How do images help coreference?
A case study on the multi-modal Tell-me-more dataset

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

Abstract

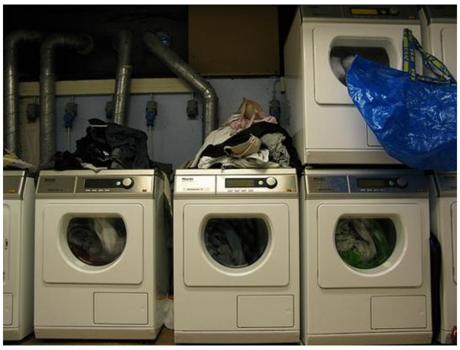
This case study analyses the errors and overall performance of the textual coreference resolution system LINK-APPEND (Bohnet et al., 2023) on the long-text image description dataset Tell-me-more (Ilinykh et al., 2019). We also identify challenges in processing and annotating coreference in multi-modal domain. Our analysis is based on a single dataset, but it invites further discussion on the importance of visual knowledge in modelling coreference.

1 Introduction

Coreference resolution (CR) is a task in which a model links different linguistic expressions referring to the same entity together. The task is typically divided into two sub-tasks: first, mentions are classified into referential or non-referential; second, mentions referring to the same entity are clustered or grouped together into *coreference* chains. One of the challenges that coreference resolution models face is that resolving coreference can require relying on extra-textual information and clues. Consider the following example:

(1) The red apple on the right has something like an apple company logo on it. The "Tasty Food Company" apple must be very tasty[1].

While the model can use syntactic cues to recognize that "it" and "the red apple" refer to the same entity, identifying "The Tasty Food Company" apple as the same red apple from the first sentence requires both world knowledge and an image showing the apple with the company logo. Additionally, if multiple apples are present in the context, "the red apple on the right" could help distinguish between them if an image were available. These different interpretations suggest that visual features could help in identifying antecedents, i.e, the specific entities or objects to which mentions refer back.



1. It's a picture of what look like **[washing machines]**[2].
2. There are three of **[them]**[2] in a row, plus **[one stacked on top]**[1].
3. A big blue bag is hanging from **[the top right washing machine]**[1]. There are four large silver pipes/tubes coming out of the wall and running behind **[the machines]**[2].
4. There's a pile of clothes stacked on top of the two left washing machines. **[They]**[2] all have clear doors so you can see there's also clothing inside **[them]**[2].
5. The angle of the picture means you can't see the floor of the room.

Figure 1: Image and its description from the Tell-me-more dataset (Ilinykh et al., 2019). We show two coreference chains that can be correctly formed in the context of the image. Relying on text alone makes the task of coreference resolution in this example challenging.

In this case study, **we explore whether visual information (as found in images) is useful for the coreference resolution task**. Our analysis is limited to one small dataset and our goal is to study examples of automatically identified coreference chains and estimate the extent to which visual information can help a neural coreference resolution system. We use the Tell-me-more dataset (Ilinykh et al., 2019), which consists of multi-sentence image descriptions of house environments. These descriptions were generated by Amazon Mechanical Turk (AMT) crowdworkers, who were shown an image and asked to describe it in a way that

[1]Made up example by the authors. "Tasty Food Company" refers to an imaginary company that sells apples.

would help someone identify it within a larger set of images. The descriptions in the dataset include referring expressions to objects in the image, and many of these expressions corefer with each other, e.g., “*There is [a pantry] in the kitchen. There is a white door to [the pantry].*".

As coreference model, we study the output of the decoder-based state-of-the-art neural coreference system, LINK-APPEND (Bohnet et al., 2023). LINK-APPEND adopts language models for coreference resolution through a text generation task. This method trains the model to list all referring expressions in each sentence of a document while the document is being generated. Inspired by transition-based parsing, the LINK-APPEND system links each identified mention to an antecedent to form a new set or appends it to an already existing set of coreferring expressions. LINK-APPEND performs well primarily because it leverages pretrained knowledge from the 13-billion-parameter multilingual T5 model (Xue et al., 2021). Importantly, the model has been fine-tuned on textual corpora that do not include accompanying images for grounding the texts. Our results suggest that the LINK-APPEND model, which has been trained and tested on text-only datasets for coreference, struggles with chains that can be resolved by referencing the image, e.g., chains 1 and 2 in Figure 1. Our analysis is promising for future experiments and the development of new multimodal coreference resolution systems, offering a potential means to enrich their world knowledge.

2 Background

Some of the recent neural coreference resolution systems are encoder-only and do not frame coreference through text generation. Examples include LingMess (Otmazgin et al., 2023) and Maverick (Martinelli et al., 2024). LingMess achieves better results on several coreference datasets, while Maverick is more resource-efficient and faster during inference. Despite the strengths of the encoder-based coreference resolution models, decoder-based models such as LINK-APPEND are incremental. This property makes them more similar to how humans process coreference in real-world text production tasks, as research shows that humans resolve referring expressions incrementally (Altmann and Steedman, 1988).

