Tell Me More: A Dataset of Visual Scene Description Sequences

Nikolai Ilinykh
Dialogue Systems Group
Bielefeld University
nikolai.ilinykh@uni-bielefeld.de

Sina Zarrieß *
Digital Humanities
University of Jena
sina.zarriess@uni-jena.de

David Schlangen *
Computational Linguistics
University of Potsdam
david.schlangen@uni-potsdam.de

Abstract

We present a dataset consisting of what we call *image description sequences*. These multi-sentence descriptions of the contents of an image were collected in a pseudo-interactive setting, where the describer was told to describe the given image to a listener who needs to identify the image within a set of images, and who successively asks for more information. As we show, this setup produced nicely structured data that, we think, will be useful for learning models capable of planning and realising such description discourses.

1 Introduction

Talking about what one sees brings together several core competences of situated agents: to *understand* the world in terms of objects and their attributes and mutual relations, and to be able to *name* these objects, attributes, and relations, and to *compose* linguistic expressions from that, for the given addressee and under the constraints of the given communicative intention.

Many of the decisions involved in this do not only require general visual and linguistic competences, but are well-known to be affected by the task, the context and the intended addressee. Consequently, recent progress in the area of NLG, Language & Vision has been made by moving from generic settings like image captioning (Lin et al., 2014; Chen et al., 2015; Hodosh et al., 2013; Plummer et al., 2015) to *task-oriented* settings like referring expression generation (Kazemzadeh et al., 2014; Yu et al., 2016) or interactive visual question answering (Das et al., 2017; De Vries et al., 2017). As shown by Ilinykh et al. (2018), task-based image descriptions substantially differ in terms of their linguistic properties (e.g. occurrence of referring expressions, attribute types) from their "neutral" counterparts.

An orthogonal development has been to move towards longer units of text as the desired output. A few datasets exist that pair longer natural language texts (like full *paragraphs*) with single images that they are meant to describe (Krause et al., 2017; Lin et al., 2015). These constitute a challenging testbed for state-of-the-art models in NLG where common tasks from Language & Vision need to be connected to core aspects of text generation such as content selection, text structuring, or aggregation.[1] On the "interactivity" dimension, however, these datasets constitute a step back to a monological setting. While the instructions to the annotators were to imagine that they describe an image for an imagined partner, they were allowed to edit the paragraph in the usual way, thus creating something that is more akin to a text than to a task-oriented contribution to an interaction.

We present a task and a dataset that is meant to combine aspects of those mentioned above. We have collected *image description sequences*, which are sequences of expressions that collectively are meant to single out one image from an (imagined) set of other similar images. These sequences were produced in a monological setting, but with the instruction to imagine they were provided to a partner who successively asked for more information (hence, "tell me more").[2] We believe that such setting at least partially resembles dialogical interaction between humans, and, therefore, we refer to a single expression in a sequence as *a turn*. In the user interface, this sequential / incremental aspect was stressed by offering separate text input fields, rather than one block.

*Work done while at Bielefeld University.

[1] See (Gatt and Krahmer, 2018) for a survey on this traditional area in NLG.

[2] This setting is somewhat similar to that of Lin et al. (2015), who collected texts meant to describe a scene to someone who can't see it, but it is tuned even more towards (imagined) interaction. We also collected data for about 4 times as many images.

1: *This is a large bedroom with two large windows, a bed, and a two person chaise lounge.*
2: *The windows have striped curtains in front of them and a curtain rod that goes over both windows.*
3: *There is a ceiling light and fan in the center of the room.*
4: *There are two large pictures above the bed and dark colored nightstands on both sides.*
5: *There are table lights on the nightstands and several plants throughout the room.*

Figure 1: An image / description sequence pair

As the example in Figure 1 illustrates, the sequences bring together higher level summarising descriptions ("a large bedroom") with more detailed descriptions of individual objects in the scene and their relations (e.g., "a curtain rod that goes over both windows"), and they form mini-*discourses* that are cohesive (co-references, e.g. "a bed" – "the bed") and coherent (elaborations of descriptions of individual objects followed by descriptions of other objects). The sequence as a whole can be seen as providing a single fine-grained description which is delivered in *installments* (Clark, 1996).[3]

The research questions to which we aim to contribute are: How is the selection made of objects, attributes, and relations that are to be mentioned? How is the selection serialised and prioritized to form the sequence, and how are later turns in the sequence influenced by earlier ones? Ultimately, we want models derived from this dataset to also contribute to interactive description generation where parts of the sequence may come from different participants. More immediately, however, the combination of visual grounding and successive discourse planning seems already challenging.

