# Supplementary material: Who are you referring to? Coreference resolution in image narrations

Arushi Goel[1], Basura Fernando[2], Frank Keller[1], and Hakan Bilen[1]
[1]School of Informatics, University of Edinburgh, UK
[2]CFAR, IHPC, A*STAR, Singapore

## 1. Annotation Details

**Localized Narratives dataset.** Tuset *et al*. [11] proposed the Localized Narratives dataset, a new form of multimodal image annotations connecting vision and language. In particular, the annotators describe an image with their voice while simultaneously hovering their mouse over the region they are describing. Hence, each image is described with a natural language description attending to different regions of the image. In addition to textual descriptions (obtained using speech-to-text conversion), they additionally provide mouse traces for the words.

The Localized Narratives dataset is built on top of COCO [7], Flickr30k [10], ADE20k [14] and Open Images [6]. The statistics of the individual datasets are shown in Table 1.

| Localized Narratives Subsets [11] | #images | #captions | #words/capt. |
|---|---|---|---|
| COCO | 123,287 | 142,845 | 41.8 |
| Flickr30k | 31,783 | 32,578 | 57.1 |
| ADE20k | 22,210 | 22,529 | 43.0 |
| Open Images | 671,469 | 675,155 | 34.2 |

Table 1: Statistics of Localized Narratives for COCO, Flickr30k, ADE20k, and Open Images.

**Annotation tool and analysis.** We develop an HTML-based interface on the Label Studio annotation tool [1]. Figure 1 shows the annotation interface from Label Studio. We hired 6 high-quality annotators (all from computer science background) for an average of 54 hours of annotation time. The annotators were trained with the exact description of the task and given a pilot study before proceeding with the complete annotations. The pilot study was useful to correct and retrain annotators if needed. As shown in Figure 1, the annotators had to select a mention in the caption with a given label (C1, C2, etc.) in Step 1 and draw a bounding box in the image for the selected mention in Step 2 (with the same label).

For Step 1, if the mention is coreferring then it is selected with the same label to define coreference chains. It is important to note that the captions are pre-marked with noun phrases parsed from [2]. The annotators are instructed to correct the phrases if they are wrong (*e.g*. for a mention glass windows, the parser parses *glass* and *windows* as two different mentions rather than belonging to the same label/cluster) and remove the phrases that do not correspond to a region in the image.

In Step 2, if there are plural mentions such as *two men*, we ask the annotators to draw two separate bounding boxes for this. In the case of mentions such as *several people*, if the people are less than five, they are instructed to draw separate bounding boxes otherwise a group bounding box (covering all the people).

Given the challenging nature of the task, we doubly annotate 30 images with coreference chains and bounding boxes to compute the inter-annotator agreement. More specifically, for the coreference chain we compute *Exact Match* which denotes whether the coreference chains annotated by the two annotators are the same. We get an exact match of 79.9% in the coreference chains, which is a high agreement given the complexity of the task. For the bounding box localization, we compute the Intersection over Union (IoU) to compute the overlap between the two annotations. It is considered to be correct/matching if the IoU is above 0.6. We achieve bounding box accuracy of 81% on this subset of images. This analysis shows good agreement between the annotators given the subjective nature and complexity of the task.

**Coreferenced Image Narratives dataset.** In total, we annotate all the 1000 test images and 880 validation images (out of 1000) in the Flickr30k dataset. The text descriptions from the Localized Narratives dataset are very noisy with a lot of words/sequence of words. We manually filter phrases such as - *in this image, in the front, in the background, we can see, i can see, in this picture*. If there are some other mentions that are pre-marked and not filtered, we ask the annotators explicitly to filter them out. By doing this, we make sure that the dataset is clear of any unnecessary and noisy mentions.


Figure 1: Annotation interface from Label Studio.

All the words that are marked as mentions and are not noun phrases (as detected by the part of speech tagger [2]) are considered as pronouns *e.g*. *them, they, their, this, that, which, those, it, who, he, she, her, him, its*.

