# Automatic Annotation of Structured Facts in Images

**Mohamed Elhoseiny**[1,2], **Scott Cohen**[1], **Walter Chang**[1], **Brian Price**[1], **Ahmed Elgammal**[2]
[1]Adobe Research [2]Department of Computer Science, Rutgers University

## Abstract

Motivated by the application of fact-level image understanding, we present an automatic method for data collection of structured visual facts from images with captions. Example structured facts include attributed objects (e.g., <flower, red>), actions (e.g., <baby, smile>), interactions (e.g., <man, walking, dog>), and positional information (e.g., <vase, on, table>). The collected annotations are in the form of fact-image pairs (e.g.,<man, walking, dog> and an image region containing this fact). With a language approach, the proposed method is able to collect hundreds of thousands of visual fact annotations with accuracy of 83% according to human judgment. Our method automatically collected more than 380,000 visual fact annotations and more than 110,000 unique visual facts from images with captions and localized them in images in less than one day of processing time on standard CPU platforms. We will make the data publically available.

## 1 Introduction

People generally acquire visual knowledge by exposure to both visual facts and to semantic or language-based representations of these facts, e.g., by seeing an image of "a person petting dog" and observing this visual fact associated with its language representation . In this work, we focus on methods for collecting structured facts that we define as structures that provide attributes about an object, and/or the actions and interactions this object may have with other objects. We introduce the idea of automatically collecting annotations for second order visual facts and third order visual facts where second order facts <S,P> are attributed objects (e.g., <S: car, P: red>) and single-frame actions (e.g., <S: person, P: jumping>), and third order facts specify interactions (i.e., <boy, petting, dog>). This structure is helpful for designing machine learning algorithms that learn deeper image semantics from caption data and allow us to model the relationships between facts. In order to enable such a setting, we need to collect these structured fact annotations in the form of (language view, visual view) pairs (e.g., <baby, sitting on, chair> as the language view and an image with this fact as a visual view) to train models.

(Chen et al., 2013) showed that visual concepts, from a predefined ontology, can be learned by querying the web about these concepts using image-web search engines. More recently, (Divvala et al., 2014) presented an approach to learn concepts related to a particular object by querying the web with Google-N-gram data that has the concept name. There are three limitations to these approaches. (1) It is difficult to define the space of visual knowledge and then search for it. It is further restricting to define it based on a predefined ontology such as (Chen et al., 2013) or a particular object such as (Divvala et al., 2014). (2) Using image search is not reliable to collect data for concepts with few images on the web. These methods assume that the top retrieved examples by image-web search are positive examples and that there are images available that are annotated with the searched concept. (3) These concepts/facts are not structured and hence annotations lacks information like "jumping" is the action part in <person, jumping >, or "man' and "horse" are interacting in <person, riding, horse >. This structure is important for deeper understanding of visual data, which is one of the main motivations of this work.

The problems in the prior work motivate us to propose a method to automatically annotate struc-



*Proceedings of the 5th Workshop on Vision and Language*, pages 1–9,
Berlin, Germany, August 12 2016. 

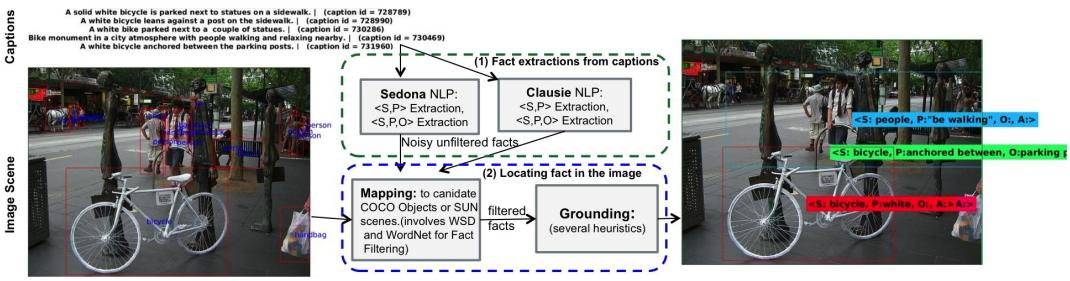



Figure 1: Structured Fact Automatic Annotation

tured facts by processing image caption data since facts in image captions are highly likely to be located in the associated images. We show that a large quantity of high quality structured visual facts could be extracted from caption datasets using natural language processing methods. Caption writing is free-form and an easier task for crowd-sourcing workers than labeling second- and third-order tasks, and such free-form descriptions are readily available in existing image caption datasets. We focused on collecting facts from the MS COCO image caption dataset (Lin et al., 2014) and the newly collected Flickr30K entities (Plummer et al., 2015). We automatically collected more than 380,000 structured fact annotations in high quality from both the 120,000 MS COCO scenes and 30,000 Flickr30K scenes.

