Who are you referring to?

Citation for published version:

Digital Object Identifier (DOI):
10.1109/ICCV51070.2023.01399

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
2023 IEEE/CVF International Conference on Computer Vision (ICCV)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Who are you referring to? Coreference resolution in image narrations

Arushi Goel\textsuperscript{1}, Basura Fernando\textsuperscript{2}, Frank Keller\textsuperscript{1}, and Hakan Bilen\textsuperscript{1}

\textsuperscript{1}School of Informatics, University of Edinburgh, UK
\textsuperscript{2}CFAR, IHPC, A*STAR, Singapore.

Abstract

Coreference resolution aims to identify words and phrases which refer to same entity in a text, a core task in natural language processing. In this paper, we extend this task to resolving coreferences in long-form narrations of visual scenes. First we introduce a new dataset with annotated coreference chains and their bounding boxes, as most existing image-text datasets only contain short sentences without coreferring expressions or labeled chains. We propose a new technique that learns to identify coreference chains using weak supervision, only from image-text pairs and a regularization using prior linguistic knowledge. Our model yields large performance gains over several strong baselines in resolving coreferences. We also show that coreference resolution helps improving grounding narratives in images.

1. Introduction

Consider the image paired with the long-form description in Figure 1, an example from the Localized Narratives [48]. Can you tell whether the woman who is wearing spectacles refers to a person or another woman in the text? We are remarkably good in identifying referring expressions (or mentions) and determining which of them corefer to the same entity, a task that we regularly perform when we read text or engage in conversation. The text-only version of this problem is known as coreference resolution (CR) [29, 30, 57], a core task in natural language processing (NLP) with a large literature. Most text-only CR methods are either rule-based [29, 50] using heuristics such as pronoun match or exact match based on part of speech tagging, or are learned on large annotated text datasets from domains such as news text or Wikipedia articles [5, 31, 30, 21]. State-of-the-art methods [30, 21] fail to resolve coreferences correctly in image narrations for few reasons. First, CR in image narrations often require image understanding (see Fig. 1). Neural networks trained on text datasets [49, 9] suffer from poor transferability and a significant performance drop when applied to image narrations because of domain shift. Image narrations are unstructured and can be noisy, unlike the well-edited text used during training (such as news or Wikipedia). Moreover, standard image-text datasets [34, 26, 8, 47] only contain short descriptions with very few or no cases of coreference, thus, are not suitable for training text-only CR models.

Some prior work have looked at visual CR for specific tasks. [51] and [54] link character mentions in TV shows or movie descriptions to character occurrences in videos. More recently, the Who’s Waldo dataset [13] links person names in the caption to their occurrence in the image. However, these methods rely on a limited set of object categories and referring expression types (see Table 2 discussed below), require annotated training data and therefore cannot
be applied to long-form unconstrained image narrations that include open-world object categories and multiple types of referring expressions such as pronouns (she), common nouns (another woman), or proper nouns (Peter).

In this paper, we look at the problem of CR in image narrations, i.e., resolving the coreference of mentions in narrative text that is paired with an image. As the prior benchmarks in this domain are limited to either a small vocabulary of objects or specific referring expression types, we introduce a new dataset, Coreferenced Image Narratives, CIN, which augments the rich long-form narrations in the existing Localized Narratives dataset [48]. We add coreference box annotations and ground each chain by linking it to a bounding box in the corresponding image.

Manually annotating the whole dataset [48] is expensive, hence these annotations are provided only for evaluation and are not available for training. To cope with this setting, we propose a weakly supervised CR method that learns to predict coreference chains from only paired image-text data. Our key idea is to learn the linking of the mentions to image regions in a joint multi-modal embedding space and use the links to form coreference chains during training. To this end, we propose a multimodal pipeline that represents each modality (image regions, text mentions and also mouse traces, additionally provided by [48]) with a modality-specific encoder and then exploit the cross-modal correlations between them to resolve coreference. Finally, inspired from the rule-based CR [29], we incorporate linguistic rules to our learning formulation in a principled way. We report extensive experiments on CIN and demonstrate that our method not only brings significant improvements in CR but also large gains in weakly supervised narrative grounding, a form of disambiguation that has been underexplored in visual grounding.

To summarize our contributions, we introduce (1) the new task of resolving coreferences in multimodal long form textual descriptions (narrations), (2) a new dataset, CIN, that enables the evaluation of coreference chains in text and the localization of bounding boxes in images, which is provided with multiple baselines and detailed analysis for future work, (3) a new method that learns to resolve coreferences while jointly grounding them from weak supervision and exploiting linguistic knowledge, (4) a rigorous experimental evaluation showing significant improvement over the prior work not only in CR but also in weakly supervised grounding of complex phrases in narrative text.

