Supplementary Materials: Resolving Vision and Language Ambiguities Together: Joint Segmentation & Prepositional Attachment Resolution in Captioned Scenes

Gordon Christie[1,*], Ankit Laddha[2,*], Aishwarya Agrawal[1], Stanislaw Antol[1]
Yash Goyal[1], Kevin Kochersberger[1], Dhruv Batra[1]
[1]Virginia Tech [2]Carnegie Mellon University
ankit1991laddha@gmail.com
{gordonac,aish,santol,ygoyal,kbk,dbatra}@vt.edu

In this document, we provide a full description of the datasets, including a detailed description of the dataset curation process.

Datasets. Access to rich annotated image + caption datasets is crucial for performing quantitative evaluations. Since our work is the first to study the problem of joint segmentation and PPAR, no standard datasets for this task exist so we had to curate our own annotations for PPAR on three image caption datasets – ABSTRACT-50S [1], PASCAL-50S [1] (which expands the UIUC PASCAL sentence dataset [2] from 5 captions per image to 50), and PASCAL-Context-50S [3]) (which uses the PASCAL Context image annotations and the same sentences as PASCAL-50S). Our annotations are publicly available on the authors webpages. To curate the PASCAL PPAR annotations, we first select all sentences that have preposition phrase attachment ambiguities. We then plotted the distribution of prepositions in these sentences. The top 7 prepositions are used, as there is a large drop in the frequencies beyond these. The 7 prepositions are: "on", "with", "next to", "in front of", "by", "near", and "down". We then further sampled sentences to ensure uniform distribution across prepositions. We perform a similar filtering for the ABSTRACT-50S dataset (we use top-6 prepositions). We consider a preposition to be ambiguous if there are at least 2 parsings where one of the two objects in the preposition dependency (object1, preposition, object2) is same in the parsings while the other object is different (*e.g.* (dog on couch) and (woman on couch)). To summarize the statistics of all three datasets:

1. **ABSTRACT-50S** [1]: 25,000 sentences (50 per image) with 500 images from abstract scenes made from clipart. Filtering for captions containing the top-6 prepositions resulted in 399 sentences describing 201 unique images. These 6 prepositions are: "with", 'next to", "on top of", "in front of", "behind", and "under". There are 502 total prepositions, 406 ambiguous prepositions, 80.88% ambiguity rate and 60 sentences with multiple ambiguous prepositions. The Abstract Scenes dataset [4] contains synthetic images generated by human subjects via a drag-and-drop clipart interface. The subjects are given access to a (random) subset of 56 clipart objects that can be found in park scenes, as well as two characters, Mike and Jenny, with a variety of poses and expressions. Example scenes can be found in Figure 1. The motivation is to allow researchers to focus on higher-level semantic understanding without having to deal with noisy information extraction from real images, since the entire contents of the scene are known exactly, while also providing a dense semantic space to study (due to the heavily constrained world). We used the dataset in precisely this way to first test out the PPAR module in isolation to demonstrate that this problem can be helped by a sentence's corresponding image.

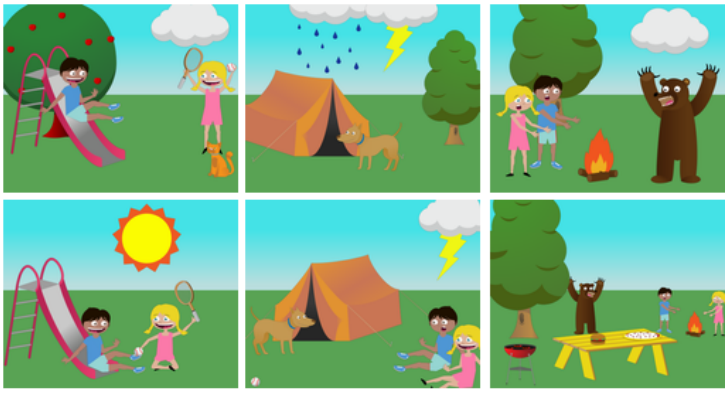

Figure 1: We show some example scenes from [4]. Each column shows two semantically similar scenes, while the different columns show the diversity of scene types.

2. **PASCAL-50S** [1]: 50,000 sentences (50 per image) for the images in the UIUC PASCAL sentence dataset [2]. Filtering for the top-7 prepositions resulted in a total of 30 unique images, and 100 image-caption pairs, where ground-truth PPAR were carefully annotated by two vision + NLP graduate students. There are 213 total prepositions, 147 ambiguous prepositions,

69.01% ambiguity rate and 35 sentences with multiple ambiguous prepositions.

3. **PASCAL-Context-50S** [3]: We use images and captions from PASCAL-50S, but with PASCAL Context segmentation annotations (60 categories instead of 21). This makes the vision task more challenging. Filtering this dataset for the top-7 prepositions resulted in a total of 966 unique images and 1822 image-caption pairs. Ground truth annotations for the PPAR were collected using Amazon Mechanical Turk. Workers were shown an image and a prepositional attachment (extracted from the corresponding parsing of the caption) as a phrase ("woman on couch"), and asked if it was correct. There are 2,540 total prepositions, 2,147 ambiguous prepositions, 84.53% ambiguity rate and 283 sentences with multiple ambiguous prepositions.