Existing work on coreference in the language-and-vision domain focuses mainly on visual dialogue (Kottur et al., 2018; Li and Moens, 2021). In these studies images are paired with a history of question-and-answer pairs, creating a relatively straightforward coreference scenario. For instance, questions typically involve a pronoun whose antecedent can be found in the preceding utterance: “*There is [a boat] on the water. What colour is [it]? [It] is green.*”. This contrasts with the example in 1, where the image is necessary to identify the antecedent of *the Tasty Food Company apple*. The Tell-me-more dataset offers a more complex multimodal scenario in which text alone is not always sufficient to resolve coreference. The example in Figure 1 demonstrates that the image is needed to link, for example, “one stacked on top” with “the top right washing machine”.

3 Coreference annotations

We use existing coreference annotations in the Tell-me-more dataset collected with two human experts and described in Loáiciga et al. (2022)[2]. We remove instances with missing annotations and collect 536 annotations for image-description pairs.

Modest in size, this annotation covers a diverse array of coreference types (e.g., anaphora, bridging) alongside links to objects in images. This is important because it allows us to compute the standard coreference metrics, a feature rarely addressed in coreference resolution work in the multimodal domain. Previous work has focused on grounding pronouns in images (Yu et al., 2019), noun phrases and pronouns (Lu et al., 2022) or looked at the domain of visual dialogue (Dobnik and Loáiciga, 2019).

There are two important properties of the annotations to keep in mind. First, two annotators were free to determine boundaries of each mention and to decide which ones belong to a single chain As reported in the annotation description, this has resulted in imperfect matches and variation between identified boundaries of mentions. Second, the images that the annotators were provided with included bounding boxes from a pre-trained object detector. Out of those many bounding boxes annotators were required to freely choose which bounding box refers to which mention. Such degree of freedom in the annotation process could result in inconsistencies in the annotated text and in how objects in the image are linked with men-

[2]The annotations are publicly available at https://zenodo.org/records/7084861

Metric	object-based			text-based		
	Recall	Precision	F1	Recall	Precision	F1
Mention identification	46.95	71.25	56.60	76.68	57.65	**65.82**
MUC (Vilain et al., 1995)	41.49	61.90	49.68	67.02	49.00	**56.61**
B^3 (Bagga and Baldwin, 1998)	41.27	62.47	49.71	68.48	48.60	**56.85**
$CEAF_e$ (Luo, 2005)	43.93	67.91	53.35	71.47	55.67	**62.59**
LEA (Moosavi and Strube, 2016)	37.45	56.69	45.10	62.61	43.44	**51.29**
CoNLL Score (Pradhan et al., 2011)			50.91			**58.68**

Table 1: Automatic evaluation of the performance of LINK-APPEND coreference resolution system. The system's performance is compared against two sets of coreferences: one derived from matches between bounding boxes of described objects (object-based), and another one is constructed with human annotations (text-based).

tions. To identify such inconsistencies, we perform two types of analyses: i) **text-based**, which considers the coreference chains with the same set id, and ii) **object-based**, which examines the coreference chains whose mentions are linked to the same bounding box in the image.

According to Loáiciga et al., if mentions are co-referential in the text, they are also linked to their corresponding bounding boxes in the image, when available during annotation. However, we found inconsistencies in the resulting annotations as the two annotation sets (text-based and object-based) are not entirely identical. We found that there are 797 unique mentions that appear in both sets across all annotated documents, while 921 mentions appear in either text-based set or object-based set. There are 57 mentions unique to the object-based set and 864 mentions unique to the text-based set. Upon manual inspection of sets, we saw that one reason for such a large number of non-overlapping mentions is disagreement between annotators on mention boundaries as also reported by Loáiciga et al., e.g., "something green" vs. "something green that" for a single mention based on results from two annotators. This is an interesting finding, as it suggests that either (1) the annotation task is complex and hard to frame in a simple and intuitively clear way, (2) not every co-referential mention in text might be linked with the corresponding bounding box in the image (which is a problem of the bounding box/object extractor), or (3) vice versa, not every bounding box that refers to mentions appearing multiple times in texts has been annotated by the annotators (annotation mistake). By creating two different held-out sets of coreference we study inconsistencies in the annotations that can be used to evaluate coreference models and identify points that would allow us to improve future annotation guidelines.