The main contribution of this paper is an accurate, automatic, and efficient method for extraction of structured fact visual annotations from image-caption datasets, as illustrated in Fig. 1. Our approach (1) extracts facts from captions associated with images and then (2) localizes the extracted facts in the image. For fact extraction from captions, We propose a new method called *SedonaNLP* for fact extraction to fill gaps in existing fact extraction from sentence methods like Clausie (Del Corro and Gemulla, 2013). SedonaNLP produces more facts than Clausie, especially <subject,attribute> facts, and thus enables collecting more visual annotations than using Clausie alone. The final set of automatic annotations are the set of successfully localized facts in the associated images. We show that these facts are extracted with more than 80% accuracy according to human judgment.

## 2 Motivation

Our goal by proposing this automatic method is to generate language&vision annotations at the fact-level to help study language&vision for the sake of structured understanding of visual facts. Existing systems already work on relating captions directly to the whole image such as (Karpathy et al., 2014; Kiros et al., 2015; Vinyals et al., 2015; Xu et al., 2015; Mao et al., 2015; Antol et al., 2015; Malinowski et al., 2015; Ren et al., 2015). This gives rise to a key question about our work: why it is useful to collect such a large quantity of structured facts compared to caption-level systems?

We illustrate the difference between caption-level learning fact-level learning that motivates this work by the example in Fig 1. Caption-level learning systems correlate captions like those on top of Fig. 1(top-left) to the whole image that includes all objects. Structured Fact-level learning systems are instead fed with localized annotations for each fact extracted form the image caption; see in Fig. 1(right), Fig. 6, and 7 in Sec. 6. Fact level annotations are less confusing training data than sentences because they provide more precise information for both the language and the visual views. **(1)** From the language view, the annotations we generate is precise to list a particular fact (e.g., <bicycle,parked between, parking posts>). **(2)** From the visual view, it provide the bounding box of this fact; see Fig 1. **(3)** A third unique part about our annotations is the structure: e.g., <bicycle,parked between, parking posts> instead of "a bicycle parked between parking posts".

Our collected data has been used to develop methods that learn hundreds of thousands of image facts, as we introduced and studied in (Elhoseiny et al., 2016a). The results shows that fact-level learning is superior compared to caption-level learning like (Kiros et al., 2015), as shown in Table 4 in (Elhoseiny et al., 2016a) (16.39% accuracy versus 3.48% for (Kiros et al., 2015)). It further shows the value of the associated structure in the (16.39% accuracy versus 8.1%) in Table 4(Elhoseiny et al., 2016a)). Similar results also shown on a smaller scale in Table 3 in (Elhoseiny et al.,

2016a).

## 3 Approach Overview

We propose a two step automatic annotation of structured facts: (i) Extraction of structured fact from captions, and (ii) Localization of these facts in images. First, the captions associated with the given image are analyzed to extract sets of clauses that are considered as candidate <S,P>, and <S,P,O> facts.

Captions can provide a tremendous amount of information to image understanding systems. However, developing NLP systems to accurately and completely extract structured knowledge from free-form text is an open problem. We extract structured facts using two methods: Clausie (Del Corro and Gemulla, 2013) and Sedona( detailed later in Sec 4); also see Fig 1. We found Clausie (Del Corro and Gemulla, 2013) missed many visual facts in the captions which motivated us to develop Sedona to fill this gap as detailed in Sec. 4.

Second, we localize these facts within the image (see Fig. 1). The successfully located facts in the images are saved as fact-image annotations that could be used to train visual perception models to learn attributed objects, actions, and interactions. We managed to collect 380.409 high-quality second- and third-order fact annotations (146,515 from Flickr30K Entities, 157,122 from the MS COCO training set, and 76,772 from the MS COCO validation set). We present statistics of the automatically collected facts in the Experiments section. Note that the process of localizing facts in an image is constrained by information in the dataset.