2. Related Work

Text-only CR in NLP has a long history of rule-based and machine learning-based approaches. Early methods [20, 50] used hand-engineered rules to parse dependency trees, which outperformed all learning-based methods at the time. Recently, neural network methods [62, 61, 12, 21, 30] have achieved significant performance gains. The key idea is to identify all mentions in a document using a parser and then learn a distribution over all the possible antecedents for each mention. SpanBERT [21] uses a span-based masked prediction objective for pre-training and shows improvements on the downstream task of CR. Stolfo et al. [56], on the other hand, transfer the pretrained representations using rules for CR. It is worth noting that all these learning-based approaches either require large pretraining data or training data annotated with gold standard (ground-truth) coreference chains, such as OntoNotes [49] or PreCo [9].

Visual CR includes learning to associate people or characters mentioned in the text with images or videos [51, 54, 13]. Kong et al. [24] exploit CR to relate texts to 3D scenes. Another direction is to resolve coreferences in visual dialog [25] for developing better question-answering systems. Unlike these works, we focus on learning coreferences from long unconstrained image narrations using weak supervision. A related group of work [64, 66, 33, 16] aims to ground phrases in image parts. In visual phrase grounding [64, 35, 10, 65, 16, 22, 32], the main objective is to localize a single image region given a textual query. These models are trained on visual grounding datasets such as ReferItGame [23], Flickr30K Entities [47], or RefCOCO [65]. However, due to short captions, the grounding of text boils down to mostly salient objects in images. In contrast, grounding narrations which aims at capturing all image regions is significantly more challenging and cannot be effectively solved with those prior methods.

Weakly supervised grounding learning to ground from image-text pairs only, has recently been used in [37, 39, 38, 36, 59] for referring expression grounding. These methods use phrase reconstruction from visual region features as a training signal. Other methods [60, 18, 15] use contrastive learning by creating many negative queries (based on word replacement) or by mining negative image regions for a given query. Wang et al. [60] is a strong method in this domain, hence we establish it as a baseline in our experiments. Liu et al. [39] parses sentences to scene graphs for capturing visual relation between mentions to improve phrase grounding. However, this cannot be directly applied to our task, as parsing scene graphs from narrations is typically very noisy and incomplete. Wang et al. [59] aims to learn/predict object class labels from the object detector during training and inference respectively. Due to the open-vocabulary setting in our dataset, we directly rely on predictions from the detector and use them as features to avoid the complexity of open-vocabulary object detection. Furthermore, as we show in the experiments that grounding is useful to anchor mentions but it is not sufficient to resolve coreferences without prior linguistic knowledge. Thus, our

---

1Our code and dataset will be made publicly available.
method also employs contrastive learning but for learning CR from weak supervision.

3. Coreferenced Image Narratives

Our CIN dataset contains 1880 images from the Localized Narratives dataset [48] that come with long-form text descriptions (narrations) and mouse traces. These images are originally a subset of the test and validation set of the Flickr30k dataset [47]. We annotated this subset with coreference chains and bounding boxes in the image that are linked with the textual coreference chains, and use them only for validation and testing. Note that we also include singletons (i.e., coreference chains of length one). Fig. 1 shows an example image from CIN.

Annotation procedure. The annotation involved three steps: (1) marking the mentions (sequences of words) that refer to a localized region in the image, (2) identifying coreference chains for the marked mentions, including (a) pronominal words such as that refer to a localized region in the image, (2) identifying coreference chains of length one). Fig. 1 shows an example image from CIN.

Comparison to existing datasets. In Table 2, we compare our proposed CIN dataset to other CR datasets. This comparison shows that most of the other datasets are either from a restricted domain (shopping, indoor scenes, etc.), have limited mention types referring to either only people or limited object categories, or do not cover all possible referring expression types such as common nouns (a person), proper nouns (Peter) and pronouns (he).

4. Method

4.1. Text-only CR

Given a sentence containing a set of mentions (i.e., referential words or phrases), the task of CR is to identify which mentions refer to the same entity. This is fundamentally a clustering problem [57]. In this work, we use an off-the-shelf NLP parser [2] to obtain the mentions. Formally, let \( S = \{m_1, m_2, \ldots, m_{|S|}\} \) denote a sentence with \(|S|\) mentions, where each mention \( m \) contains a sequence of words, \( \{w_1, w_2, \ldots, w_{|m|}\} \). We assign a label \( y_{ij} \) to each mention pair \( (m_i, m_j) \), which is set to 1 when the pair refers to the same entity, and to 0 otherwise. We wish to learn a compatibility function, a deep network \( f \) that scores high if a pair refers to the same entity, and low otherwise.

