UNIVERSITY OF GOTHENBURG

On Visual Coreference Chains Resolution

Simon Dobnik and Sharid Loáiciga

Centre for Linguistic Theory and Studies in Probability (CLASP)
FLOV, University of Gothenburg
`name.surname@gu.se`

1 Aims

- Situated dialogue: **language and vision**
- **(Co)reference** of descriptions depends on
  - previous dialogue context
  - joint visual attention (Clark and Wilkes-Gibbs, 1986)
- Contexts are evolving
- **A pilot study:** application and error analysis of a textual coreference tool applied to visual dialogue
  - Differences between the domains?
  - Relation between information expressed in language and in vision?
  - Adaptation of coreference tools for visual dialogue?

2 Textual coreference resolution

- A hard task: 0.73 F-score with DNNs (Lee et al., 2018) and 0.63 F-score on the CoNLL2012 dataset
- Use two existing systems from Stanford CoreNLP: parse text, identify mentions, and build a co-referential chain
  - (Lee et al., 2011): a series of filters with patterns
  - (Clark and Manning, 2015): classifiers with a scoring function to combine their outputs
  - Trained on the news domain: OntoNotes (Pradhan et al., 2013)

3 Visual dialogue

- Turns pronounced by different speakers
- "Messy": incomplete and continued utterances, lack of sentence boundaries, personalised spelling, speaker-relative pronouns, etc. (Byron, 2003)
- Different mechanisms for object reference resolution
  - Kelleher (2006): a model of attention in visual dialogue on objects based on linguistic and visual attention scores
- Co-reference chains are not standardly modelled in dialogue

4 The cups corpus

- Corpus of free conversations over perceptual scenes
- Similar to Map Task (Anderson et al., 1991) but conversational roles may change freely
- Swedish: 985 turns and English: 598 turns
- Use the English part, e.g.

A hej
B hej
A först och frömst...
A first of all
A I see lots of cups and containers on the table
B me too
A some white, some red, some yellow, some blue
B I see six white ones
B me too
A i see seven
A but maybe we should move in one direction...
B ok, lets do that

