Tell Me More: A Dataset of Visual Scene Description Sequences

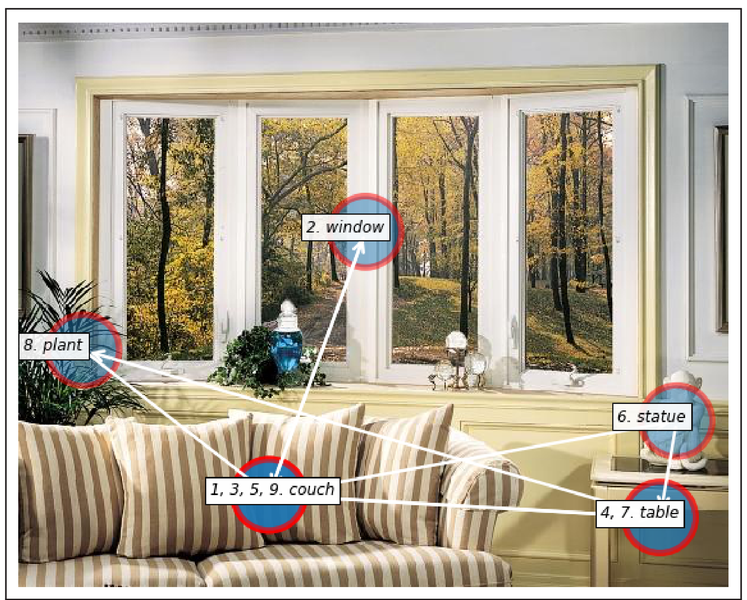
Nikolai Ilinykh, Sina Zarrieß, David Schlangen
nikolai.ilinykh@uni-bielefeld.de, sina.zarriess@uni-jena.de, david.schlangen@uni-potsdam.de

INLG 2019
29 October - 1 November 2019
Tokyo, Japan

Summary

- **Dataset:** we introduce a new dataset of **image description sequences (IDS)**, which are sequences of expressions that collectively are meant to single out one image from an (imagined) set of other similar images. We demonstrate that such descriptions form a single coherent text.
  The data is available at **https://github.com/clp-research/image-description-sequences**.
- **Data Collection Set-up:** 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"). 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*.
- **Research Questions:**
  1. How is the selection made of objects, attributes, and relations that are to be mentioned in the sequence?
  2. How is the selection serialised and prioritized to form the sequence?
  3. How are later turns in the sequence influenced by earlier ones?

Grounding Phrases in Images



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]

- **Idea:** In sequences, humans typically describe relations between objects and might refer to some objects multiple times. We link nouns in the sequences and labels of objects in the corpus via string matching / word similarity to demonstrate that (a) there is indeed some structure to the way objects are (re)introduced, (b) the mentioned objects are image-grounded.
- 30198 nouns were linked with 45324 objects (each noun is linked to 1.5 objects on average).
- The best linking method: precision **0.77**, recall **0.64**, f-score **0.70** (tested on the manually annotated dataset of 44 description sequences).
- General notation: correct links are in blue, incorrect ones are in red.

The linking method is noisy: it is not capable of resolving ambiguity between objects of the same type. However, we are still able to show that sequences are indeed ordered, using the current method.

Description Sequence

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.*

- 4410 images of house indoor / outdoor environments (ADE20k)
- 37 annotated objects per image
- 22050 individual sentence descriptions
- 1.01 sentences per turn
- 208778 tokens and 5124 token types
- 9 tokens on average per turn
- 297 unique AMT workers with top five completing over 80 tasks each

What is mentioned first?

The largest proportion of first turns provides the scene type, compared to the following turns.

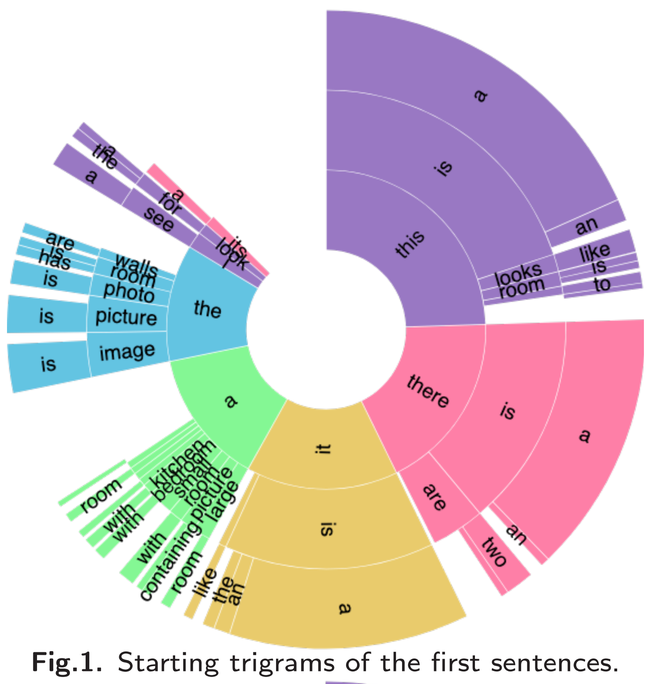


Fig.1. Starting trigrams of the first sentences.