[3]But note that this is just an approximation, for the sake of allowing for a more controlled data collection. A truly interactive setting, such as in Ilinykh et al. (2019), will turn up additional phenomena like clarification requests and corrections, from which we wanted to abstract away here.

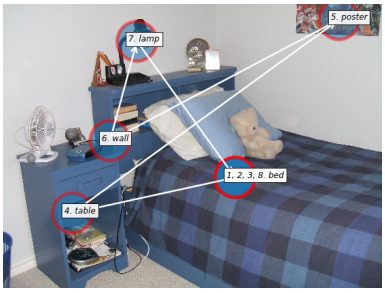
2 The Dataset

2.1 Data Collection

Images As our material on the visual side, we used a part of the ADE20k corpus (Zhou et al., 2017), which consists of images of indoor and outdoor environmental scenes that come with pixel-level object labels. We chose visual scenes as image subject matter, rather than the more event or single-object oriented settings that dominate other corpora, because scenes afford a natural high-level categorisation (e.g., "a bathroom") that triggers expectations about objects that are present (e.g., "a sink"), while at the same time still allowing for a wide variety in how they are composed (e.g., what shape or colour the sink has, what material it is made of, where it is placed). This turns the task into a fine-grained classification task, where unlike in other such settings—e.g., the CUB corpus of images of bird species, (Wah et al., 2011)—there is no single label that fully categorises the instance. To further reinforce this, we used only such images which belong to one of the 35 house-related image categories specified in the 'indoor/home or hotel' section of the SUN image hierarchy (Xiao et al., 2016), which this corpus follows; the corpus as a whole contains also more esoteric scene categories where these expectations may not hold.

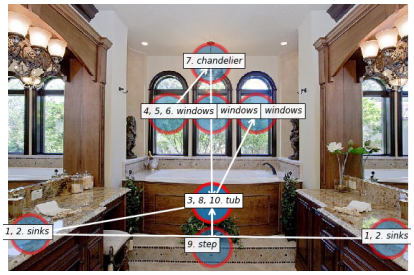
We have noticed that the first largest category ("bedroom") is oversampled with nearly twice as many images as the second largest category in each scene set; we hence reduced this to the same size as the next largest categories (bathroom, living room, kitchen). In total, we selected 4,410 images of house indoor and outdoor visual scenes, for which the corpus provides 165,088 annotated objects (for an average of 37 objects per image). The data has been divided into three disjoint subsets: 3528 images in the train set, 441 in the validation and test sets (80/10/10).

Crowd-sourcing The data collection has been conducted on Amazon Mechanical Turk (AMT). We created a task (according to AMT terminology, a HIT), in which workers were presented with an image that they could zoom into, a set of instructions, and 5 text fields in which to enter the subsequent turns. Providing separate text fields was meant to encourage the workers to indeed treat the turns as separate, and set up a small obstacle discouraging editing of earlier turns.



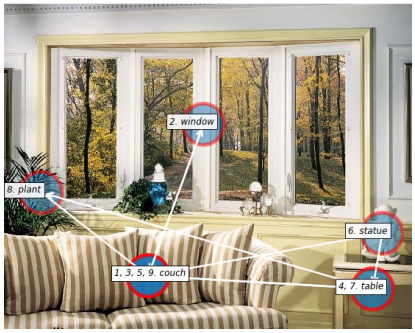
1: This is a boys bedroom with a teddy bear on the *bed*[1].
2: The *bed*[2] has a plaid blue *bed*[3] spread.
3: The head board and side *table*[4] are painted blue.
4: There is a *poster*[5] on the *wall*[6].
5: A *lamp*[7] is situated over the *bed*[8] to read with.

(a) Scene type: kids' room



1: It is a very fancy bathroom.
2: There are twin *sinks*[1, 2] across from each other.
3: There is a deep soaking *tub*[3] in front of 3 domed *windows*[4, 5, 6].
4: There is a very fancy *chandelier*[7] over the *bathtub*[8] and everything is done in brown woods and granite.
5: There is a *step*[9] up to the *bathtub*[10].

(b) Scene type: bathroom



1: there is a *couch*[1] white striped brown
2: there is a *window*[2] behind the *couch*[3] the view outside is beautiful
3: there is a side *table*[4] beside the *couch*[5]
4: there is a *statue*[6] on the side *table*[7]
5: there is a *plant*[8] behind the *couch*[9]

(c) Scene type: living room

Figure 2: Examples of visual scenes with corresponding IDS. The results of the linking are superimposed on the images, with the numbers indicating the position in the sequence and the nouns shown in italics. Nouns where a (correct) link could not be established are shown in red.