Given a training set \( D \) that contains \(|D|\) sentences with their corresponding labels, one can learn \( f \) by optimizing a binary cross-entropy loss:

\[
\min_{f} \sum_{S \in D} \sum_{i=0}^{S-1} \sum_{j=i+1}^{S} \log(y_{ij}(\sigma(f(m_i, m_j)))) - \frac{1}{2} + \frac{1}{2}
\]

where \( \sigma \) is the sigmoid function. Note that prior methods
[29, 30, 21] require large labeled datasets for training and are limited to only a single modality, text. These methods typically also combine the learning with fixed rules based on recency and grammatical principles [29].

4.2. CR in image narrations

Problem definition. Next we extend the text-only CR to image-text data in the absence of coreference labels. Let \((I, S)\) denote an image-text pair where \(S\) describes an image \(I\) as illustrated in Figure 1, and assume that coreference labels for mention pairs are not present. As in Sec. 4.1, our goal is to identify the mentions that refer to the same entity in an image-text pair. Each image is defined by \(|I|\) regions \(I = \{r_1, r_2, \ldots, r_{|I|}\}\) which are obtained by running the pretrained object detector (trained on the COCO [34] and Visual Genome [26] dataset) in [53] on the image. Each region \(r\) is described by its bounding box coordinates \(b\), the text embedding for the detected object category \(o\), and the visual features \(v\). More details are provided in Section 4.3.

Weak supervision. We use ‘weak supervision’ to refer to a setting where no coreference label for mention pairs and no grounding of mentions \((i.e.,\) bounding boxes are not linked to phrases in the text) are available. Moreover, in contrast to the output space of the object detector (a restricted set of object categories), the sentences describing our images come from unconstrained vocabulary. Hence, an object instance in a sentence can be referred to with a synonym or may not even be present in the object detector vocabulary [34, 27]. Finally, the object detector can only output category-level labels and hence cannot localize object instances based on the more specific instance-level descriptions provided by the sentences. For instance in Figure 1, a person and the woman both are labeled as person by the object detector.

In addition to image and text, we explore the use of an auxiliary modality, mouse trace segments provided in [48]. Each mouse trace includes a sequence of 2D points over time that relate to a region in the image when describing the scene. As the text in Localized Narratives is transcription of the speech of the annotators, the mouse traces are synced with spoken words, which we denote as \(T = \{t_1, t_2, \ldots, t_{|T|}\}\) where \(|T| = |S|\). These features are stacked with textual features (see Section 4.3).

In the weakly supervised setting, the key challenge is to replace the coreference label supervision with an alternative one. We hypothesize that each mention in a coreferring pair corresponds to (approximately) the same image region, and it is possible to learn a joint image-text space which is sufficiently rich to capture such correlations. Concretely, let \(g(m, r)\) denote an auxiliary function that is instantiated as a deep network and outputs a score for the mention \(m\) being located at region \(r\) in image \(I\). This grounding score for each mention can be converted into probability values by normalizing them over all regions in the image:

\[
\tilde{g}(m, r) = \frac{\exp (g(m,r))}{\sum_{r' \in I} \exp (g(m,r'))}.
\]  

The compatibility function \(f\) can be defined as a sum product of a pair’s grounding probabilities over all regions:

\[
f(m, m') = \sum_{r \in I} \tilde{g}(m, r) \tilde{g}(m', r).
\]

In words, mention pairs with similar region correlations yield bigger compatibility scores and are hence more likely to corefer to each other. The key idea is that we employ the grounding for mentions as anchors to relate coreferring mentions \((e.g.,\) a person and the woman\). At test time, we compute \(f(m, m')\) for all the pairs and threshold them to predict their pairwise coreference labels.

As no ground-truth bounding box for each mention is available for learning the grounding \(g\), we pose grounding as a weakly supervised localization task as in [18, 60]. To this end, we impute the missing bounding boxes by taking the highest scoring region for a given mention \(m\) at each training iteration:

\[
r_m = \arg \max_{r \in I} g(m, r).
\]

Then we use \(r_m\) as the pseudo-truth to learn \(g\) as following:

\[
\min_g \sum_{(I,S) \in D} \sum_{m \in S} -\log \left( \frac{\exp (g(m, r_m))}{\sum_{r' \in I} \exp (g(m, r_m'))} \right)
\]

where \(r_m = \arg \max_{r \in I} g(m, r)\) is the highest scoring region in image \(I\) for mention \(m\). For each mention, we treat the highest scoring region in the original image as positive and other highest scoring regions across different images as negatives, and optimize \(g\) for discriminating between the two. However, as the denominator requires computing \(g\) over all training samples at each iteration, which is not computationally feasible, we instead sample the negatives only from the randomly sampled minibatch.