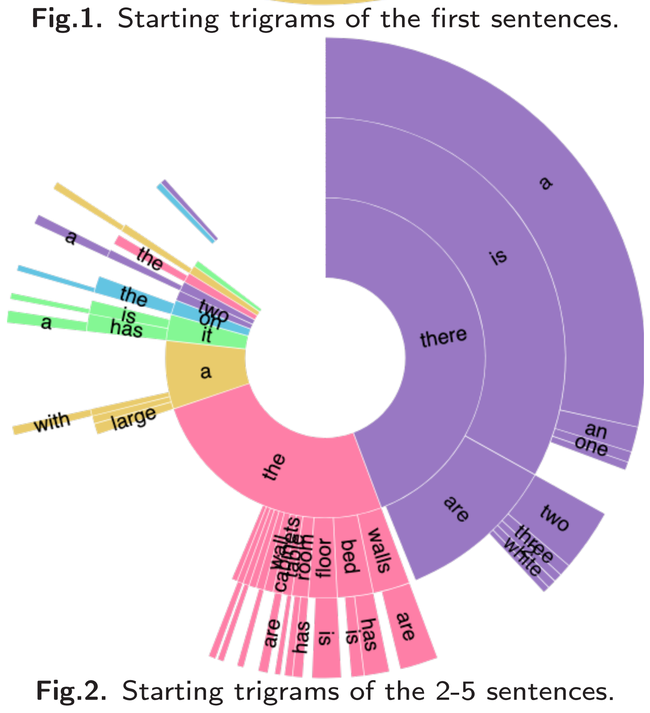


Fig.2. Starting trigrams of the 2-5 sentences.

Sequences are highly descriptive

Description sequences vs. traditional captions

Traditional captions (in the spirit of MSCOCO) were collected for almost 10% of the images (441 image, in particular):

65% of the objects mentioned in captions were also mentioned in the sequences, but conversely, only **32%** of the objects mentioned in the sequences were mentioned in the captions (that is, their coverage in terms of objects is higher).

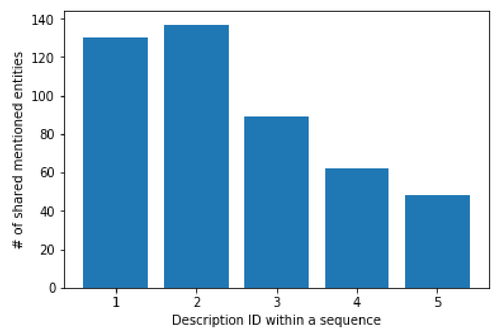
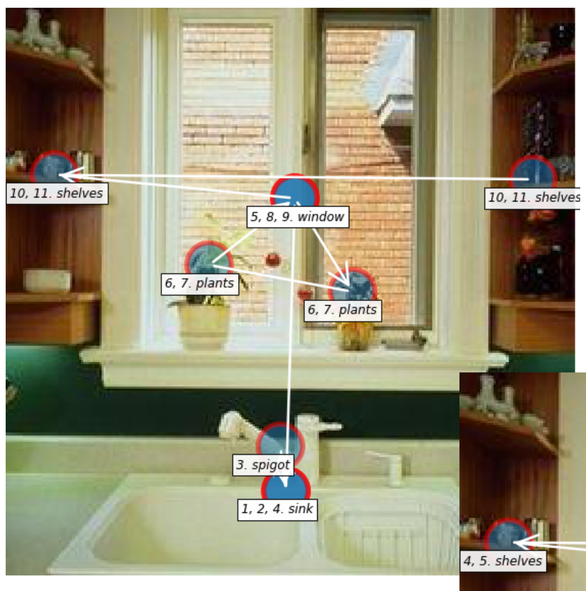


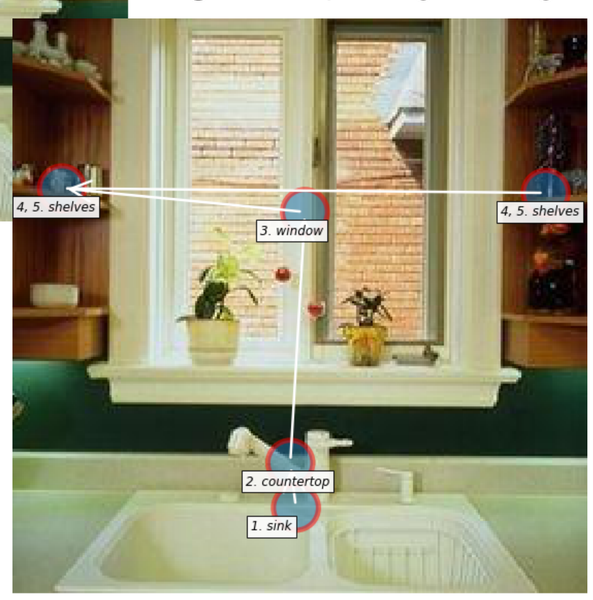
Fig.3 Sequence position of objects mentioned both in sequence and caption.



Description sequence: Picture looks like you are standing right in front of *the sink*[1]. *The sink*[2] and *the spigot*[3], blush the sprayer are all in white. Looks like all in plastic, Zero chrome. Beyond *the sink*[4] is *a window*[5] vertically divided into two panes. There are *two very small plants*[6, 7] on *the window sill*[8]. On each side of *the window*[9] are *small wooden shelves*[10, 11] in natural wood.

Fig.4 Example of grounding objects in sequences.

On average, **3.7** links to image objects were found in captions. In comparison, in sequences **10** links to image objects were established on average.



Caption: *A white sink*[1] is mounted on *a white countertop*[2] and sits under *a window*[3] with *wood shelves*[4, 5] on either side of it.

Fig.5 Example of grounding objects in captions.