For each image, we collected a single description sequence. Table 1 in Appendix A gives the set of instructions that the workers were presented with when working on the HIT. We restricted the worker's location to English-speaking countries only and only used workers who had previously successfully completed more than 3000 HITs. For each sequence, we payed $0.15. In total, 297 workers submitted HITs with the top five participants completing over 80 tasks each. For comparison, in a separate task, we randomly chose almost 10% of the images in our subcorpus and collected traditional captions for them, using the COCO instructions (Chen et al., 2015) (at least 8 words, don't start with "there is", don't mention things that can't be seen).

General statistics Overall, we collected a sequence for each of the 4,410 images in our set, with a fixed number of 5 turns each and consequently resulting in 22,050 collected turns. A few turns contained more than one sentence, as indicated by using the nltk sentence splitter (Bird et al., 2009), yielding an average of 1.01 sentences per turn. There are 208,778 tokens altogether in the corpus, realising 5,124 token types. On average, each turn contains 9 tokens.

Preprocessing While there are various methods for unsupervised learning to ground phrases in images (e.g., Rohrbach et al., 2015; Wang et al., 2016), for now we went a simpler route and made use of the object level annotations provided by the ADE corpus. More specifically, we tried to link nouns in the sequences with labels of objects in the corpus, going first by string matches and then by similarity in a vector space (where we used the GoogleNews vectors provided by Mikolov et al. (2013)). For plural nouns, we took their singular form; if the image had a set of objects marked with the same label, a matching noun would be linked to all of them. If singular nouns matched more than one object, we selected the one with the biggest bounding box. Out of all noun candidates for linking, 30,198 nouns were linked with 45,324 annotated objects (that is, on average each noun was linked to 1.5 objects). On a manually annotated small set (44 sequences), the best linking method (testing various weightings and similarity thresholds) reached a precision of 0.77, recall 0.64, and 0.70 f-score; the numbers provided next are hence somewhat noisy. We note that our linking method ignores any adnominal dependents (determiners, adjectives, etc.), and, therefore, has no capability to resolve ambiguity between objects of the same type, but with different attributes ("red chair" vs. "blue chair"). We leave this for the future work. Some examples with the computed links are shown in Figure 2.

3 Data Analysis

Sequence Structure Looking at the validation set, we noticed that there is a characteristic struc-

ture to the sequences. Typically, the first turn provides a classification of the scene using the expected labels ("a bathroom", "a boys bedroom", as in Figure 2), with subsequent turns providing additional information about selected objects. Figure 3 shows the frequency distribution of initial 3-grams for the first turn (left) and for the remaining turns together, confirming this impression.

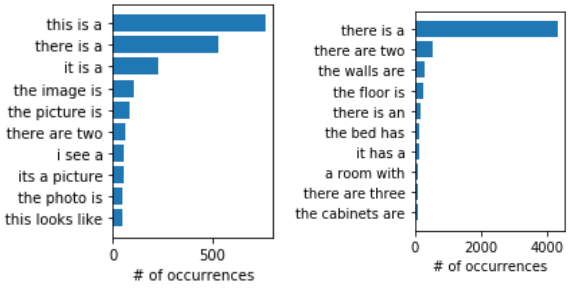


Figure 3: Frequency distribution of the first 10 trigram prefixes of the first sentences in first turn (left) compared to those in other sequence positions (right).

Our elicitation method enforced a natural split into discourse segments, by providing 5 separate text fields. We expect that the intentional structure, in the sense of Grosz and Sidner (1986), is flat, i.e. the whole sequence only serves the purpose of describing the image. The attentional structure of which object is in the center of attention, however, can be expected to be richer, as we discuss next.

What gets mentioned? Using the linking method described above, we associate on average 10 objects in each scene to nouns in the description sequence. The ADE annotation contains on average 37 objects per image. This indicates that the describers did not simply describe the images exhaustively.

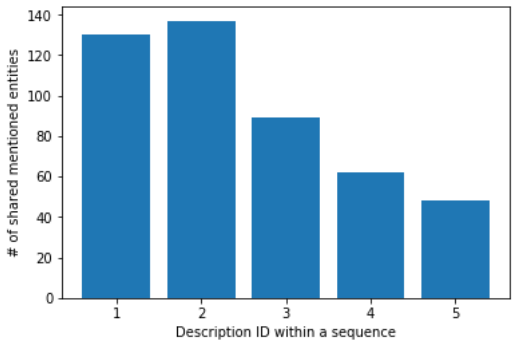


Figure 4: Position in the sequence of objects mentioned both in sequence and caption.

