Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models Supplementary

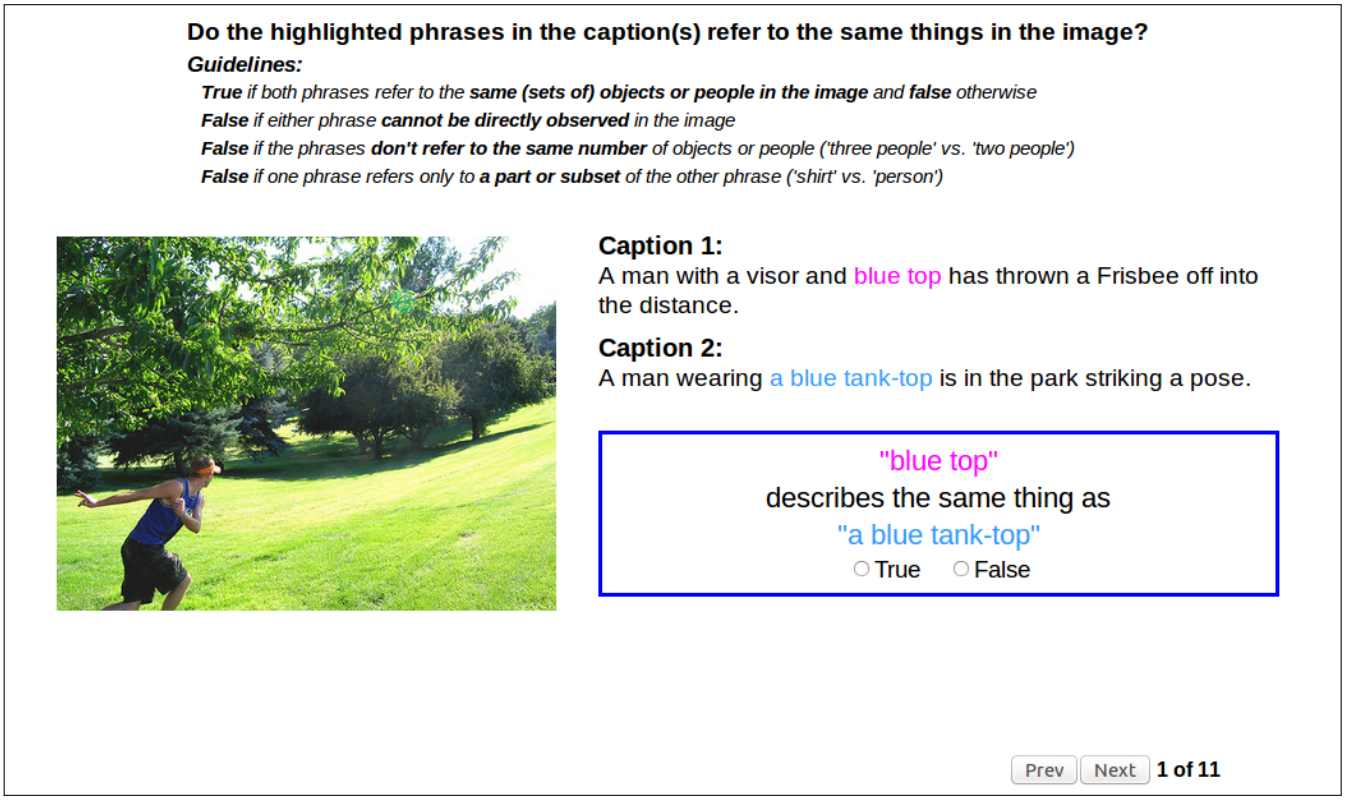
Bryan A. Plummer[†] Liwei Wang[†] Chris M. Cervantes[†] Juan C. Caicedo[*]
Julia Hockenmaier[†] Svetlana Lazebnik[†]
[†]Univ. of Illinois at Urbana-Champaign [*]Fundación Univ. Konrad Lorenz

Contents

1. Annotation Task Guidelines

For each task in our annotation pipeline described in Section 2 of the paper, we provided a set of guidelines which we displayed with each question. Along with these guidelines we provided a link to examples showing a worker questions, the expected answer, and explanation of how the guidelines were being applied. Below is a screenshot of the interface for each qualifying task showing how each task was presented along with the guidelines provided to workers.

Binary Coreference Link Annotation Interface



Coreference Chain Verification Interface

Determine if all the phrases are associated to the same thing. Each color highlights a different phrase

Guidelines:

*Select **True** when **all** highlighted phrases are related to the **same concept or object**.*
*Select **False** if you find **at least one** highlighted phrase that **does not correspond** to the same thing as the others.*
*Read all captions to make sure that you **understand the context** of each highlighted phrase.*

Image Captions:

- An adult riding **a bike** on a beach with many visible vapour trails in the sky.
- A person rides **his bicycle** in the sand beside the ocean.
- A man riding **his bike** on a beach by the ocean.
- A man rides **a bike** under a blue and white sky.
- A man on **a bicycle** riding on a beach.