Linguistic constraints. Learning the associations between textual and visual features helps with disambigating coreferring mentions, especially when mentions contain visually salient and discriminative features. However, resolving coreferences when it comes to pronouns \((e.g.,\) her, their\)) or ambiguous phrases \((e.g.,\) one man or another man\) remains challenging. To address such cases, we propose to incorporate a regularizer into the compatibility function \(f(m, m')\) based on various linguistic priors. Concretely, we construct a look-up table for each mention pair \(q(m, m')\) based on the following set of rules [29]:

\[
\text{Linguistic constraints.}
\]

...
Figure 3: Overview of our pipeline. Our model encodes the image regions obtained from an object detector using the image encoder. We parse text mentions and mouse traces from the sentence description, which are then encoded using a text and trace encoder respectively. Finally, a joint text-trace encoder learns a joint embedding for text and traces. A cross-attention module attends to the words given an image region and then we compute the joint probability of the paired mentions, thus forming coreference chains.

(a) Exact String Match. Two mentions corefer if they exactly match and are noun phrases (not pronouns).

(b) Pronoun Resolution. Based on the part-of-speech tags for the mentions, we set \( q(m, m') \) to 1 if \( m \) is a pronoun and \( m' \) is the antecedent noun that occurs before the pronoun.

(c) Distance between mentions. Smaller distance is more likely to indicate coreference since mentions to occur close together if they refer to the same entity.

(d) Last word match. In certain cases, the entire phrases might not match but only the last word of the phrases.

(e) Overlap between mentions. If two mentions have one or more overlapping words, then they are likely to corefer.

Finally, we include \( q(m, m') \) as a regularizer in Eq. (5):

\[
\min_g \sum_{I,S} \sum_{m \in S} \left( - \log \frac{\exp(g(m, r_m))}{\sum_{m' \in D \setminus I} \exp(g(m, r_{m'}))} \right) + \lambda \sum_{m' \in S} \| f(m, m') - q(m, m') \|_F^2
\]

where \( \lambda \) is a scalar weight for the Frobenius norm term. Note that \( f \) is a function of \( g \) (see Eq. (3)). We show in Section 6 that incorporating this term results in steady and significant improvements in CR performance.

4.3. Network modules

Our model (illustrated in Figure 3) consists of an image encoder \( e_i \) and text encoder \( e_t \) to extract visual and linguistic information respectively, and a cross-attention module \( a \) for their fusion.

Image encoder \( e_i \) takes in a \( d_v \)-dimensional vector for each region \( r \) that consists of a vector consisting of bounding box coordinates \( b \in \mathbb{R}^4 \), text embedding for the detected object category \( o \in \mathbb{R}^{d_v} \) and visual features \( v \in \mathbb{R}^{d_v} \). The regions are extracted from a pretrained object detector \([53]\) for the given image \( I \). The image encoder applies a nonlinear transformation to this vector to obtain a \( d \)-dimensional embedding for each region \( r \).

Text encoder \( e_t \) takes in the multiple mentions from a parsed multi-sentence image description \( S \) produced by an NLP parser \([2]\) and outputs a \( d \)-dimensional embedding for each word in the parsed mentions. Note that the parser does not only extract nouns but also pronouns as mentions.

Mouse trace encoder \( e_m \) takes in the mouse traces for each mention parsed above after it is preprocessed into a 5D vector of coordinates and area, \( (x_{min}, x_{max}, y_{min}, y_{max}, \text{area}) \) \([45]\) and outputs a \( d_m \)-dimensional embedding. In \([7, 48]\), mouse trace embeddings have been exploited for image retrieval, however, we use them to resolve coreferences. We concatenate each mention embedding extracted from \( e_t \) with the mouse trace encoding \( e_m \), denoted as \( e_{tm} \) and apply additional nonlinear transformations (Joint encoder in Fig. 3) before feeding into the cross-attention module.

Cross-attention module \( a \) takes in the joint text-trace embeddings for all the words in each mention and returns a \( d \)-dimensional vector for each \( m \) by taking a weighted average of them based on their correlations with the image regions. Concretely, in this module, we first compute the correlation between each word \( w \) (or joint word-mouse trace) and all regions, take the highest correlation over the regions through an auxiliary function \( \hat{a} \):

\[
\hat{a}(w) = \max_{r \in I} \left( \frac{\exp(e_{tm}(w) \cdot e_i(r))}{\sum_{r' \in I} \exp(e_{tm}(w) \cdot e_i(r'))} \right)
\]
where \( \cdot \) is dot product. The transformation can be interpreted as probability of word \( w \) being present in image \( I \). Then we compute a weighted average of the word embeddings for each mention \( m \):

\[
a(m) = \sum_{w \in m} \bar{a}(w)e_{tm}(w).
\]

(8)

Similarly, \( a(m) \) can be seen as probability of mention \( m \) being present in image \( I \).