We predict coreference chains by feeding each text into the LINK-APPEND model. We then compare the model-predicted coreference chains with the two sets of chains (text-based and object-based) and compute the standard coreference metrics. Metrics for automatic coreference resolution compare two sets of coreference chains, one is typically generated by a model and the second one is the gold coreference. These metrics typically compute precision (i.e., how many of the coreference chains identified by the system are actually correct), recall (i.e., how many of the actual coreference chains in the ground-truth data are identified by the model), and the F1 score, which is a combination of the two. The CoNLL score (the average of MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and $CEAF_e$ (Luo, 2005)) is also widely reported. For automatic evaluation of coreference we rely on the CoVal package (Moosavi and Strube, 2016)[3].

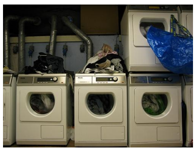
4 Results

As Table 1 shows, LINK-APPEND is generally better at predicting mentions and coreference chains when it is compared against the text-based set. When the ground-truth is changed to the object set (i.e., the one that is based on annotated links between mentions and objects), the model's performance drops in F1 score, recall and CoNLL score. These results could mean that either the model is not predicting mentions and chains that can also be linked with the image or the differences between the two sets of human annotations are too large or both. Example 3a shows that a human did not an-

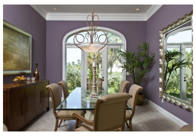
[3]Available at `https://github.com/ns-moosavi/coval/tree/master?tab=readme-ov-file`.

Statistic	model-generated	text-based	object-based
Total number of coreference chains	793	1018	513
Average number of coreference chains per document	1.48	1.90	0.96
Average length of coreference chains	2.42	2.52	2.47

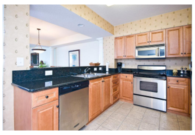
Table 2: Statistics about coreference chains in three different sets.

	model-generated	text-based	object-based
	[It] 's a picture of what look like [washing machines]. There are three of [them] in a row, plus one stacked on top.	[It] 's [a picture] of what look like [washing machines]. There are three of [them] in a row, plus one stacked on top.	It's a picture of what look like washing machines. There are three of them in a row, plus [one] stacked on top.
	A big blue bag is hanging from the top right washing machine.	A big blue bag is hanging from the top right washing machine.	A big blue bag is hanging from [the top right washing machine].
	There are four large silver pipes/tubes coming out of the wall and running behind [the machines].	There are four large silver pipes/tubes coming out of the wall and running behind the [machines].	There are four large silver pipes/tubes coming out of the wall and running behind the machines.
	There's a pile of clothes stacked on top of the two left washing machines. [They] all have clear doors so you can see there's also clothing inside [them].	There's a pile of clothes stacked on top of the two left washing machines. [They] all have clear doors so you can see there's also clothing inside [them].	There's a pile of clothes stacked on top of the two left washing machines. They all have clear doors so you can see there's also clothing inside them.
	The angle of [the picture] means you can't see the floor of the room.	The angle of [the picture] means you can't see the floor of the [room].	The angle of the picture means you can't see the floor of the room.

(a) Example doc_ann1-84.

	model-generated	text-based	object-based
	[Elegant Room with [a HUGE archway window on one wall]].	[Elegant Room] with [a HUGE archway window] on one wall.	Elegant Room with [a HUGE archway window] on one wall.
	and the mirror that reflect [the window] makes [it] look like [it] has two windows but only has one.	and [the mirror] [that] reflect [the window] makes [it] look like [it] has two windows but only has [one].	and the mirror that reflect [the window] makes it look like it has two windows but only has [one].
	The wall paint is a dark gray-purple color.	The wall paint is a dark gray-purple color.	The wall paint is a dark gray-purple color.
	The dining table is glass.	The dining table is glass.	The dining table is glass.
	The dining set seats six.	The dining set seats six.	The dining set seats six.

(b) Example doc_ann1-423.

	model-generated	text-based	object-based
	It 's a kitchen that overlooks into [a table area].	[It] 's [a kitchen that overlooks into [a table area]].	It 's a kitchen that overlooks into a table area.
	The countertops are shiny black and the cupboards are a caramel color. All lower cupboards and upper cupboards along back wall.	[The countertops] are shiny black and [the cupboards] are [a caramel color]. All lower cupboards and upper cupboards along [back wall].	The countertops are shiny black and the cupboards are a caramel color. All lower cupboards and upper cupboards along [back wall].
	The right side has a dishwasher built in and a sink along the top. There is also a ledge and opening cut out to look into [a table area].	The right side has [a dishwasher] built in and a sink along [the top]. There is also a ledge and opening cut out to look [into dining area].	The right side has a dishwasher built in and a sink along the top. There is also a ledge and opening cut out to look into dining area.
	The back wall has oven/range combo built into the bottom and microwave built into the top.	[The back wall] has [oven/range combo] built into the bottom and [microwave] built into the top.	[The back wall] has oven/range combo built into the bottom and microwave built into the top.
	All the appliances are stainless steel, floor is tan laminate tile, and there is [wallpaper all over with circleish shapes on [it]].	[All the appliances] are [stainless steel], [floor] is [tan laminate tile], and there is [wallpaper] all over with circleish shapes on [it].	All the appliances are stainless steel, [floor] is [tan laminate tile], and there is wallpaper all over with circleish shapes on it.