Do all highlighted phrases refer to the same thing? ○ True ○ False

Prev Next

1 of 10 images

Add Comments (Optional)

Box Requirement Interface

Answer questions about a phrase in an image caption.

Guidelines:

*Select a **box can be drawn** if there are specific location(s) of an image the phrase refers to*

*Select **scene or place** when the phrase doesn't refer to something specific in the image, but rather the image as a whole*

*Select **no box** if what the phrase refers to cannot be seen in the image*

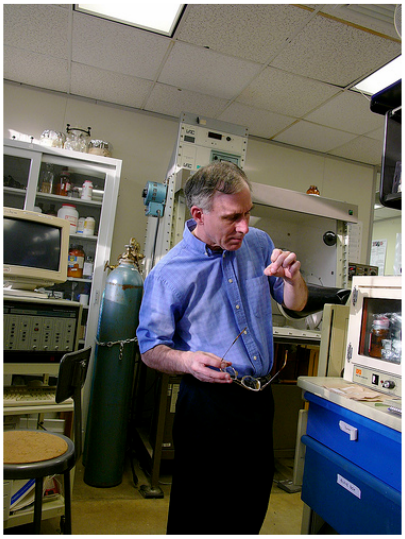


Image Caption:
A gray and black-haired male is holding his glasses in one hand while looking at something in the other hand ; surrounded by numerous amounts of machines.

For the phrase "one hand":

- At least one box can be drawn
- Refers to a scene or place
- No box can be drawn

Prev Next

1 of 7 annotations.

Add Comments (Optional)

Box Drawing Interface

Draw a bounding box for a phrase in an image caption.

Guidelines:

*Include **all visible parts** and draw as **tightly as possible***

*If there are multiple instances that **can** be easily separated, pick **only one** to draw (any one)*

*If there are multiple instances that **cannot** be easily separated, draw **one box** for the group*

***Do not draw** a box for instances that already have boxes*

Image Caption:
A black and white dog standing near a man.

Draw a bounding box for:
"A black and white dog"

Add Box | Hide Prev Boxes | Delete Box

Check here if the phrase can't be observed in the image or if every instance already has a bounding box