For MS COCO, the dataset contains object annotations for about 80 different objects as provided by the training and validation sets. Although this provides abstract information about objects in each image (e.g., "person"), it is usually mentioned in different ways in the caption. For the "person" object, "man", "girl", "kid", or "child" could instead appear in the caption. In order to locate second- and third-order facts in images, we started by defining visual entities. For the MS COCO dataset (Lin et al., 2014), we define a visual entity as any noun that is either (1) one of the MS COCO dataset objects, (2) a noun in the WordNet ontology (Miller, 1995; Leacock and Chodorow, 1998) that is an immediate or indirect hyponym of one of the MS COCO objects (since WordNet is searchable by a sense and not a word, we perform word sense disambiguation on the sentences using a state-of-the-art method (Zhong and Ng, 2010)), or (3) one of scenes the SUN dataset (Xiao et al., 2010) (e.g., a "restaurant"). We expect visual entities to appear either in the S or the O part (if exists) of a candidate fact. This allows us to then localize facts for images in the MS COCO dataset. Given a candidate third-order fact, we first try to assign each S and O to one of the visual entities. If S and O elements are not visual entities, then the fact is ignored. Otherwise, the facts are processed by several heuristics, detailed in Sec 5. For instance, our method takes into account that grounding the plural "men" in the fact <S:men, P: chasing, O: soccer ball > may require the union of multiple "man" bounding boxes.

In the Flickr30K Entities dataset (Plummer et al., 2015), the bounding box annotations are presented as phrase labels for sentences (for each phrase in a caption that refers to an entity in the scene). A visual entity is considered to be a phrase with a bounding box annotation or one of the SUN scenes. Several heuristics were developed and applied to collect these fact annotations, e.g. grounding a fact about a scene to the entire image; detailed in Sec 5.

## 4 Fact Extraction from Captions

We extract facts from captions using Clausie (Del Corro and Gemulla, 2013) and our proposed SedonaNLP system. In contrast to Clausie, we address several challenging linguistic issues by evolving our NLP pipeline to: 1) correct many common spelling and punctuation mistakes, 2) resolve word sense ambiguity within clauses, and 3) learn a common spatial preposition lexicon (e.g., "next_to", "on_top_of", "in_front_of") that consists of over 110 such terms, as well as a lexicon of over two dozen collection phrase adjectives (e.g., "group_of", "bunch_of", "crowd_of", "herd_of"). For our purpose, these strategies allowed us to extract more interesting structured facts that Clausie fails at which include (1) more discrimination between single versus plural terms, (2) extracting positional facts (e.g., next_to). Additionally, SedonaNLP produces attribute facts that we denote as <S, A>; see Fig 4. Similar to some existing systems OpenNLP (Baldridge, 2014) and ClearNLP (Choi, 2014), the SedonaNLP

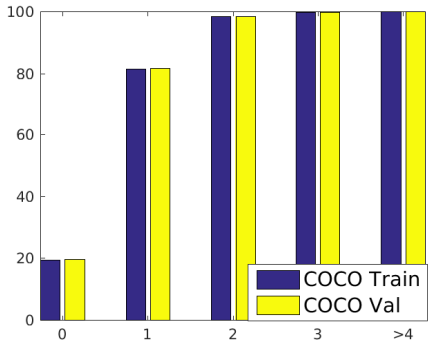
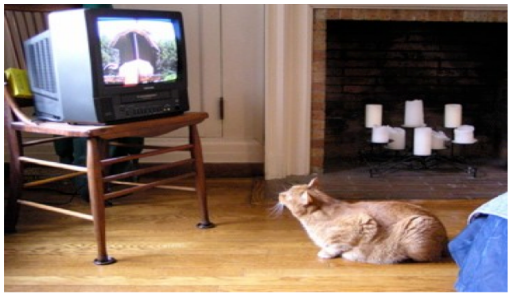



Figure 3: Accumulative Percentage of SP and SPO facts in COCO 2014 captions as number of verbs increases

platform also performs many common NLP tasks: e.g., sentence segmentation, tokenization, part-of-speech tagging, named entity extraction, chunking, dependency and constituency-based parsing, and coreference resolution. SedonaNLP itself employs both open-source components such as NLTK and WordNet, as well as internally-developed annotation algorithms for POS and clause tagging. These tasks are used to create more advanced functions such as structured fact annotation of images via semantic triple extraction. In our work, we found SedonaNLP and Clausie to be complementary for producing a set of candidate facts for possible localization in the image that resulted in successful annotations.