Dataset Curation and Annotation. The subsets of the PASCAL-50S and ABSTRACT-50S datasets were carefully curated by two vision + NLP graduate students. The subset of the PASCAL-Context-50S dataset was curated by Mechanical Turk workers. The following describes the dataset curation process for each datset.

1. **PASCAL-50S:** For PASCAL-50S we first obtained sentences that contain one or more of 7 prepositions (*i.e.*, "with", "next to", "on top of", "in front of", "behind", "by", and "on") that intuitively would typically depend on the relative distance between objects. Then we look for sentences that have preposition phrase attachment ambiguities, *i.e.*, sentences where the parser output has different sets of prepositions for different parsings. Due to our focus on PP attachment, we do not pay attention to other parts of the sentence parse, so the parses can change while the PP attachments remain the same, as in Figure **??** and Figure **??**. The sentences obtained are further filtered to obtain sentences in which the objects that are connected by the preposition belong to one of the 20 PASCAL object categories. Since our Module 1 is semantic segmentation and not instance-level segmentation, we restrict the dataset to sentences involving prepositions connecting two different PASCAL categories. Thus, our final dataset contains 100 sentences describing 30 unique images and contains 16 of the 20 PASCAL categories as described in the paper. We then manually annotated the ground truth PP attachments. Such manual labeling by student annotators with expertise in NLP takes a lot of time, but results in annotations that are linguistically high-quality, with any inter-human disagreement resolved by strict adherence to rules of grammar.
2. **ABSTRACT-50S:** We first obtained sentences that contain one or more

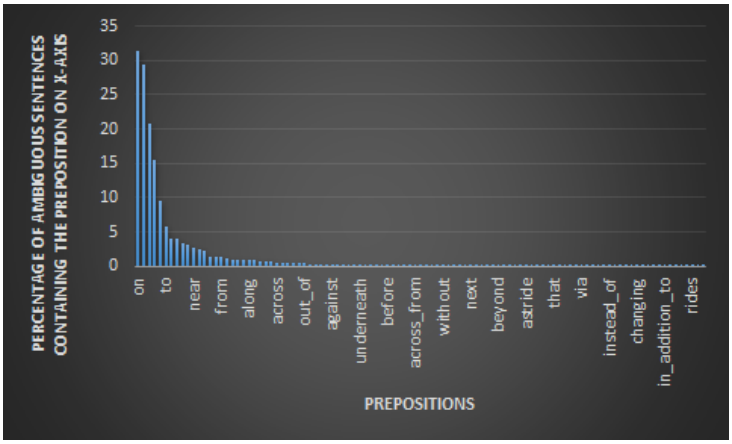

Figure 2: We show the percentage of ambiguous sentences in PASCAL-Context-50S dataset before filtering for prepositions. We found that there was a drop in the percentage of sentences for prepositions that appear in the sorted list after "down". So, for the PASCAL-Context-50S dataset we only keep sentences that have one or more visual prepositions in the list of prepositions upto "down".

of 6 prepositions (*i.e.*, "with", "next to", "on top of", "in front of", "behind", "under"). Due to the semantic differences between the datasets, not all prepositions found in one were present in the other. Further filtering on sentences was done to ensure that the sentences contain at least one preposition phrase attachment ambiguity that is between the clipart noun categories (*i.e.*, each clipart piece has a name, like "snake", that we search the sentence parsing for). This filtering reduced the original dataset of 25,000 sentences and 500 scenes to our final experiment dataset of 399 sentences and 201 scenes. We then manually annotated the ground truth PP attachments.

3. **PASCAL-Context-50S**. For PASCAL-Context-50S, we first selected all sentences that have preposition phrase attachment ambiguities. We then plotted the distribution of prepositions in these sentences (see Figure 2). We found that there was a drop in the percentage of sentences for prepositions that appear in the sorted list after "down". Therefore, we only kept sentences that have one or more 2-D visual prepositions in the list of prepositions upto "down". Thus we ended up with the following 7 prepositions: "on", "with", "next to", "in front of", "by", "near", and "down". We then further sampled sentences to ensure uniform distribution across prepositions. Unlike PASCAL-50S, we did not filter sentences based on whether

the objects connected by the prepositions belong to one of 60 PASCAL Context categories or not. Instead, we used the Word2Vec [5] similarity between the objects in the sentence and the PASCAL Context categories as one of the features. Thus, our final dataset contains 1822 sentences describing 966 unique images.

