On Visual Coreference Chains Resolution

Simon Dobnik **Sharid Loáiciga**
Department of Philosophy, Linguistics, Theory of Science (FLoV)
Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
{simon.dobnik,sharid.loaiciga}@gu.se

Abstract

We explore to what degree an existing textual coreference resolution tools can be applied to visual dialogue data. The analysis of error of the coreference system (i) demonstrates the extent to which such data differs from the written document texts that these tools are typically applied on; (ii) informs about the relation between information expressed in language and vision; and (iii) suggests further directions in which coreference tools should be adapted for visual dialogue.

1 Introduction and Related Work

"Situated" dialogue involves language and vision. An important aspect of processing situated dialogue is to resolve the reference of linguistic expressions. The challenging aspect is that descriptions are local to the current dialogue and visual context of the conversation (Clark and Wilkes-Gibbs, 1986) and that not all information is expressed linguistically as a lot of meaning can be recovered from the joint visual and dialogue attention. Coreference resolution has been studied and modelled extensively in the textual domain where the scope of the processing coreference is within a document. Robust coreference resolution for dialogue systems is a very much needed task. The aim of this paper is to provide a preliminary investigation of to what degree an existing off-the-shelf textual coreference resolution tool can be used in the domain of the visual dialogue.

Given its popularity in contexts with scarce amounts of training data, such as dialogue systems, we use the Lee et al.'s 2011 sieve-based system here. For comparison, we also use Clark and Manning's 2015 mention-pair system. Both are freely available through the Stanford CoreNLP distribution.

Unlike the neatly structured written text which is organised in documents, dialogue data is messy. The text is structured in turns that are pronounced by different speakers, and sentence boundaries are not clear (cf. Byron (2003) for an overview). Work on referring expressions generation (e.g. Krahmer and van Deemter, 2011; Mitchell et al., 2012; Xu et al., 2015; Lu et al., 2017), on its part, does not typically involve dialogue or the notion of coreference chain – a central construct for coreference resolution systems. Furthermore, coreference resolution tools for dialogue are often custom built to the specific needs of companies or datasets (Rolih, 2018; Smith et al., 2011).

2 Data Processing

The dataset We take the English subsection of the Cups corpus (Dobnik et al., 2015) which consists of two dialogues between two participants with 598 turns in total. The goal of this corpus is to sample how participants would refer to things in a conversation over a visual scene. A virtual scene involving a table and cups has been designed in with a 3-d modelling software (Figure 1). Some cups have been removed from each participant's view and the participants are instructed to discuss over a computer terminal in order to find the cups that each does not see. The task therefore resembles the Map Task (Anderson et al., 1984, 1991).

Annotation In this pilot study two annotators additionally annotated the first 100 turns of the GU-EN-P1 dialogue for coreference chains. The annotation follows the CoNLL format with the last column containing the coreference chains (Pradhan et al., 2011). Each chain is assigned a number id, where the first and the last tokens of a mention within the chain are identified with opening and closing brackets, as illustrated in Figure 2. In this example, the mentions 'cups and containers', 'some white, 'some red', 'some yellow', and 'some blue', all belong to the same chain.

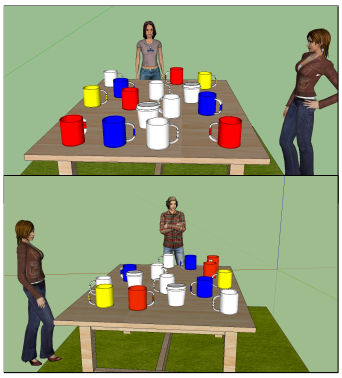

Figure 1: Scene as seen by Participants 1 and 2.