Varying degrees of success have been achieved in extracting and representing structured triples from sentences using <subject, predicate, object> triples. For instance, (Rusu et al., 2007) describe a basic set of methods based on traversing the parse graphs generated by various commonly available parsers. Larger scale text mining methods for learning structured facts for question answering have been developed in the IBM Watson PRISMATIC framework (Fan et al., 2010). While parsers such as CoreNLP (Manning et al., 2014) are available to generate comprehensive dependency graphs, these have historically required significant processing time for each sentence or have traded accuracy for performance. In contrast, SedonaNLP currently employs a shallow dependency parsing method that runs in some cases 8-9X faster than earlier cited methods running on identical hardware. We choose a shallow approach with high, medium, and low confidence cutoffs after observing that roughly 80% of all captions consisted of 0 or 1 Verb expressions (VX); see Fig. 3 for MSCOCO dataset (Lin et al., 2014). The top 500 image caption syntactic patterns we observed can be found on our supplemental materials (Elhoseiny et al., 2016b). These syntactic patterns are used to learn rules for automatic extraction for not only <S,P,O>, but also <S,P>, and <S,A>, where <S,P>, are subject-action facts and <S,A> are subject-attribute facts. Pattern examples and statistics for MS COCO are shown in Fig. 5.



Caption 1: A cat onthe floor watching a tv on a chair.
Caption 2: A fat cat in the living room watching the tv.

```
Caption 1 (Processing)
 1. A cat onthe floor watching a tv on a chair
                 | |
 2. A cat on the floor watching a tv on a chair.

 3. A/DT cat/NN on/IN the/DT floor/NN watching/VBG a/DT tv/NN on/IN
    a/DT chair/NN ./.
 4. NX( A/DT cat/NN )     IX( on/IN )    NX( the/DT floor/NN )
    VX( watching/VBG )
    NX( a/DT tv/NN )     IX( on/IN )    NX( a/DT chair/NN )
 5a. Subject    : NX( A/DT cat/NN ) IX( on/IN ) NX( the/DT floor/NN)
 5b. Predicate: VX( watching/VBG )
 5c. Object     : NX( a/DT tv/NN )  IX( on/IN ) NX( a/DT chair/NN )
 5d. <A cat on the floor; watching; a tv on a chair>
 6. <cat; watching; tv>
 7. <cat; on; floor>
 8. <tv; on chair>
```

```
Extracted Facts

Caption 1| nVX01,nIN02 | <S;P;O> | ID NX IN NX VX=VBG NX IN NX
       <cat/NN on/IN floor/NN; watching/VBG; tv/NN on/IN chair/NN>
       <cat/NN; watching/VBG; tv/NN>
Caption 1| nVX01,nIN02 | <s;r;o>
       <cat; on; floor>
Caption 1| nVX01,nIN02 | <s;r;o>
       <tv; on; chair>
Caption 2| nVX01,nIN01 | <S;P;O> | ID NX IN NX VX=VBG NX
       <fat/JJ cat/NN in/IN living/JJ room/NN; watching/VBG; tv/NN>
       <cat/NN in/IN room/NN; watching/VBG; tv/NN>
       <cat/NN; watching/VBG; tv/NN>
Caption 2| nVX01,nIN01 | <S;A           >
       <cat; fat>
Caption 2| nVX01,nIN01 | <S;A           >
       <room; living>
Caption 2| nVX01,nIN01 | <s;r;o>
       <fat cat; in; living room>
```

Figure 4: Examples of caption processing and <S,P,O> and <S,P> structured fact extractions.

In SedonaNLP, structured fact extraction was accomplished by learning a subset of abstract syntactic patterns consisting of basic noun, verb, and preposition expressions by analyzing 1.6M caption examples provided by the MS COCO, Flickr30K, and Stony Brook University Im2Text caption datasets. Our approach mirrors existing known art with the addition of internally-developed POS and clause tagging accuracy improvements through the use of heuristics listed