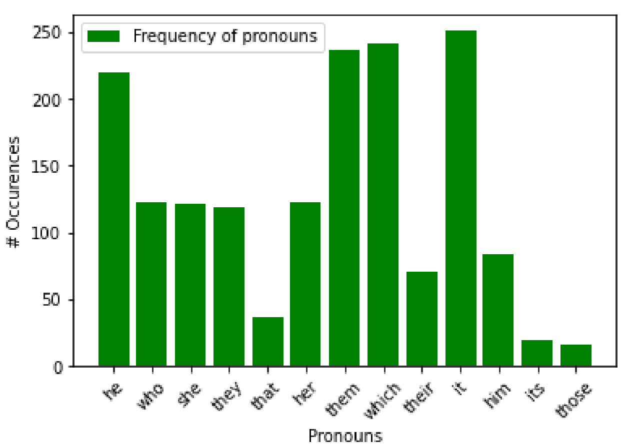



Figure 2: Total number of occurrences of pronouns in Coreferenced Image Narratives .

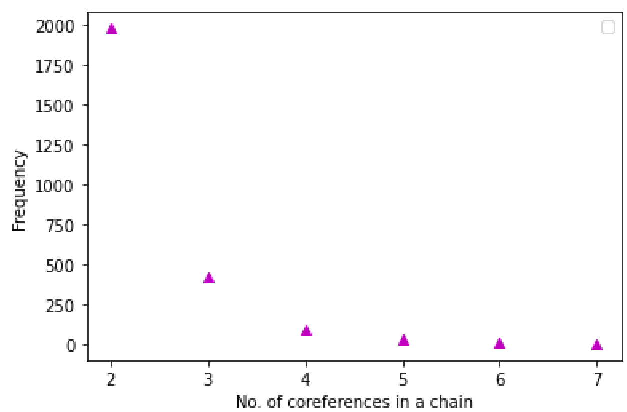



Figure 3: Number of coreference chains with 2 or more than 2 mentions in a chain in Coreferenced Image Narratives .

**Statistics for the Coreferenced Image Narratives .** In Figure 2, we show the statistics for the frequency of pronouns in the dataset. Few pronouns (*e.g*. he, it, them) are more frequent than the others. Overall, the occurrence of pronouns is frequent to conduct a fair evaluation of the coreference based models. Similarly in Figure 3, we evaluate how many mentions occur in the coreference chains.

Coreference chains with 2 and 3 mentions have a very high frequency in the dataset. There are few chains that have longer mentions (*e.g.* 6 and 7). Hence, we can safely conclude that the dataset is a powerful tool to evaluate coreference chains and learn complex coreferencing and grounding models. Moreover, the average length of the mentions (excluding pronouns) is 1.93.

## 2. Evaluation Metrics

In this section, we discuss in detail the evaluation metrics used for CR and narrative grounding. For CR, we use the MUC and the BLANC metrics, which are discussed below.

*(a) MUC F-measure.* It measures the number of coreference links (pairs of mentions) common to the predicted $R$ and ground-truth chains $K$. It involves computing the partitions with respect to the two chains:

$$\text{MUC-R} = \frac{\sum_{i=1}^{N_k}(|K_i| - |p(K_i)|)}{\sum_{i=1}^{N_k}(|K_i| - 1)}, \tag{1}$$

$$\text{MUC-P} = \frac{\sum_{i=1}^{N_r}(|R_i| - |p^{'}(R_i)|)}{\sum_{i=1}^{N_r}(|R_i| - 1)} \tag{2}$$

where $K_i$ is the $i^{th}$ ground-truth chain and $p(K_i)$ is the set of partitions created by intersecting $K_i$ with the output chains; $R_i$ is the $i^{th}$ output chain and $p^{'}(R_i)$ is the set of partitions created by intersecting $R_i$ with the ground-truth chains; and $N_k$ and $N_r$ are the total number of ground-truth and output chains, respectively.

*(b) BLANC.* Let $C_k$ and $C_r$ be the pairs of coreference links respectively, and $N_k$ and $N_r$ be the set of non-coreference links in the ground-truth and output respectively. The BLANC Precision and Recall for coreference links are calculated as follows:

$R_c = \frac{|C_k \cup C_r|}{|C_k|}$ and $P_c = \frac{|C_k \cup C_r|}{|C_r|}$, where $R_c$ and $P_c$ are the recall and precision respectively.

Similarly, recall $R_n$ and precision $P_n$ for non-coreference links ($N_k$ and $N_r$) are computed. The overall precision and recall are:

BLANC-R $= \frac{(R_c+R_n)}{2}$ and BLANC-P $= \frac{(P_c+P_n)}{2}$, respectively.

