Comparing Listener Gaze with Predictions of an Incremental Reference Resolution Model

Casey Kennington
CITEC, Bielefeld University
ckennington@
cit-ec.uni-bielefeld.de

David Schlangen
Bielefeld University
david.schlangen@
uni-bielefeld.de

Abstract

In situated dialogue, listeners resolve referring expressions incrementally and their gaze often attends to the described objects in the context. We have looked at how listener gaze compares to a statistical reference resolution model that works incrementally. We found that listeners gaze at referred objects even before a referring expression begins, suggesting that salience and prior information is important in reference resolution models.

1 Introduction

Listeners interpret what they are hearing as an utterance is unfolding (Tanenhaus et al., 1995). Furthermore, listeners don't sit idly as they listen: she attends to (i.e., gazes at) objects which are being described, resolving the utterance by finding objects with properties that match the ongoing description (Spivey et al., 2002). Here, we report ongoing work in comparing listener gaze and an incremental reference resolution model, desribed below.

2 Data and Model

Data Imagine playing a game with a friend. The goal is to put together a puzzle to form a shape. However, there are some constraints: you can manipulate the pieces, but cannot see the goal shape; your friend can see the goal shape, but cannot manipulate the pieces; you must work toether to form the goal shape. This was the setting for several recent corpora (Tokunaga et al., 2012; Spanger et al., 2010), a setting which produced a rich number of (exophoric and anaphoric) REs. The IUM has been tested against these data and found to work well in resolving which piece was being referred. For the work presented here, we used the data described in (Spanger et al., 2010) where there are always 7 pizzle pieces. An example is in Figure 1.

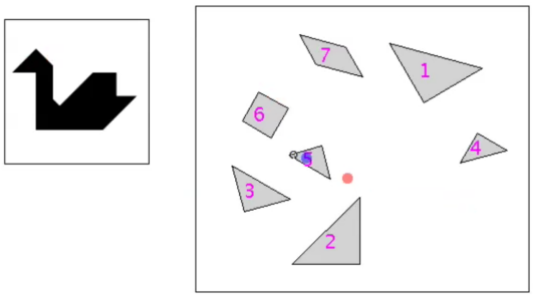

Figure 1: Example Puzzle Board; the goal shape is the swan in the top left, a shared work area is the large board on the right, the mouse cursor and OP gaze (blue dot) are on object 5

Model We apply an *incremental update model* (IUM) (Kennington et al., 2013) requires a set of objects I that could be referred, properties R that belong to those objects, and an (ongoing) utterance U. Formally:

$$P(I|U) = \frac{P(I)}{P(U)} \sum_{r \in R} P(U|R = r)P(R = r|I) \quad (1)$$

Where the posterior $P(I|U)$ at one step becomes the prior $P(I)$ in the next, uniform at first. A classifier can perform $P(U|R)$ by learning co-occurrence between utterances (n-grams or as abstract semantic representations) where the properties are the class labels. The set of properties can be visual such as color or abstract (e.g., Edinburgh has the property of being `Edinburgh`).

3 Comparison with Listener Gaze

We looked at comparing the IUM and listener gaze (OP-GAZE, short for *operator gaze*), as way of determining how well the model might work in an interactive dialogue system. Specifically, we compare the following (here, IUM accuracy measures

whether the argmax of the returned distribution matches the reference):

- RE-level accuracy of IUM and the % of REs where the OP-GAZE looked at the referred object during the REs.
- Incremental (word-level) accuracy of IUM and the % of words where the OP-GAZE looked at the referred object during the word.
- Common words which caused IUM to resolve the reference, and words that caused OP-GAZE to look at the referred object.

Results For RE-level accuracy, OP-GAZE looked at the referred object in 77% of the REs and IUM picked the correct reference object 78% of the time. Figure 2 shows the incremental comparison for referring expressions of length 4-6. Included is a 1.5-second window before the beginning of the RE.

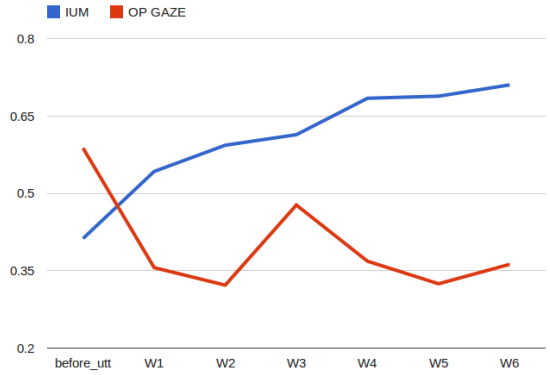

Figure 2: IUM and OP-GAZE accuracy during each word for REs of length 4-6.

Figure 3 shows the final metric. OP-GAZE gazed the most at the referred object 1.5s before the RE and for up to 1s after the REs 50% of the time. The word *there* was uttered often when a mouse pointer was over an object (exophoric reference), and *triangle* was the most common object shape; these helped IUM more than causing OP-GAZE to gaze at the referred object.

4 Discussion and Conclusion

Given this setup, IUM seems to process its input differently than listeners (where gaze is the window into how the listener is processing REs). This is perhaps not too surprising given that IUM starts with an almost empty context, whereas the

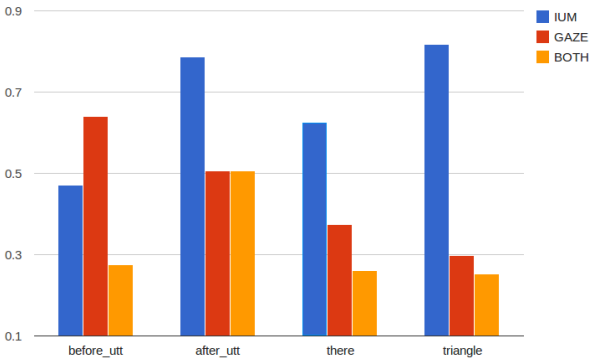

Figure 3: Comparison of IUM accuracy and OP-GAZE accuracy for some common words.

listeners are embedded in the interaction context and thus presumably have much stronger priors on what might be referred. This suggests that models of reference resolution would benefit from incorporating prior information from salient features in a context.

In future work, we will explore in more detail what exactly constitutes these priors, for which the present work forms a good starting point.

References

Casey Kennington, Spyros Kousidis, and David Schlangen. 2013. Interpreting Situated Dialogue Utterances: an Update Model that Uses Speech, Gaze, and Gesture Information. In *SIGdial 2013*.

Philipp Spanger, Masaaki Yasuhara, Ryu Iida, Takenobu Tokunaga, Asuka Terai, and Naoko Kuriyama. 2010. REX-J: Japanese referring expression corpus of situated dialogs. *Language Resources and Evaluation*, 46(3):461–491, December.

Michael J Spivey, Michael K Tanenhaus, Kathleen M Eberhard, and Julie C Sedivy. 2002. Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45(4):447–481.

Michael K Tanenhaus, Michael J Spivey-Knowlton, Eberhard Kathleen M, and Julie C Sedivy. 1995. Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science*, 268:1632–1634.

Takenobu Tokunaga, Ryu Iida, Asuka Terai, and Naoko Kuriyama. 2012. The REX corpora : A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 422–429.