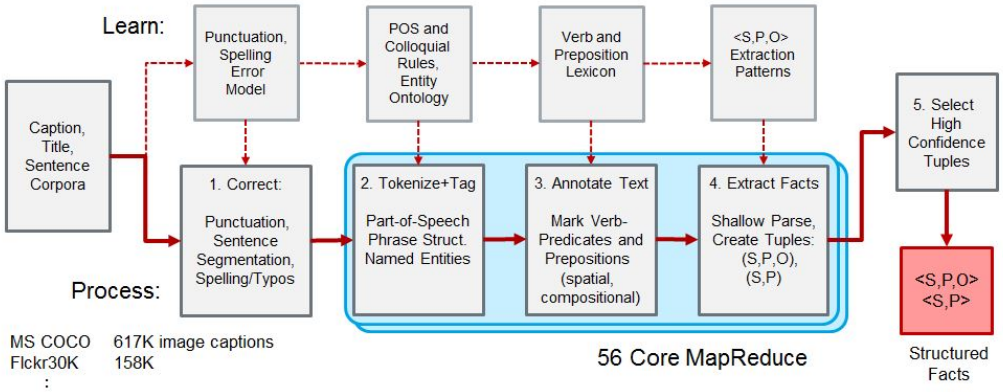



Figure 2: SedonaNLP Pipeline for Structured Fact Extraction from Captions

```
NX VX IN NX                 #   1  10.19/10.19 NX(a blue street sign) VX(sitting) IN(under) NX(a camera)
NX VX IN NX IN NX           #   2   9.73/19.92 NX(a brown cat) VX(stares) IN(at) NX(something) IN(in) NX(the field)
NX VX NX IN NX              #   3   7.05/26.97 NX(some sheep) VX(eating) NX(grass) IN(in_front_of) NX(a rock)
NX IN NX IN NX              #   4   5.65/32.62 NX(a) IN(round) NX(blue street sign) IN(with) NX(a white arrow)
NX IN NX VX IN NX           #   5   3.27/35.89 NX(a sign) IN(in_front_of) NX(a fence) VX(laced) IN(with) NX(shrubbery)
NX IN NX                    #   6   3.00/38.89 NX(a orange cat) IN(with) NX(green eyes and long whiskers)
NX VX NX IN NX IN NX        #   7   1.77/40.67 NX(a) VX(very close) NX(shot) IN(of) NX(a cat's face) IN(in_front_of)
                                               NX(the camera)
NX IN NX IN NX IN NX        #   8   1.76/42.43 NX(a toddler reaches) IN(into) NX(a bowl) IN(of) NX(grapes) IN(in)
                                               NX(a sink)
NX IN NX CC NX              #   9   1.70/44.13 NX(a bathroom) IN(with) NX(two sinks mirrors) CC(and) NX(some bottles)
NX IN NX VX NX              #  10   1.69/45.82 NX(a person) IN(on) NX(a skate board) VX(does) NX(a trick)
```

Figure 5: Examples of the top observed Noun (NX), Verb (VX), and Preposition (IN) Syntactic patterns.

below to reduce higher occurrence errors due to systematic parsing errors: (i) Mapping past participles to adjectives (e.g., stained glass), (ii) Denesting existential facts (e.g., this is a picture of a cat watching a tv.), (iii) Identifying auxiliary verbs (e.g., do verb forms).

In Fig. 4, we show an example of extracted <S,P,O> structured facts useful for image annotation for a small sample of MS COCO captions. Our initial experiments empirically confirmed the findings of IBM Watson PRISMATIC researchers who indicated big complex parse trees tend to have more wrong parses. By limiting a frame to be only a small subset of a complex parse tree, we reduce the chance of error parse in each frame (Fan et al., 2010). In practice, we observed many correctly extracted structured facts for the more complex sentences (i.e., sentences with multiple VX verb expressions and multiple spatial prepositional expressions) – these facts contained useful information that could have been used in our joint learning model but were conservatively filtered to help ensure the overall accuracy of the facts being presented to our system. As improvements are made to semantic triple extraction and confidence evaluation systems, we see potential in several areas to exploit more structured facts and to filter less information. Our full <S,P,O> triple and related tuple extractions for MS COCO and Flickr30K datasets are available in the supplemental material (Elhoseiny et al., 2016b).

## 5 Locating facts in the Image

In this section, we present details about the second step of our automatic annotation process introduced in Sec. 3. After the candidate facts are extracted from the sentences, we end up with a set $\mathbf{F}_s = \{\mathbf{f}_l^i\}, i = 1 : N_s$ for statement $s$, where $N_s$ is the number of extracted candidate fact $\mathbf{f}_l^i, \forall i$ from the statement $s$ using either Clausie (Del Corro and Gemulla, 2013) or Sedona-3.0. The localization step is further divided into two steps. The mapping step maps nouns in the facts to candidate boxes in the image. The grounding step processes each fact associated with the candidate boxes and outputs a final bounding box if localization is successful. The two steps are detailed in the following subsections.