For evaluating narrative grounding in images, we consider a prediction to be correct if the IoU (Intersection over Union) score between the predicted bounding box and the ground truth box is larger than 0.5 [13, 4]. Following [5], if there are phrases with multiple ground truth boxes (*e.g.* several people), we use the any-box protocol *i.e.*, if any ground truth bounding box overlaps the predicted bounding box, it is a correct prediction. We report percentage accuracy for evaluating narrative grounding.

## 3. Implementation details

**Inputs and modules.** For the image modeling, we extract bounding box regions, visual features, and object class labels using the Faster-RCNN object detector [12]. We use Glove embeddings [9] to encode the object class labels and the mentions from the textual branch. For the mouse traces, we follow [11] and extract the trace for each word in the sentence and then convert it into bounding box coordinates for the initial representation. All the modules *i.e.*, image encoder, text encoder, trace encoder, and joint text-trace encoder are a stack of two transformer encoder layers. Each transformer encoder layer includes a multi-head self-attention layer and an FFN. There are two heads in the multi-head attention layer, and two FC layers followed by ReLU activation layers in the FFN. The output channel dimensions of these two FC layers are 2048 and 1024, respectively. The input to the joint text-trace encoder comes from the separate text and trace encoder branches. We add a special embedding to the learned embeddings following [3] to distinguish between the two modalities (text and trace) in the transformer encoder.

**Training details.** The whole architecture is trained end-to-end with the AdamW [8] optimizer. We train the transformer encoders with the learning rate of 3e-5, batch size of eight, weight decay of 0.01 and the loss coefficient $\lambda$ of 0.001. We train the model for 60 epochs and choose the best performing model based on the validation set.

## 4. Zero-shot results on Flickr30k dataset [10]

| Method | zs-MUC-R | zs-MUC-P | zs-MUC-F1 | zs-Grounding Acc. (%) |
|---|---|---|---|---|
| VinVL | 59.16 | 60.78 | 57.24 | - |
| MAF$^{\dagger}$ | 61.97 | 68.46 | 63.91 | 57.1 |
| Ours (w/o MT) | 70.11 | 68.67 | 68.48 | 59.4 |

Table 2: Zero-shot performance on the Flickr30k entities dataset.

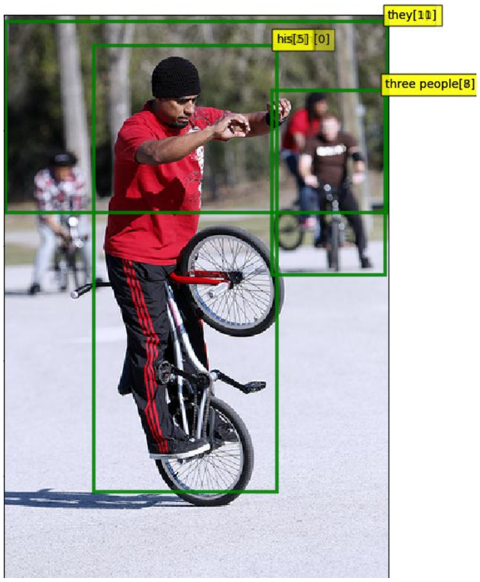
In Tab. 2, we evaluate our model and baselines using the zero-shot setting on the Flickr30k entities dataset [10] for CR and grounding. These results indicate that our method better generalizes to unseen CR chains and narrative grounding than the baselines.

## 5. Additional Qualitative Results

In Fig. 4, we show additional qualitative results from our proposed method. The model correctly chains mentions and grounds them to the correct entities in the image even for complex and ambiguous cases. Our model finds coreferences for people (*e.g. [a man, his]*) or for objects (*e.g. [a barbecue grill, it]*). Moreover, it also finds links for plurals such as *[two men, them]*. There is a huge potential in