To investigate whether the sequences tend to mention salient objects first, we compared them to the captions that we collected for a subset of the corpus. In the collected captions, on average 3.7 links to image objects were found.

65% of all objects mentioned in all captions were also mentioned in the sequences, but only 32% of the objects mentioned in the sequences were mentioned in the captions. Figure 4 shows that objects mentioned in both caption and sequence tend to occur early in the sequence.

Co-reference We expected that this dataset would also provide interesting data for learning to co-refer, see e.g. Lin et al. (2015). The examples in Figure 2 illustrate the phenomenon. In each of the sequences, there is an object that is referred to repeatedly, and in all of them, it is the central or most salient one (the bed, the bathtub, the couch). As it should be, first mentions are indefinite NPs ("a deep soaking tub") and repeated mentions definite NPs ("the bathtub"). The exceptions are interesting as well: In Figure 2a, the definite "the bed" can be resolved as given with a bridging inference (Clark, 1977) linking it to "bedroom", and "the headboard" to "the bed". Pronominal co-reference occurs as well, with at least one pronoun, as determined via their POS tag as assigned by SpaCy (Honnibal and Montani, 2017), occurring in 42% of the sequences. We leave to future work an analysis of a larger part of the corpus in terms of the *centering* process assumed by Grosz et al. (1995) to underly local coherence.

4 Conclusions

We have presented a dataset of in-depth descriptions of images of typical domestic scene types. The description sequences were elicited in a pseudo-interactive setting under the pretense of helping an imagined addressee to do a task, namely to identify the described image within a set of similar images. We have shown that the resulting data is rich in referring expressions, and poses interesting discourse planning challenges from the perspective of natural language generation. We hope that the data will be useful for training models that can perform in actual interactive settings and can realise descriptions of scenes in *installments*, similar to previous work on collaborative reference to objects (Fang et al., 2014; Zarrieß and Schlangen, 2016). Whether that is the case remains to be seen in future work.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Herbert Clark. 1977. Bridging. In Phillip N. Johnson-Laird and P. C. Wason, editors, *Thinking: Readings in Computer Science*, pages 411–420. Cambridge University Press, Cambridge, UK.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*.

Rui Fang, Malcolm Doering, and Joyce Y Chai. 2014. Collaborative models for referring expression generation in situated dialogue. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Barbara Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Comput. Linguist.*, 12(3):175–204.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2018. The task matters: Comparing image captioning and task-based dialogical image description. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 397–402, Tilburg University, The Netherlands. Association for Computational Linguistics.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meetup! a corpus of joint activity dialogues in a visual environment. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2019 / LondonLogue)*, London, UK.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 787–798, Doha, Qatar.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Computer Vision and Patterm Recognition (CVPR)*.

Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. 2015. Generating multi-sentence natural language descriptions of indoor scenes. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 93.1–93.13. BMVA Press.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C.Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693, pages 740–755. Springer International Publishing.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*, abs/1505.04870.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2015. Grounding of textual phrases in images by reconstruction. *CoRR*, abs/1511.03745.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Mingzhe Wang, Mahmoud Azab, Noriyuki Kojima, Rada Mihalcea, and Jia Deng. 2016. Structured matching for phrase localization. In *Computer Vision – ECCV 2016*, pages 696–711, Cham. Springer International Publishing.

Jianxiong Xiao, Krista A. Ehinger, James Hays, Antonio Torralba, and Aude Oliva. 2016. Sun database: Exploring a large collection of scene categories. *Int. J. Comput. Vision*, 119(1):3–22.

L. Yu, P. Poirson, S. Yang, A.C. Berg, and Berg T.L. 2016. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, volume 9906 of *Lecture Notes in Computer Science*. Springer.

Sina Zarrieß and David Schlangen. 2016. Easy things first: Installments improve referring expression generation for objects in photographs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 610–620, Berlin, Germany. Association for Computational Linguistics.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

A Instructions

Imagine that you are talking to someone over the phone, who sees a set of images of places. What you see is one image from this set. Your partner has to pick out the image that you see out of the set that they see.
What would you say?
Imagine that is a game and the both of you want to do this as quickly as possible. There are several text fields here. Imagine that your partner can't immediately find the image, and you want to offer more information, or phrase what you've said differently, until they find the image.
We will pay you (and accept your results) only if you (a) fill our all the text fields with descriptions, (b) provide reasonable descriptions, (c) properly follow the instructions. You can get an image zoomed in by clicking on it.

Table 1: The set of instructions for the non-interactive data collection. Workers additionally saw an image and five text fields that they were supposed to fill with descriptions.