### 5.1 Mapping

The mapping step starts with a pre-processing step that filters out a non-useful subset of $\mathbf{F}_s$ and produces a more useful set $\mathbf{F}_s^*$ that we try to locate/ground in the image. We perform this step by performing word sense disambiguation using the state-of-the-art method (Zhong and Ng, 2010).

The word sense disambiguation method provides each word in the statement with a word sense in the wordNet ontology (Leacock and Chodorow, 1998). It also assigns for each word a part of speech tag. Hence, for each extracted candidate fact in $\mathbf{F}_s$ we can verify if it follows the expected part of speech according to (Zhong and Ng, 2010). For instance, all S should be nouns, all P should be either verbs or adjectives, and O should be nouns. This results in a filtered set of facts $\mathbf{F}_s^*$. Then, each S is associated with a set of candidate boxes in the image for second- and third-order facts and each O associated with a set or candidate boxes in the image for third-order facts only. Since entities in MSCOCO dataset and Flickr30K are annotated differently, we present how the candidate boxes are determined in each of these datasets.

**MS COCO Mapping:** Mapping to candidate boxes for MS COCO reduces to assigning the S for second-order and third-order facts, and S and O for third-order facts. Either S or O is assigned to one of the MSCOCO objects or SUN scenes classes. Given the word sense of the given part (S or O), we check if the given sense is a descendant of MSCOCO objects senses in the wordNet ontology. If it is, the given part (S or O) is associated with the set of candidate bounding boxes that belongs to the given object (e.g., all boxes that contain the "person" MSCOCO object is under the "person" wordnet node like "man", 'girl', etc). If the given part (S or O) is not an MSCOCO object or one of its descendants under wordNet, we further check if the given part is one of the SUN dataset scenes. If this condition holds, the given part is associated with a bounding box of the whole image.

**Flickr30K Mapping:** In contrast to MSCOCO dataset, the bounding box annotation comes for each entity in each statement in Flickr30K dataset. Hence, we compute the candidate bounding box annotations for each candidate fact by searching the entities in the same statement from which the clause is extracted. Candidate boxes are those that have the same name. Similarly, this process assigns S for second-order facts and assigns S and O for second- and third-order facts.

Having finished the mapping process, whether for MSCOCO or Flickr30K, each candidate fact $\mathbf{f}_l^i \in \mathbf{F}_s^*$, is associated with candidate boxes depending on its type as follows.

**<S,P> :** Each $\mathbf{f}_l^i \in \mathbf{F}_s^*$ of second-order type is associated with one set of bounding boxes $\mathbf{b}_S^i$, which are the candidate boxes for the S part. $\mathbf{b}_O^i$ could be assumed to be always an empty set for second-order facts.

**<S,P,O> :** Each $\mathbf{f}_l^i \in \mathbf{F}_s^*$ of third-order type is associated with two sets of bounding boxes $\mathbf{b}_S^i$ and $\mathbf{b}_S^i$ as candidate boxes for the S and P parts, respectively.

### 5.2 Grounding

The grounding process is the process of associating each $\mathbf{f}_l^i \in \mathbf{F}_s^*$ with an image $\mathbf{f}_v$ by assigning $\mathbf{f}_l$ to a bounding box in the given MS COCO image scene given the $\mathbf{b}_S^i$ and $\mathbf{b}_O^i$ candidate boxes. The grounding process is relatively different for the two dataset due to the difference of the entity annotations.

**Grounding: MS COCO dataset (Training and Validation sets)**

In the MS COCO dataset, one challenging aspect is that the S or O can be singular, plural, or referring to the scene. This means that one S could map to multiple boxes in the image. For example, "people" maps to multiple boxes of "person". Furthermore, this case could exist for both the S and the O. In cases where either S or O is plural, the bounding box assigned is the union of all candidate bounding boxes in $\mathbf{b}_S^i$. The grounding then proceeds as follows.

**<S,P> facts:**

(1) If the computed $\mathbf{b}_S^i = \varnothing$ for the given $\mathbf{f}_l^i$, then $\mathbf{f}_l^i$ fails to ground and is discarded.