**Scoring function** \( g(m, r) \) can be written as a dot product between the output of the attention module and region embeddings:

\[
g(m, r) = a(m) \cdot e_{i}(r).
\]

(9)

While taking a dot product between the two embeddings seemingly ignores the correlation between text and image data, the region embedding \( e_{i}(r) \) encodes the semantic information about the detected object category in addition to other visual features and hence results in a high score only when the mention and region are semantically close. Further implementation details about the modules can be found in Section 5 and the supplementary.

5. Experiments

We train our models on the Flickr30k subset of the Localized Narratives [48] which consists of 30k image-narration pairs, and evaluate on the proposed CIN dataset, which contains 1000 and 800 pairs for test and validation respectively.

**Evaluation metrics.** To evaluate the CR performance, we use the standard link-based metrics MUC [58] and BLANC [52]:

- **(a) MUC F-measure** counts the coreference links (pairs of mentions) common to the predicted chain \( R \) and the ground-truth chain \( K \) by computing MUC-R (recall) and MUC-P (precision).
- **(b) BLANC** measures the precision (BLANC-P) and recall (BLANC-R) between the ground-truth and predicted coreference links and also between non-coreferent links.
- **(c) Narrative grounding.** For evaluating narrative grounding in images, we consider a prediction to be correct if the IoU (Intersection over Union) between the predicted bounding box and the ground truth box is larger than 0.5 [60, 18]. We report percentage accuracy for evaluating narrative grounding for both noun phrases and pronouns. Further details about the metrics is in supplementary material.

**Inputs and modules.** For the image modeling, we extract bounding box regions, visual features and object class labels using the Faster-RCNN object detector [53]. For the text modeling, we use Glove embeddings [46] to encode the object class labels and the mentions from the textual branch. For the mouse traces, we follow [48] and extract the trace for each word in the sentence and then convert it into bounding box coordinates for the initial representation. The model discussed in Sec. 4 referred to as ‘Ours’ in Sec. 6 uses the transformer backbone for the image, text and trace encoders (more details in supplementary).

**Baselines.** We consider the following baselines to fairly compare and evaluate our proposed method:

- **(a) Text-only CR:** For all these methods, we directly evaluate the coreference chains using the narration only without the image. (1) **Rule-based [29]:** In this method, a multisieve rule based system is used to find mentions in the sentence and the coreference chains, (2) **Neural-Coref [30]:** Instead of rules, this method is trained end-to-end using a neural network on a large corpus of wikipedia data to detect mentions and coreferences, and (3) **Similarity-based:** We compute cosine similarity between mentions using Glove word features and threshold them to get coreference chains.

- **(b) Visual grounding:** The baselines discussed below are not trained for CR and hence we post-process their output in order to evaluate for CR. (1) **GLIP [32]:** GLIP is trained on large-scale image-text paired data with bounding box annotations and shows improvement on object detection and visual phrase grounding. To evaluate it for CR, we predict bounding boxes for the mentions in the narratives from GLIP. If the IoU overlap between the mentions is greater than 0.7, then we consider them to form a coreference chain. (2) **MAF [60]:** MAF is a weakly supervised phrase grounding method, originally trained on the Flickr30k-Entities [47]. We train this model on narrations data and evaluate CR by computing Eq. (3). (3) **MAF++:** We retrain the MAF† model on the narrations with our regularization term. Architecturally our method differs from the MAF† in two aspects: i) we employ a transformer to encode visual and text features unlike the MLP in theirs and ii) we attend to the mouse traces when present (not present in MAF) and word features jointly whereas they directly compute the similarity function.

6. Results

Coreference resolution. In Table 3, we report CR performance of the baselines and our method. Our method significantly outperforms all the text-only and the grounding baselines on all the metrics. The text-only CR baselines in the first three rows fail to effectively resolve conferences from narrations. It is important to note that relatively high number in BLANC scores (compared to MUC) occurs because this measure also counts non-coreferent links (i.e., mentions that are not paired with anything), whereas MUC only measures pairs that are resolved.

The rule-based method [29] uses exact match noun
Table 3: CR performance on CIN dataset. MT denotes mouse trace and \( ^\dagger \) denotes our trained model.

