On Visual Coreference Chains Resolution

Simon Dobnik and Sharid Loáiciga

CLASP
University of Gothenburg
name.lastname@gu.se

1. Introduction

"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. Co-reference resolution has been studied and modelled extensively in the textual domain where the scope of the processing co-reference is within a document. Robust co-reference resolution for dialogue systems is a very much needed task. In this paper we explore to what degree an existing textual co-reference resolution tool can be applied to visual dialogue data. The analysis of error of the co-reference system (i) demonstrates the extent to which such data differs from the written document texts where these tools apply; (ii) provides about the relation between information expressed in language and vision; and (iii) suggests further directions in which co-reference tools should be adapted for visual dialogue.

2. Related Work

Textual coreference resolution is a hard task in its own. Before current end-to-end neural systems raised the state of the art to up to 0.72 F-score in 2017, co-reference resolution success was around 0.63 F-score on the CoNLL2012 dataset. The best performing system to this day for English is that of Lee et al. (2018), who reports an F-score of up to 0.73 in the same dataset. If we compare these scores with other NLP tasks such as named entity recognition or parsing (both with more than 90% accuracy), they appear low.

Given its popularity in contexts with scarce amounts of training data, such as dialogue systems, here we use the Lee et al. (2011)' sieve-based system. For comparison, we also use Clark and Manning (2015)'s mention-pair system. Both are freely available through the Stanford CoreNLP distribution. Building on the output of a parser, they both first identify mentions and then decide if these mentions belong to the same co-referential chain, i.e, they all refer to the same entity. The first achieves this decision making through a series of filters for matching different patterns and the second with two classifiers and a scoring function to combine their outputs.

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 (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 co-reference chain –a central construct for co-reference resolution systems. Furthermore, co-reference resolution tools for dialogue are often custom built to the specific needs of companies or datasets (Rolih, 2018; Smith et al., 2011).

Our aim is to treat vision and language in a uniform manner. For example, (Kelleher, 2006) describes a model of attention in visual dialogue where the attention score is calculated for objects as the weighted integration of linguistic and visual attention scores which are then used in a ranked resolution of reference. (Stoia et al., 2006) proposes a similar model for the domain of route instructions. In all these models, the notion of co-reference chain is not taken into account as in the textual co-reference resolution domain.

The aim of this paper is to provide a preliminary investigationi of to what degree an existing off-the-shelf textual co-reference resolution tool can be used in the domain of the visual dialogue.

3. Data Processing

3.1 Method

The dataset We take the English subsection of the Cups corpus (Dobnik et al., 2015) which consists of two dialogues, each involving two participants, resulting in 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 a 3-d modelling software and two avatars have been placed at the opposite side of this table representing the conversation participants. A third avatar who is a passive observer of the scene is standing at the side. A screenshot of the scene from each participants view is taken and furthermore some cups have been removed from each participants view but which the other participant can see (Figure 1). The participants are instructed to discuss over a computer terminal their view of the virtual world with each other in order to find the cups that each does not see. An example of the ellicited dialogues is given in example (1).

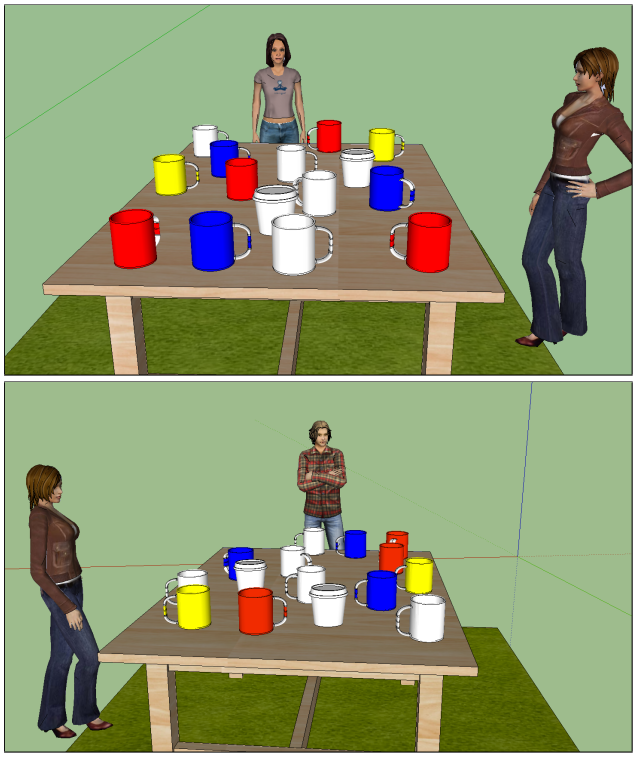

Figure 1: The table scene as seen by Participants 1 and 2 respectively.