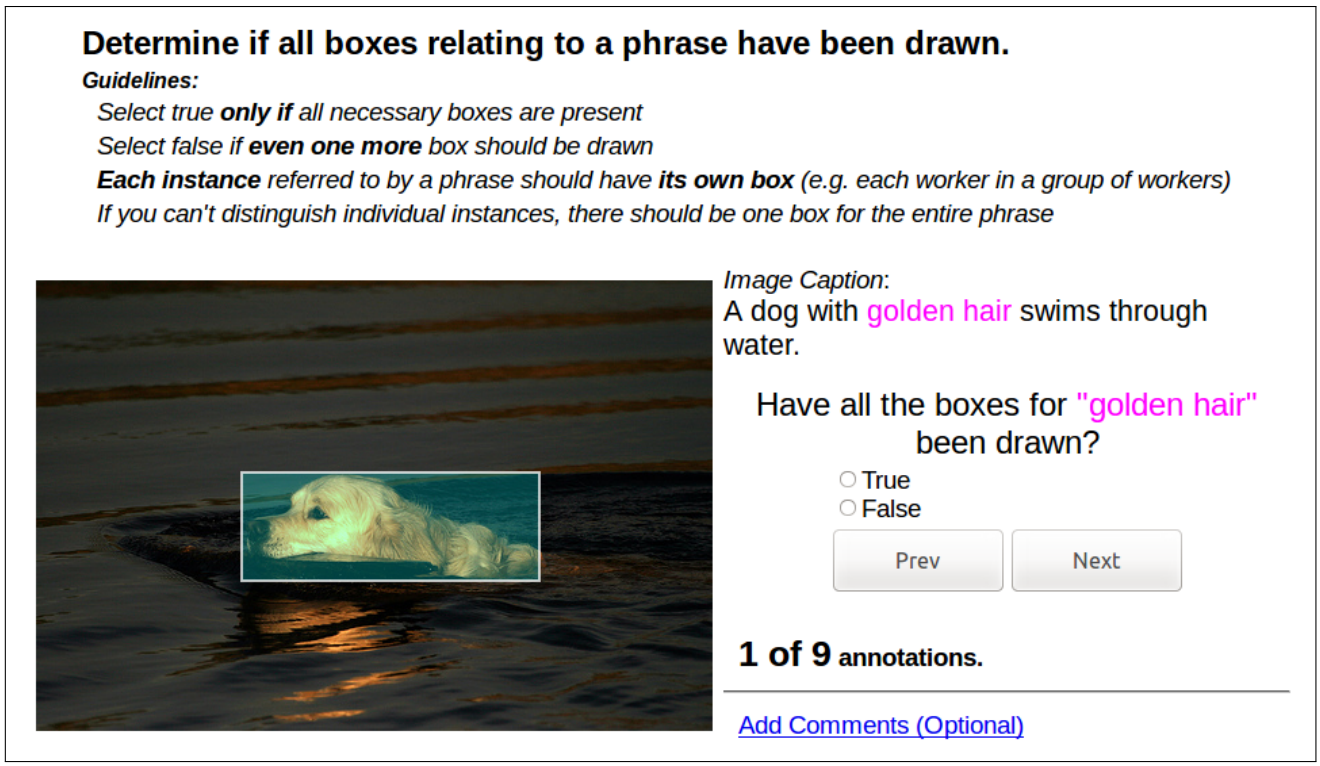
Prev | Done | Next

1 of 9 annotations.

Add Comments (Optional)

Box Quality Interface

Box Coverage Interface



2. Crowdsourcing Statistics

Here we provide further details about our annotation process described in Section 2 of the paper and the performance of the workers on each task.

	Avg Time (s)	Annos per Task	Min Performance	Avg Worker Quality	% Rejected	Num Workers
Coreference Links	75	10	80%	90.6%*	2*	587
Coreference Verify	95	5	83%	90.6%*	2*	239
Box Requirement	81	10	83%	88.4%	< 1	684
Box Drawing	134	5	70%	82.4%	38.3	334
Box Quality	110	10	78%	88.0%	52.7	347
Box Coverage	91	10	80%	89.2%	35.4	624

*combined

Table 1: Per task crowdsourcing statistics about our annotation process. Average Worker Quality was computing using the average accuracy of workers on verification questions (or approved annotations in the Box Drawing task). Min Performance is the Worker Quality score a worker must maintain to remain approved to do our tasks.

3. Trusted Workers vs. Post Hoc Verification

In this section we provide additional discussion into the motivation behind using Trusted Workers which is described in Section 2.3 of the paper.

Initially we attempted to use verification questions (questions for which we know the answers) to filter out good annotations post hoc. Rather than tasks containing 2% verification questions, they contained 20% verification questions, and were evaluated on a per worker basis in batches. While this process produces satisfactory results for the first three steps of the annotation pipeline (Conreference Links, Coreference Verify, and Box Requirement), we were not able to successfully apply this model to the last three steps.

This appears to be due, in part, to the relative difficulty and attention to details required in the steps relating to box drawing. Not only does someone have to read and understand the sentence and how it relates to the image being annotated, but must also be careful about the placement of the boxes being drawn in the last three steps. This increased difficulty led to a much smaller portion of workers successfully completing the tasks (see rejection rates in Table 1). Even our attempts to change the qualification task to be more detailed had little effect on worker performance. In doing so, a post hoc evaluation of responses to these tasks would lead to either higher costs (if you were to pay workers for poorly completed tasks) or greatly reduced completion rates for a batch of annotations in tasks proving difficult for workers (due to workers not wanting to risk doing a task they may not get paid for).

By using a list of Trusted Workers to pre-filter who can do our tasks, we not only were able to limit monetary cost of poorly performing workers, but also increased the annotation completion rate for each of our tasks. This model was also cheaper due to fewer verification questions being embedded in each task.

4. Dataset Statistics

This section extends Section 2.4 of the paper to provide additional insight into the makeup of the Flickr30k Entities dataset.