<table>
<thead>
<tr>
<th>Method</th>
<th>Text</th>
<th>Image</th>
<th>MT</th>
<th>MUC-R</th>
<th>MUC-P</th>
<th>MUC-F1</th>
<th>BLANC-R</th>
<th>BLANC-P</th>
<th>BLANC-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-Based [29]</td>
<td>✓    │ ×</td>
<td>×</td>
<td>5.6</td>
<td>10.13</td>
<td>6.4</td>
<td>3.3</td>
<td>4.1</td>
<td>4.9</td>
<td></td>
</tr>
<tr>
<td>Neural-Coref [30]</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>0.11</td>
<td>0.17</td>
<td>0.13</td>
<td>1.59</td>
<td>36.99</td>
<td>3.23</td>
</tr>
<tr>
<td>Similarity-based</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>7.07</td>
<td>14.43</td>
<td>9.06</td>
<td>37.48</td>
<td>65.17</td>
<td>45.98</td>
</tr>
<tr>
<td>GLIP [32]</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>0.13</td>
<td>0.12</td>
<td>0.12</td>
<td>21.71</td>
<td>61.40</td>
<td>31.66</td>
</tr>
<tr>
<td>MAF(^\dagger) [60]</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>25.86</td>
<td>10.18</td>
<td>13.21</td>
<td>37.68</td>
<td>61.14</td>
<td>58.17</td>
</tr>
<tr>
<td>MAF++</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>19.07</td>
<td>15.62</td>
<td>15.65</td>
<td>41.25</td>
<td>65.04</td>
<td>47.21</td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>22.07</td>
<td>17.10</td>
<td>17.58</td>
<td>42.72</td>
<td>65.92</td>
<td>48.29</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>24.87</td>
<td>18.34</td>
<td>19.19</td>
<td>43.81</td>
<td>66.35</td>
<td>48.53</td>
</tr>
</tbody>
</table>

Next we compare our method to the visual grounding baselines that use both image and text input. Our method also outperform these baselines: Though GLIP is pretrained on large-scale data with ground-truth boxes for each object in captions, these captions are usually short and do not contain multiple mentions of entities, unlike in our data. Hence GLIP acts more like an object detector, fails to link coreferring pairs (low MUC scores) and merely identifies singletons (higher BLANC scores). While it is nontrivial to finetune GLIP on our data without groundtruth boxes, we finetuned MAF on our data, as its training does not require groundtruth boxes; we denote this as MAF\(^\dagger\). This is the strongest baseline on our task, as training it on narrations including the pronouns reduces the domain gap and enables it to resolve coreferences well. However, this method obtains low precision by incorrectly linking visually similar mentions (that do not belong together) such as trees, plant, flowers. When the training is regularized with the linguistic priors from our method, denoted as MAF++, its performance significantly improves on both MUC and BLANC. The constraint helps to push away the negative mentions (trees, plant, etc.) and encourage the model to learn unique embeddings for them. Due to the self-attention in the transformer architectures, Ours without mouse traces (MT) achieves better performance than MAF++, a simple MLP baseline. The performance difference between our method without using mouse-traces and MAF++ can be explained by the better architecture described previously. Finally, our method achieves the best performance gains in CR thanks to the mouse traces and improved architecture over MAF.

Ablation on mouse traces In Table 3, we also analyze the contribution of modeling mouse traces (second last row). Adding the mouse traces improves performance on CR across all metrics. We hypothesize that the mouse traces provide a strong discriminative location prior to the textual mentions, which helps the model to learn a better compatibility score. To visualize qualitatively, consider the example in Figure 3, the same mention this person points to two different visual regions – one with the person holding the ball and the other person standing next to the door. In such cases, mouse traces provide a strong signal for disambiguation. But in many cases, mouse traces are noisy and can link mentions that are very close to each other in the image, referring to two different regions. In the above example, mouse traces for these persons and this person have a significant overlap and hence act as a noisy prior. Therefore, without the visual/image region features, it is very challenging to address the problem with mouse traces alone.

Narrative grounding Not only does our method show performance gains on CR but also outperform the baselines on another challenging task of narrative grounding. Table 4 compares results from our methods and baselines. We directly compare with the weakly supervised method for a fair comparison. MAF\(^\dagger\) [60] is originally evaluated on the Flickr30k-Entities [47] dataset where the textual descriptions are significantly shorter (i.e. single sentence) than the image narrations in our dataset. The performance of
in the middle of the picture, we see a person who is wearing the costumes is walking on the road[1].
on the left side, we see the people are walking.
at the bottom, we see an object which looks like a water sprinkler.
in front of them, we see a baby trolley and a baby is sitting in the baby trolley.