(1) A hej
B hej
A först och främst...
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

Annotation In this pilot study two annotators annotated the first 100 turns of the GU-EN-P1 dialogue for co-reference chains as described in Pradhan et al. (2011). The annotation follows the CoNLL format with the last column containing the co-reference chains. 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 example (2). In this example, the mentions 'lots of cups and containers' , 'some white, 'some red', 'some yellow', and 'some blue', all belong to the same chain.

This is the standard scheme used on textual data consisting of documents, but presented two challenges for our annotation: (i) in the dialogue data descriptions are made by two conversational participants 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 context; and (ii) a description 'the red cup' does not have a unique referent through the dialogue but this changes depending on the previous state of the dialogue and the focus on the scene. Both facts are related to our earlier observation that in visual dialogue information is not only communicated in words but also relying on joint attention.

(2)	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)

Hence, the annotators also used a visual representation of the scene and descriptions were identified as belonging to the same co-reference 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), as well as person A and B, 'Katie' and the table as frequently used parts of the scene such as B's-left, Katie's-right. However, 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. We expect that the challenge of this data and annotation for a textual co-reference system will be the fact the co-reference chains may be very long, e.g. 'I' and 'you' for the entire length of the dialogue. Also, the co-reference chains may be threaded as the same objects may be discussed again in another section of the dialogue. As the dialogue participants do not see exactly the same scene and they see it from a different perspective they may not be referring to the same object although they might believe so.

3.2 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 co-reference scorer provided with the CoNLL12 data computes the standard measures MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF

(Luo, 2005), and BLANC (Recasens and Hovy, 2010)). However, this scorer searches for complete co-reference links, and since the system was unable to find any of the gold links in our data, this oficial scorer produced appalling negative results.

A major cause behind this inability to identify the co-reference chains accurately lies on the dynamic nature of this particular type of dialogue text. For instance, the pronouns 'I' and 'me' refer to either Participant A or B, changing their reference actively as the participants use them, but the systems grouped all pronouns 'I' and 'me' into the same chain (and therefore the same entity) because they have identical forms which is one strong feature for determining co-reference in these systems. This problems 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. Some multi-word mentions such as 'a white funny top' or 'the third row from you' were also correctly analysed, suggesting further that the quality of the parser and the mention identification component was acceptable.

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 at the mention level reveals that a major problem was the correct identification of the mention span. For instance, in one sentence, the gold the mentions 'left' and 'red mug' were annotated, but the system identified 'her left' and 'a read mug' instead, producing a complete mistmatch. 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.

4. Conclusions

The results of our pilot study show that at least the two co-reference resolution systems tested cannot handle visual dialogue data. We expect that the created annotations will help us create a system able to simultaneously model both the language and visual components of this dataset. Current approaches to combining vision and language, e.g. (Xu et al., 2015; Lu et al., 2017) demonstrate that successful deep learning models involving vision and language can be built in the domain of static image captioning. Co-reference resolution (or generation) is a further step where such systems would be applied in a dynamic context. One difficulty that we expect for unsupervised approaches is that co-reference in visual dialogue is not directly observable in features; humans use complex mechanisms of attention to reach joint understanding. This means that a large amount of quality annotated data will be required and effectively the system will have to a learn a model of attention (cf. (Dobnik and Kelleher, 2016) for a top-down mechanistic model of attention).

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566, Granada, Spain.

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.

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.

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

Simon Dobnik and John D. Kelleher. 2016. A model for attention-driven judgements in type theory with records. In Julie Hunter, Mandy Simons, and Matthew Stone, editors, *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue*, volume 20, pages 25–34, New Brunswick, NJ USA, July 16–18.

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In Christine Howes and Staffan Larsson, editors, *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 24–32, Gothenburg, Sweden, 24–26th August.

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.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana, June. 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], 6 June.

Xiaoqiang Luo. 2005. On coreference resolution perfor-

mance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT/EMNLP 2005, pages 25–32, Vancouver, British Columbia. Association for Computational Linguistics.

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.

Marta Recasens and Eduard Hovy. 2010. Coreference resolution across corpora: languages, coding schemes, and preprocessing information. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL 2010, pages 1423–1432, Uppsala, Sweden. 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.

Laura Stoia, Darla Magdalene Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2006. Noun phrase generation for situated dialogs. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 81–88, Sydney, Australia, July. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on message understanding*, MUC-6, pages 45–52, Columbia, Maryland. Association for Computational Linguistics.

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, February 11.