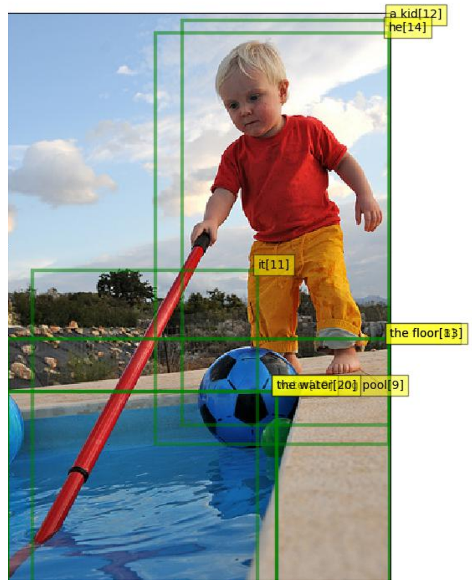
**Narration:** in this picture i can see **a man[0]** doing stunts with **a bicycle[1]**, **he[2]** is wearing **a cap[3]** on **his[5] head[4]**. i can see **three people[6-8]** in the back, **they[9-11]** are riding **bicycles[12]**. i can see **the ground[13]** at the bottom and **the trees[14]** in the background and **it[15]** looks like **grass[16]** on **the ground[17]** in the back.

**Predicted Coreference Chains:** [**a man[0], he[2], his[5]**], [**three people[6-8], they[9-11]**]



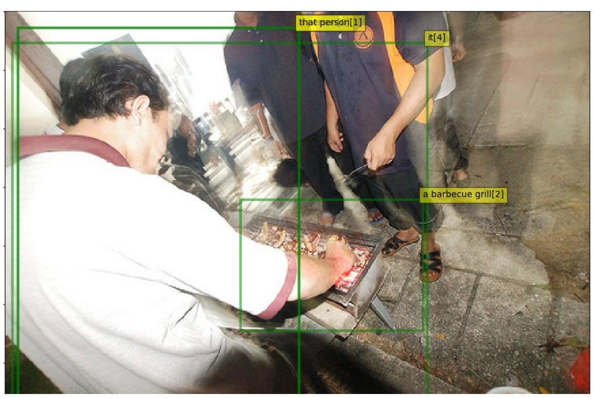

**Narration:** this image is taken outdoors. at the top of the image there is **sky with clouds[1]**. in the background we can see there are **many plants[2]** and **trees[3]**. we can see **the mesh[4]**. there are **many rocks[5]**. at the bottom of the image there is **the floor[6]**. we can see **the swimming pool[9]** with **water[10]** in **it[11]**. in the middle of the image **a kid[12]** is standing on **the floor[13]** and **he[14]** is holding **a stick[15]** in **the hand[16]** and playing. we can see **the balls[17]** in **the water[20]**.

**Predicted Coreference Chains:** [**a kid[12], he[14]**], [**the swimming pool[9], water[10], it[11], the water[20]**] , [**the floor[6], the floor[13]**]




**Narration:** on the left side of the image there is **a person[0]**. in front of **that person[1]** there is **a barbecue grill[2]** with a **food item[3]** on **it[4]**. and there are **few people[5]** standing. this is an edited image. and there is a blur background. and there are few other things in the background.

**Predicted Coreference Chains:** [**a person[0], that person[1]**], [**a barbecue grill[2], it[4]**]




**Narration:** in front of the picture, we see **two men[0]**. **the man[2]** on the left side is wearing the **spectacles[3]** and **he[4]** is trying to talk something. **the man[5]** on the right side is wearing the **goggles[6]** and **an orange cap[7]**. **it[8]** looks like **a man[9]** is holding **a wooden stick[10]**. behind **them[11-12]**, we see **the people[13]** and **some of them[14]** are wearing **the orange color caps[15]**. this picture is blurred in the background.

**Predicted Coreference Chains:** [**the man[2], he[4], a man[9]**], [**two men[0], them[11-12]**]


Figure 4: Additional qualitative results for coreference chains. For each image, we show the predicted coreference chain (mentions more than 2) and the grounding results for the corresponding mentions in the chain. The colored mentions in the descriptions are the ground-truth coreference chains.

learning to disambiguate the mentions in the descriptions and this work paves the way for future research.

## References

[1] Labelstudio. https://labelstud.io/. 1

[2] Spacy. https://spacy.io/. 1, 2

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 3

[4] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 3

[5] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 3

[6] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[9] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 3

[10] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1, 3

[11] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European conference on computer vision*, pages 647–664. Springer, 2020. 1, 3

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[13] Qinxin Wang, Hao Tan, Sheng Shen, Michael W Mahoney, and Zhewei Yao. Maf: Multimodal alignment framework for weakly-supervised phrase grounding. *arXiv preprint arXiv:2010.05379*, 2020. 3

[14] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1