We demonstrate strong experimental results in multiple settings. In the future, we plan to address the noise induced by the machine,
and narrative grounding. We visualize the narrative grounding results from our proposed method on the images. The model correctly resolves and localizes phrases such as a person, who, the people, them and a baby trolley, the baby trolley. Whereas, the model fails to ground and chain the instance a baby. It is interesting to note that our model pairs an object and water sprinkler, thereby resolving ambiguity in what the object might refer to. But it fails to add which to this coreference chain. Moreover, without the language regularizer, our method fails to link them to the people. It is very hard to learn coreferences for these pronouns as they come with a weak language prior and hence are difficult for the model to disambiguate. Our model (without regularization) misses the referring expression of the baby trolley to refer to the instance of the trolley before. With the help of rules (e.g. last token match), we can resolve these pairs more often than not. Hence, we clearly show the challenging problem of coreferences we are dealing with and indicate the great potential for developing models with strong contextual reasoning.

Table 4: Grounding accuracy (%) for noun phrases and pronouns and the overall accuracy on the CIN dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Reg</th>
<th>Noun Phrases</th>
<th>Pronouns</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAF + [60]</td>
<td>✗</td>
<td>21.60</td>
<td>18.31</td>
<td>20.91</td>
</tr>
<tr>
<td>MAF++</td>
<td>✓</td>
<td>25.58</td>
<td>22.56</td>
<td>24.91</td>
</tr>
<tr>
<td>Ours</td>
<td>✗</td>
<td>27.62</td>
<td>23.46</td>
<td>26.75</td>
</tr>
<tr>
<td>Ours (w/o Reg)</td>
<td>✓</td>
<td>30.27</td>
<td>25.96</td>
<td>29.36</td>
</tr>
</tbody>
</table>

Table 5: Our method with/without cross attention.

**Further ablations.** Table 5 compares the performance of our final method under two settings: (1) directly averaging the word features or (2) attending over the words by using the image as the query as discussed in Sec. 4. Both the narrative grounding accuracy and the coreference evaluation get a boost in performance for visually aware word features. More often than not, the word phrases are relatively short (e.g., the machine) and hence the model does not always learn to disambiguate better with attention for the grounding. On the other hand, this technique is especially useful for CR because the flow of visual information to the word features acts as a prior to cluster mentions that refer to the same region but with are referred to with different mentions/entities in the text (e.g. the machine and an equipment). We provide detailed ablations in the supplementary material.

**Qualitative results** Figure 4 qualitatively analyzes CR and narrative grounding. We visualize the narrative grounding results from our proposed method on the images. The model correctly resolves and localizes phrases such as a person, who, the people, them and a baby trolley, the baby trolley. Whereas, the model fails to ground and chain the instance a baby. It is interesting to note that our model pairs an object and water sprinkler, thereby resolving ambiguity in what the object might refer to. But it fails to add which to this coreference chain. Moreover, without the language regularizer, our method fails to link them to the people. It is very hard to learn coreferences for these pronouns as they come with a weak language prior and hence are difficult for the model to disambiguate. Our model (without regularization) misses the referring expression of the baby trolley to refer to the instance of the trolley before. With the help of rules (e.g. last token match), we can resolve these pairs more often than not. Hence, we clearly show the challenging problem of coreferences we are dealing with and indicate the great potential for developing models with strong contextual reasoning.

**7. Conclusion**

We introduced a novel task of resolving coreferences in image narrations, clustering mention pairs referring to the same entity. For benchmarking and enabling the progress, we introduce a dataset – CIN – that contains images with narratives annotated with coreference chains and their grounding in the images. We formulate the problem of learning CR by using weak supervision from image-text pairs to disambiguate coreference chains and linguistic priors to avoid learning grammatically wrong chains. We demonstrate strong experimental results in multiple settings. In the future, we plan to address the noise induced by the language rules during learning and also reduce the errors coming from the mouse traces. We hope that our proposed task definition, dataset and the weakly supervised method will advance the research in multi-modal understanding.
References


[27] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan


[59] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14090–14100, 2021. 2


Appendix

8. Annotation Details

Localized Narratives dataset. Tuset et al. [48] proposed the Localized Narratives dataset, 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 [34], Flickr30k [47], ADE20k [67] and Open Images [27]. The statistics of the individual datasets are shown in Table 6.

<table>
<thead>
<tr>
<th>Localized Narratives Subsets [48]</th>
<th>#images</th>
<th>#captions</th>
<th>#words/capt.</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCO</td>
<td>123,287</td>
<td>142,845</td>
<td>41.8</td>
</tr>
<tr>
<td>Flickr30k</td>
<td>31,783</td>
<td>32,578</td>
<td>57.1</td>
</tr>
<tr>
<td>ADE20k</td>
<td>22,210</td>
<td>22,529</td>
<td>43.0</td>
</tr>
<tr>
<td>Open Images</td>
<td>671,469</td>
<td>675,155</td>
<td>34.2</td>
</tr>
</tbody>
</table>

Table 6: 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 5 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 5, 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 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.