(c) Example doc_ann2-444.

Table 3: Three examples of coreference chains produced by the model (model-generated), found in human annotations (text-based) or extracted based linking between bounding boxes of objects and referring expressions (object-based). Mentions in the same coreference chains are coloured identically.

notate "one" in the first sentence and "the top right washing machine" in the second sentence as coreferential expressions, although the same annotator annotated these mentions with the same bounding boxes. Consistency in annotations (i.e., links between objects and mentions as well as mentions and chains in text) is crucial, as in this particular example, the image is necessary to resolve the coreference identified in the object-based set.

Looking at the text-based scores in detail, we see that LINK-APPEND identified many links in coreference chains but also made many false-positive predictions. This hints at weaknesses from the model to identify boundaries of mentions. Consider the model-generated and text-based coreference chains in Example 3b: while the model has identified the coreference chain that includes references to the room (e.g., "Elegant Room"), it fails to correctly determine mention borders (e.g., "Elegant Room with a HUGE archway window on the wall" vs "Elegant Room"). However, this can also be viewed as a problem of the annotation: "with a HUGE archway window on the wall" is an embedded clause in this case and it could have been annotated as such. Turning to the object-based set, the model shows low recall and high precision. This suggests that the model performs better at correctly identifying these coreference chains and the mentions within but still misses many chains. Example 3a and Example 3c illustrate this idea. In particular, in the latter example the model has missed a lot of coreference chains, and has not identified the chains [“back wall”, “The back wall”] and [“floor”, “the laminate tile”], while both of them are present in the annotations. Potentially, the model could learn to identify these chains by looking at the spatial arrangement of objects in the image and using general knowledge about room layout (e.g., floors can have tan laminate tile).

5 Conclusion

Our case study on coreference in the multi-modal domain shows that there is room for visual information in state-of-the-art decoder-based neural coreference systems like LINK-APPEND (Bohnet et al., 2023). We have also identified challenges in human annotation of coreference in the language-and-vision domain. Our analysis suggests that human error during annotation can occur due to the complexity of task instructions and inaccuracies in the automatic models used to generate data for annotation.

Future research will investigate different ways of integrating visual and linguistic representations in integrated embeddings in order to support multimodal coreference resolution. Another potential direction is the annotation of a larger dataset as the one used in this study is relatively small, making it challenging to train robust models that can generalize across different datasets and tasks.

The presented analysis is focused on a specific domain (detailed descriptions of house environments) and can be used as a targeted evaluation benchmark, for instance, to measure the ability of large pre-trained multi-modal models to identify co-referential mentions in a multi-modal house navigation context. Given that coreference is a central component for textual coherence and has a long-standing tradition in linguistic research, we believe that integrating multimodal information is the next step for building models capable of incorporating world knowledge.

Acknowledgments

The research in this paper is 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. The authors also thank the three anonymous reviewers for their comments.

References

Gerry Altmann and Mark Steedman. 1988. Interaction with context during human sentence processing. *Cognition*, 30(3):191–238.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566.

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Simon Dobnik and Sharid Loáiciga. 2019. On visual coreference chains resolution. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, London, United Kingdom. SEMDIAL.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Tell me more: A dataset of visual scene description sequences. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 152–157, Tokyo, Japan. Association for Computational Linguistics.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, volume 11219 of *Lecture Notes in Computer Science*, pages 160–178. Springer.

Mingxiao Li and Marie-Francine Moens. 2021. Modeling coreference relations in visual dialog. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3306–3318, Online. Association for Computational Linguistics.

Sharid Loáiciga, Simon Dobnik, and David Schlangen. 2022. Anaphoric phenomena in situated dialog: A first round of annotations. In *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 31–37, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Panzhong Lu, Xin Zhang, Meishan Zhang, and Min Zhang. 2022. Extending phrase grounding with pronouns in visual dialogues. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7614–7625, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024. Maverick: Efficient and accurate coreference resolution defying recent trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13380–13394, Bangkok, Thailand. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2023. LingMess: Linguistically informed multi expert scorers for coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2752–2760, Dubrovnik, Croatia. 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, Portland, Oregon, USA. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995.*

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. What you see is what you get: Visual pronoun coreference resolution in dialogues. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5123–5132, Hong Kong, China. Association for Computational Linguistics.