The ground truth PP attachments for these 1822 sentences were annotated by Amazon Mechanical Turk workers. For each unique prepositional relation in a sentence, we showed the workers the prepositional relation of the form primary object preposition secondary object and its associated image and sentence and asked them to specify whether the prepositional relation is correct or not correct. We also asked them to choose the third option - "Primary object/ secondary object is not a noun in the caption" in case that happened. The user interface used to collect these annotations is shown in Figure 3. We collected five answers for each prepositional relation. For evaluation, we used the majority response. We found that 87.11% of human responses agree with the majority response, indicating that even though AMT workers were not explicitly trained in rules of grammar by us, there is relatively high inter-human agreement.

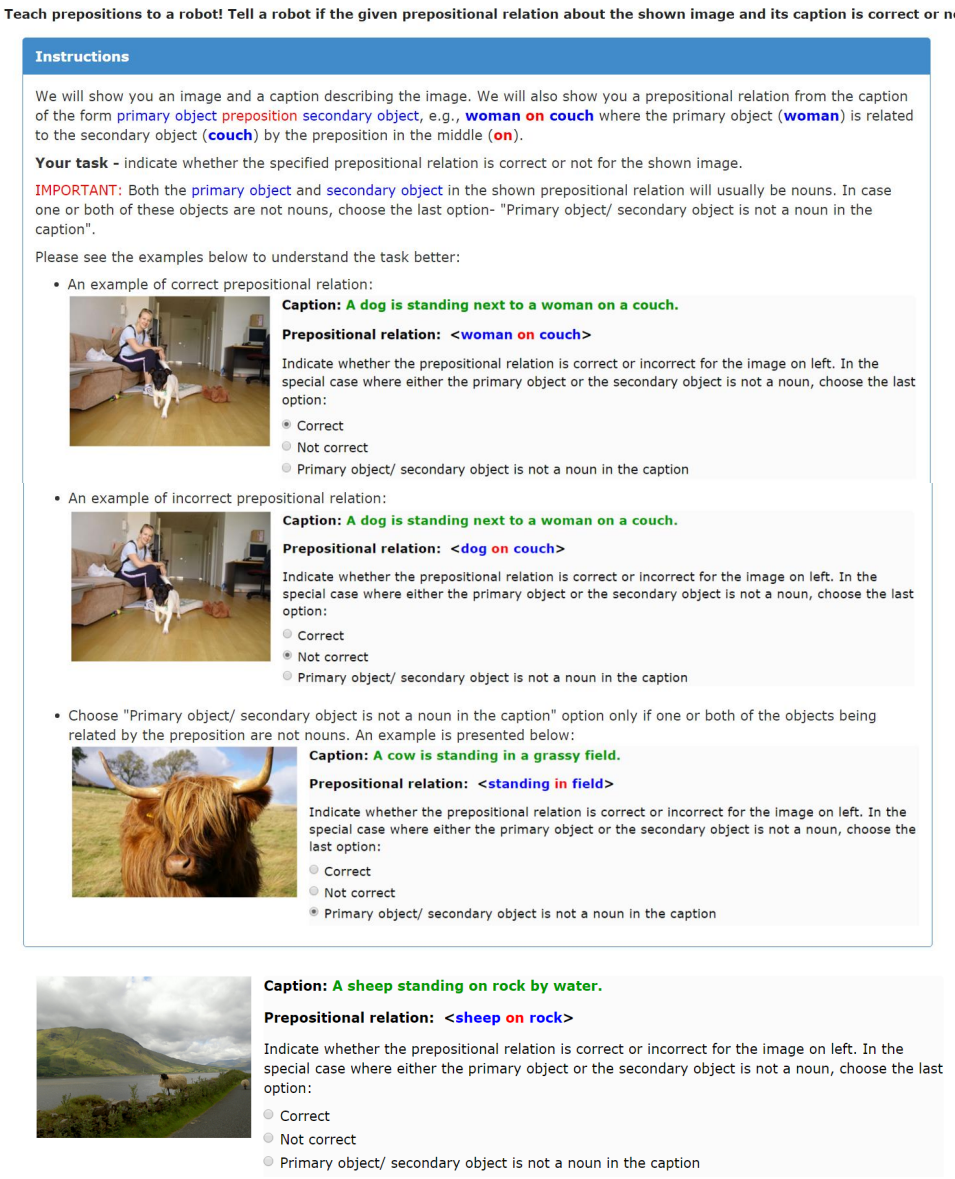

Figure 3: The AMT interface to collect ground truth annotations for prepositional relations. Five answers were collected for each prepositional relation. The majority response is used for evaluation. The AMT workers are asked to select if the preposition is correct, not correct, or that the primary or secondary object is not a noun in the caption. Examples for all three answer choices are shown in the instructions presented to the workers.

References

[1] R. Vedantam, C. L. Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: CVPR, 2014.

[2] C. Rashtchian, P. Young, M. Hodosh, J. Hockenmaier, Collecting Image Annotations using Amazon's Mechanical Turk, in: NAACL-HLT, 2010.

[3] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, A. Yuille, The role of context for object detection and semantic segmentation in the wild, in: CVPR, 2014.

[4] C. L. Zitnick, D. Parikh, Bringing Semantics Into Focus Using Visual Abstraction, in: CVPR, 2013.

[5] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: ICLR, 2013.