10.1	0.0	...	0.0
UpdateFixedObjs	0.0	3.0	3.0	0.0	...	0.0

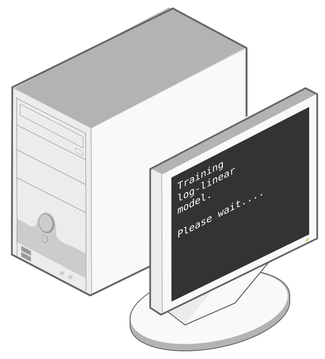
③ Feature extraction

Strategies to analyze eye-tracking information from recorded behavior

When extending the P_{obs} model, we created new feature functions over episodes, to account for the available eye-tracking data. These features are:



Looked at: counts the number of behavior states in which an object has been fixated in the current episode.
Longest Sequence: longest continuous sequence of states in which an object has been fixated.
Linear Distance: euclidean distance *dist* on screen between the gaze cursor and the center of an object.
Inv-Squared Distance: defined as $\frac{1}{1+dist^2}$



④ Model training

Defining a learning strategy to predict user interactions

We trained our models to correctly predict a target button based only on data observed up until $-d_{train}$ seconds before a button press takes place. We call this "*training at time -d_{train}*". Similarly, "*testing at time -d_{test}*" is defined as the percentage of target objects that were correctly predicted up until $-d_{test}$ seconds before a button press.
Then we trained a log-linear model, where weights assigned to each feature function were learned via optimization with the L-BFGS algorithm.

We evaluated our model with different combinations for the parameters (d_{train}, d_{test}). Given the limited amount of eye-tracking data, we applied 10-fold cross-validation. We also removed instances of insuficient length for the given parameters, and classified instances in either *easy* or *hard* based on the number of visible targets (3 or less visible objects for *easy*, or *hard* otherwise).

We trained a new classifier for each combination of the (d_{train}, d_{test}) parameters, in order to evaluate how effective our models are at predicting targets when the interval between the prediction and the action increases.

Evaluation and Results

Our eye-tracking features improve the accuracy of the previous model

We evaluated both the original P_{obs} model along with the P_{Eobs} model on the same data set. We also tested accuracy for each individual feature function, in order to test if a single function could outperform P_{obs}. In adition, we also tested two simpler versions of P_{obs} as baselines.

For each permutation of the training parameters (d_{train}, d_{test}), we obtain a set of episodes with the appropriate length. Given this set, we applied both P_{obs} and P_{Eobs} to all elements and compared their accuracies using a paired samples t-test.

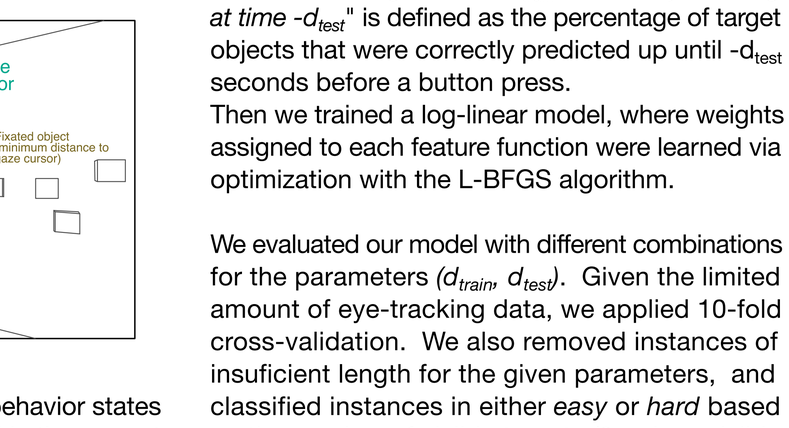
The test results show that P_{Eobs} performance is significantly better than P_{obs} ($p<0.001$). Eye-tracking features seem to be particularly helpful for predicting to which entity a RE is resolved in hard scenes.

Eye-tracking is significantly helpful 3 seconds before the button press

Results also show a peak in accuracy near the -3 sec. mark, which is consistent along both *easy* and *hard* instances. To test whether this peak is significant, we computed a 2x2 contingency table to contrast correct and incorrect predictions for P_{obs} and P_{Eobs} for all episode judgements trained at times in the [-6 sec., -3 sec.] range and tested at -3 seconds.

McNemar's test results showed that the marginal row and column frequencies are significantly different ($p<0.05$). This shows that our model is more accurate precisely at points in time when we expect fixations to a target object.

The question remains whether this effect is exclusive for our domain or whether this is consistent across interactive scenarios.



The accuracy results show our observations for $-6 \leq -d_{train} \leq -2$ and $-d_{train} \leq -d_{test} \leq 0$. The graph shows that P_{Eobs} performs similarly as P_{obs} on the easy instances. However, P_{Eobs} shows a consistent improvement on the hard instances over P_{obs}.

Future work

We have shown that the inclusion of gaze in our model improves over P_{obs} in the context of predicting the resolution of a RE. Gaze is also beneficial in an earlier stage of an interaction.
Our next step is the combination of our P_{Eobs} model with the P_{sem} semantic model of (Koller et al., 2013), in order to test how do our extended features perform in a combined model.

Testing with users in real time is also an area for future research. An implementation of P_{obs} is currently in the test phase, and an extension for P_{Eobs} will follow.
The lack of corpora collections containing eye-tracking data in interactive scenarios presents a limitation for further research. Luckily, several data collection projects are currently underway.

Selected references

Staudte et al. 2012. **Enhancing referential success by tracking hearer gaze**. In *Proceedings of the 13th Annual Meeting of the Special interest Group on Discourse and Dialogue SIGDIAL '12.*

Koller et al. 2013. **Predicting the resolution of referring expressions from user behavior.** In *Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP '13.*

For more information

Nikolina Koleva *nikkol@coli.uni-saarland.de*
Martín Villalba *martin.villalba@uni-potsdam.de*
María Staudte *masta@coli.uni-saarland.de*
Alexander Koller *alexander.koller@uni-potsdam.de*