Corefereced 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.

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.

Statistics for the Corefereced Image Narratives. In Figure 6, 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 7, 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 coreferecing and grounding models. Moreover, the average length of the mentions (excluding pronouns) is 1.93.

9. 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
1. Cluster Related Phrases with the same color

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>C2</td>
<td>C3</td>
<td>C4</td>
<td>C5</td>
<td>C6</td>
<td>C7</td>
<td>C8</td>
<td></td>
</tr>
<tr>
<td>C9</td>
<td>C10</td>
<td>C11</td>
<td>C12</td>
<td>C13</td>
<td>C14</td>
<td>C15</td>
<td>C16</td>
<td></td>
</tr>
<tr>
<td>C17</td>
<td>C18</td>
<td>C19</td>
<td>C20</td>
<td>C21</td>
<td>C22</td>
<td>C23</td>
<td>C24</td>
<td></td>
</tr>
<tr>
<td>C25</td>
<td>C26</td>
<td>C27</td>
<td>C28</td>
<td>C29</td>
<td>C30</td>
<td>C31</td>
<td>C32</td>
<td></td>
</tr>
<tr>
<td>C33</td>
<td>C34</td>
<td>C35</td>
<td>C36</td>
<td>C37</td>
<td>C38</td>
<td>C39</td>
<td>C40</td>
<td></td>
</tr>
</tbody>
</table>

Caption

in this image there is a glass and a paper on a table. behind the table there is a lady and a girl sitting on chairs. behind them there is a man and a girl sitting on chairs.

2. Annotate the clusters in the images with bounding boxes

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>C2</td>
<td>C3</td>
<td>C4</td>
<td>C5</td>
<td>C6</td>
<td>C7</td>
<td>C8</td>
<td></td>
</tr>
<tr>
<td>C9</td>
<td>C10</td>
<td>C11</td>
<td>C12</td>
<td>C13</td>
<td>C14</td>
<td>C15</td>
<td>C16</td>
<td></td>
</tr>
<tr>
<td>C17</td>
<td>C18</td>
<td>C19</td>
<td>C20</td>
<td>C21</td>
<td>C22</td>
<td>C23</td>
<td>C24</td>
<td></td>
</tr>
<tr>
<td>C25</td>
<td>C26</td>
<td>C27</td>
<td>C28</td>
<td>C29</td>
<td>C30</td>
<td>C31</td>
<td>C32</td>
<td></td>
</tr>
<tr>
<td>C33</td>
<td>C34</td>
<td>C35</td>
<td>C36</td>
<td>C37</td>
<td>C38</td>
<td>C39</td>
<td>C40</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: Annotation interface from Label Studio.

<table>
<thead>
<tr>
<th>MT Loss Function</th>
<th>CR</th>
<th>Grounding</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MUC-R</td>
<td>MUC-P</td>
</tr>
<tr>
<td>✗</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓ L1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓ MSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✗ Frobenius Norm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7: Ablation study with different regularizer types and without mouse traces.

ground-truth chains $K$. It involves computing the partitions with respect to the two chains:

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

(10)

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

(11)

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 is calculated as follows:

$$R_c = \frac{|C_k \cap C_r|}{|C_k|}$$

and

$$P_c = \frac{|C_k \cap C_r|}{|C_r|},$$

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:

$$\text{BLANC-R} = \frac{(R_c + R_n)}{2}$$

and

$$\text{BLANC-P} = \frac{(P_c + P_n)}{2},$$

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 [60, 18]. Following [22], 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.

10. 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 [53]. We use Glove embeddings [46] to encode the object class labels and the mentions from the textual branch. For the mouse traces, we follow [48] 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 [11] 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 [41] 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.

11. Ablation Study

In Table 7, first we study the impact of training with just our proposed architecture without the mouse traces and regularizer. The model suffers a drop in both the CR and grounding performance. While the model is able to learn some coreference links but it still produces a lot of false positives (lower precision scores), compared to the model trained with mouse traces. Next, we study the effect of training with different regularizer types. We achieve improvement in performance with the Frobenius norm as a constraint unlike L1 and MSE, as it imposes a stronger constraint on the learned coreference matrix. Note that the last row corresponds to our proposed model (MT+Frobenius norm).

12. Additional Qualitative Results

In Fig. 8, 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 learning to disambiguate the mentions in the descriptions and this work paves the way for future research.

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

Figure 8: 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.

Description: 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[8] on the right side is wearing the google[5] and an orange cap[6]. it[7] 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 cap[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]]

Description: 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]]

Description: 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[0], it[4]]