A	1	i	(2)	A	1	some	(5
A	2	see		A	2	white	5)
A	3	lots		A	3	,	
A	4	of		A	4	some	(5
A	5	cups	(5	A	5	red	5)
A	6	and		A	6	,	
A	7	containers	5)	A	7	some	(5
A	8	on		A	8	yellow	5)
A	9	the		A	9	,	
A	10	table	(4)	A	10	some	(5
				A	11	blue	5)
B	1	me	(1)				
B	2	too					

Figure 2: Annotation of coreference chains

This is the standard scheme used on textual data consisting of documents, but presented two challenges for annotation: (i) in the dialogue data descriptions are made by two conversational participants in turns from their own point of view hence pronouns 'I' and 'you' as well as spatial descriptions such as 'from my view' will have a different referent depending on the speaker; and (ii) a description 'the red cup' does not have a unique referent through the dialogue but this changes depending on the previous dialogue states and the focus on the scene. Hence, the annotators also used a visual representation of the scene and descriptions were identified as belonging to the same coreference chain only if they were referring to the same physical object. We assigned fixed ids to all existing objects in the scene (the cups and the table), person A and B, 'Katie' and the table as well as frequently used parts of the scene such as B's-left, Katie's-right. Dialogue participants also dynamically create 'objects' throughout the conversation that they are later referred to as normal objects, e.g. 'the empty space in front of you', 'my white ones (cups)'. For these, annotators introduced additional ids and their approximate location was marked in the representation of the scene.

2.1 Results

We run the annotated data through both the sieve-based and statistical systems from the CoreNLP distribution. Both yielded the exact same output, so our analysis does not distinguish between them.

The official coreference scorer commonly used in the domain searches for complete coreference links, and since the systems were unable to find any of the gold links in our data, the scorer produced appallingly negative results. A major cause behind this inability to identify the coreference chains accurately lies on the deictic nature of this particular type of dialogue text and the fact that it consists of speaker turns. For instance, the systems grouped all pronouns 'I' and 'me' into the same chain (and therefore the same entity) because they have identical forms which is a strong feature for assessing coreference in these systems. This problem affects basically all mentions that refer back to some description in a changing context such as 'my left' and 'your left'.

Concerning the parser, a central element to these systems, we observed that the sentences boundaries were identified often correctly (162 versus 157 in the gold), meaning that almost every turn in the dialogue was identified as a sentence. Looking at the mentions, however, from 293 manually annotated mentions distributed over 43 entities, the systems were not able to identify any of them correctly. On the contrary, the systems proposed 88 mentions and 28 entities. Further investigation reveals that a major problem was the correct identification of the mention span. For instance, in one sentence, in the gold the mentions 'left' and 'red mug' were annotated, but the system identified the maximum spans 'her left' and 'a red mug' instead. We counted only 12 mention matches due to this problem, yielding a precision of 12 / 88 = 0.14 and a recall of 12 / 293 = 0.04.

3 Conclusions

The results of our pilot study show that at least the two coreference resolution systems tested cannot handle visual dialogue data. We expect that our annotations will help us create a data-driven coreference resolution system able to simultaneously model both the language and visual components of this dataset, similar to Kelleher (2006).

Acknowledgements

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Anne Anderson, Gillian Brown, Richard Shillcock, and George Yule. 1984. *Teaching talk: Strategies for production and assessment*. Cambridge University Press, United States.

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC map task corpus. *Language and speech*, 34(4):351–366.

Donna K Byron. 2003. Understanding referring expressions in situated language some challenges for real-world agents. In *Proceedings of the First International Workshop on Language Understanding and Agents for Real World Interaction*, pages 39–47.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415. Association for Computational Linguistics.

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 24–32, Gothenburg, Sweden.

John D Kelleher. 2006. Attention driven reference resolution in multimodal contexts. *Artificial Intelligence Review*, 25(1-2):21–35.

Emiel Krahmer and Kees van Deemter. 2011. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34. Association for Computational Linguistics.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. ArXiv:1612.01887 [cs.CV].

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27. Association for Computational Linguistics.

Gabi Rolih. 2018. Applying coreference resolution for usage in dialog systems. Master's thesis, Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden.

Cameron Smith, Nigel Crook, Simon Dobnik, Daniel Charlton, Johan Boye, Stephen Pulman, Raul Santos de la Camara, Markku Turunen, David Benyon, Jay Bradley, Björn Gambäck, Preben Hansen, Oli Mival, Nick Webb, and Marc Cavazza. 2011. Interaction strategies for an affective conversational agent. *Presence: Teleoperators and Virtual Environments*, 20(5):395–411.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv*, 1502.03044v3 [cs.LG]:1–22.