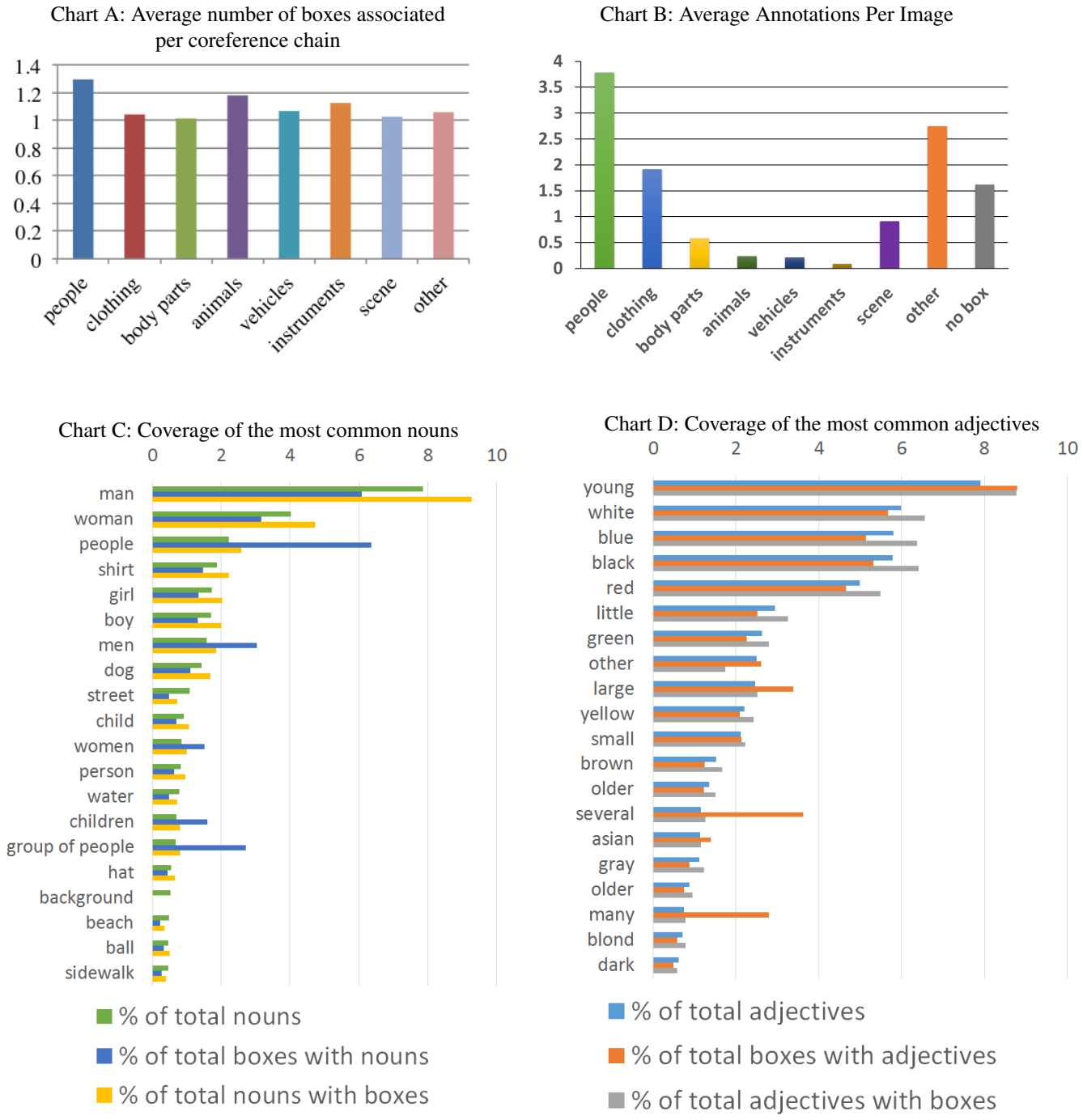


Figure 1: Analysis of the Flickr30k Entities dataset. **Chart A** shows the average number of boxes each coreference chain is associated with. **Chart B** shows average number of annotations (not including coreference links) per image by entity type. **Chart C** shows the coverage of nouns and associated boxes across the dataset. **Chart D** shows the coverage of adjectives and associated boxes across the dataset.

5. Text-to-Image Reference Resolution Evaluation Metrics

Performance was measured in two ways: recall@K and average precision . Success is achieved when a prediction has an intersection over union of at least 0.5 with the ground truth box. For each input sentence and image, the sentence was parsed to identify noun phrases. Each parsed noun phrase was then compared to the ground truth phrases, and if an exact match was present the CCA model was used to rank Edge Box Proposals. Then for each of the coarse category we computed average precision using the PASCAL method of evaluation. Recall@K was also computed for each category, and overall recall was computed by summing over the total number of successes for each category by the total number of ground truth pairings. More formally, where C is the set of M ground truth phrases in a coarse category, bb_p are the ranked list of proposals for phrase p, and $PredictionSuccess(bb_p, K)$ returns 1 if there is a successful detection within the top K proposals (no parsed phrase match always returns 0), then, $\forall p \in C$,

$$R_K = \frac{\sum_{i=1}^{M} PredictionSuccess(bb_{p_i}, K)}{M} \tag{1}$$

When computing overall recall@K we have,

$$OverallR_K = \frac{\sum_{j=1}^{C_j} \sum_{i=1}^{M_j} PredictionSuccess(bb_{p_{j,i}}, K)}{\sum_{j=1}^{C_j} M_j} \tag{2}$$

It is important to note that, although relatively uncommon, some phrases belong to multiple coarse categories and are double counted using this category based evaluation.