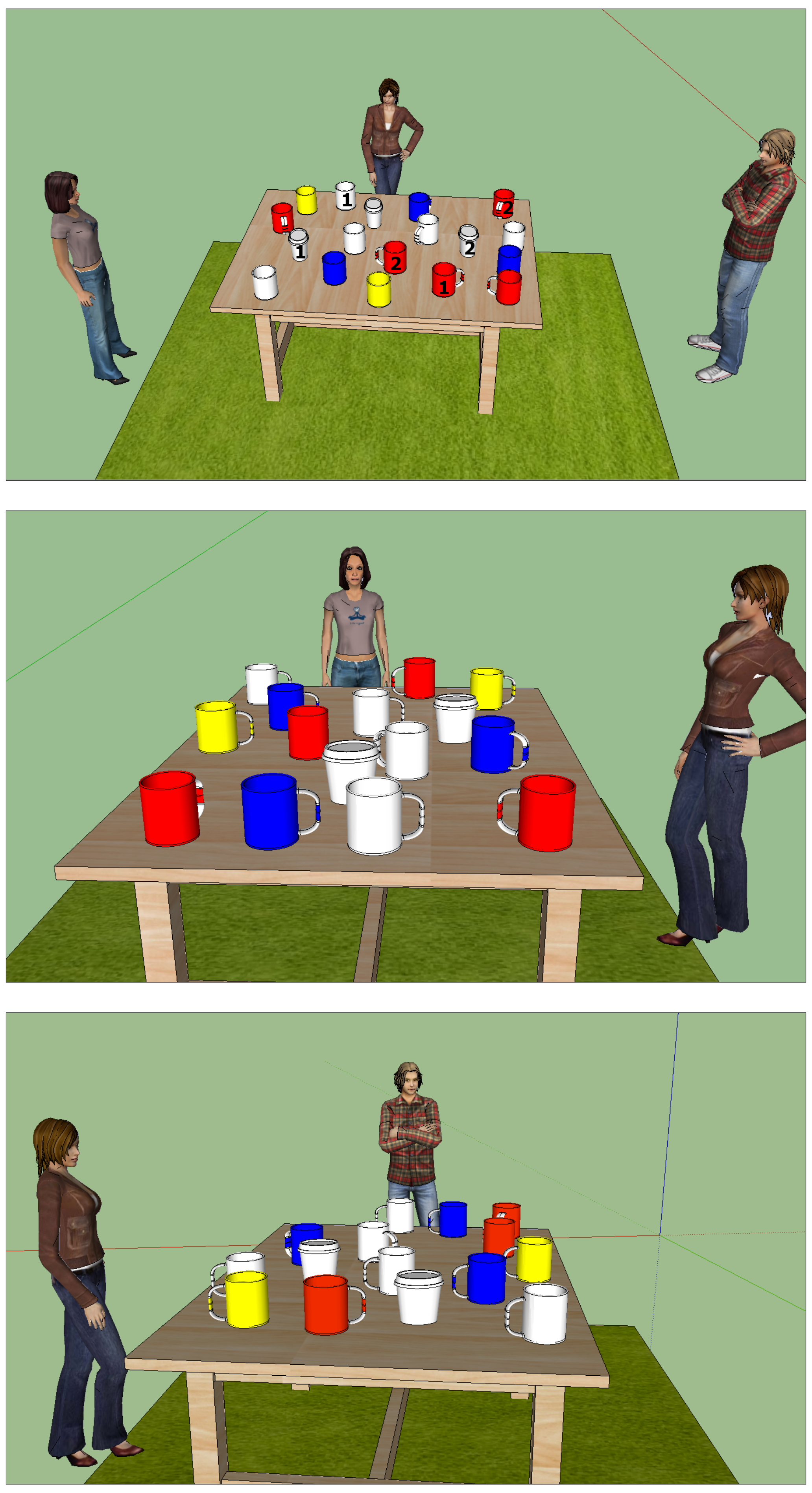

Figure 1: *The table scene from the global perspective and as seen by Participant 1 and 2. The numbers indicate objects hidden from that participant's view.*

5 Annotation of co-reference chains

- Two annotators, the first 100 turns of the GU-EN-P1 dialogue
- CoNLL format (Pradhan et al., 2011)
  - Number ID of a chain assigned to each word
  - IDs uniquely refer to physical objects in the scene, the participants, Katie, the table and the common locations (B's-left, Katie's-right)
  - Brackets indicate phrases/mentions
- Example annotation file:

```
A  1  i           (2)
A  2  see
A  3  lots         (5
A  4  of
A  5  cups
A  6  and
A  7  containers  5)
A  8  on
A  9  the
A 10  table       (4)

B  1  me          (1)
B  2  too

A  1  some         (5
A  2  white        5)
A  3  ,
A  4  some         (5
A  5  red          5)
A  6  ,
A  7  some         (5
A  8  yellow       5)
A  9  ,
A 10  some         (5
A 11  blue         5)
```

6 Challenges for annotating visual dialogue

- Descriptions are made from different points of view of participants: e.g. 'I', 'you', 'from my side'
- NPs with common nouns, e.g. 'the red cup', also have several referents over the conversation; co-reference chains are less predictable from textual features but require dialogue and visual attention
- The objects are not discussed in a particular order; the same object may be described again at a (much) later stage of the dialogue; co-reference chains may be very long
- Participants dynamically create 'new objects' representing groups of objects and locations: 'my white ones (cups)' and 'the empty space in front of you'
- As task involves ambiguity, participants may assign different reference to one particular description

7 Results

- Both sieve-based and statistical systems yielded the same output
- The systems were unable to find any of the gold links in our data
- The official co-reference scorer with the CoNLL12 data could not be used to calculate MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF (Luo, 2005), and BLANC (Recasens and Hovy, 2010)
- Errors due to the dynamic reference of descriptions in dialogue texts:
  - identical form is a strong feature for determining co-reference
  - all pronouns 'I' and 'me' assigned into the same chain
  - 'my left' and 'your left'
- Parser:
  - correct sentence boundaries: 162 vs 157 in the gold annotation
  - turns identified as sentences
  - good performance on multi-word expressions: 'a white funny top' and 'the third row from you'
- Mention identification:
  - Annotation: 293 mentions of 43 entities
  - Systems: 88 mentions and 28 entities
  - None identified correctly
  - Mention span: 'left' and 'red mug' (A) vs 'her left' and 'a read mug' (S)
  - Only 12 mention matches: precision = $12/88 = 0.14$ and recall = $12/293 = 0.04$

8 Conclusions

- The two co-reference resolution systems tested cannot handle visual dialogue data
- The annotated corpus will help us to create a system that models both vision and language components
- Co-reference is not directly observable in visual and textual features: mechanisms of attention (Dobnik and Kelleher, 2016)

CLASP centre for linguistic theory and studies in probability

The Seventh Swedish Language Technology Conference (SLTC), Stockholm 7–9 November 2018