(2) If S singular, $\mathbf{f}_v^i$ is the image region that with the largest candidate bounding box in $\mathbf{b}_S^i$.

(3) If S is plural, $\mathbf{f}_v^i$ is the image region that with union of the candidate bounding boxes in $\mathbf{b}_S^i$.

**<S,P, O> facts:**

(1) If $\mathbf{b}_S^i = \varnothing$ and $\mathbf{b}_O^i = \varnothing$, $\mathbf{f}_l^i$ fails to ground and is ignored.

(2) If $\mathbf{b}_S^i \neq \varnothing$ and $\mathbf{b}_O^i \neq \varnothing$, then bounding boxes are assigned to S and O such that the distance between them is minimized (though if S or O is plural, the assigned bounding box is the union of all bounding boxes for $\mathbf{b}_S^i$ or $\mathbf{b}_O^i$ respectively), and the grounding is assigned the union of the bounding boxes assigned to S and O.

(3) If either $\mathbf{b}_S^i = \varnothing$ or $\mathbf{b}_O^i = \varnothing$, then a bounding box is assigned to the present object (the largest bounding box if singular, or the union of all bounding boxes if plural). If the area of this region compared to the area of the whole scene is greater than a threshold $th = 0.3$, then the $\mathbf{f}_v^i$ is associ-

Table 1: Human Subject Evaluation by MTurk workers %

| Dataset (responses) | Q1 | | Q2 | | Q3 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | yes | no | Yes | No | a | b | c | d | e | f | g |
| **MSCOCO train 2014 (4198)** | 89.06 | 10.94 | 87.86 | 12.14 | 64.58 | 12.64 | 3.51 | 5.10 | 0.86 | 1.57 | 11.73 |
| **MSCOCO val 2014 (3296)** | 91.73 | 8.27 | 91.01 | 8.99 | 66.11 | 14.81 | 3.64 | 4.92 | 1.00 | 0.70 | 8.83 |
| **Flickr30K Entities2015 (3296)** | 88.94 | 11.06 | 88.19 | 11.81 | 70.12 | 11.31 | 3.09 | 2.79 | 0.82 | 0.39 | 11.46 |
| **Total** | 89.84 | 10.16 | 88.93 | 11.07 | 66.74 | 12.90 | 3.42 | 4.34 | 0.89 | 0.95 | 10.76 |

Table 2: Human Subject Evaluation by Volunteers % (This is another set of annotations different from those evaluated by MTurkers)

| Volunteers | Q1 | | Q2 | | Q3 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | yes | No | Yes | No | a | b | c | d | e | f | g |
| **MSCOCO train 2014 (400)** | 90.75 | 9.25 | 91.25 | 8.75 | 73.5 | 8.25 | 2.75 | 6.75 | 0.5 | 0.5 | 7.75 |
| **MSCOCO val 2014 (90)** | 97.77 | 2.3 | 94.44 | 8.75 | 84.44 | 8.88 | 3.33 | 1.11 | 0 | 0 | 2.22 |
| **Flickr30K Entities 2015 (510)** | 78.24 | 21.76 | 73.73 | 26.27 | 64.00 | 4.3 | 1.7 | 1.7 | 0.7 | 1.18 | 26.45 |

ated to the whole image of the scene. Otherwise, $\mathbf{f}_l^i$ fails to ground and is ignored.

**Grounding: Flickr30K dataset** The main difference in Flickr30K is that for each entity phrase in a sentence, there is a box in the image. This means there is no need to have cases for single and plural. Since in this case, the word "men" in the sentence will be associated with the set of boxes referred to by "men" in the sentences. We union these boxes for plural words as one candidate box for "men"

We can also use the information that the object box has to refer to a word that is after the subject word, since subject usually occurs earlier in the sentence compared to object. We union these boxes for plural words.

**<S,P> facts:**

If the computed $\mathbf{b}_S^i = \varnothing$ for the given $\mathbf{f}_l^i$, then $\mathbf{f}_l^i$ fails to ground and is discarded. Otherwise, the fact is assigned to the largest candidate box in if there are multiple boxes.

**<S,P, O> facts:** <S,P, O> facts are handled very similar to MSCOCO dataset with two main differences.

a) The candidate boxes are computed as described for the case of Flickr30K dataset.

b) All cases are handled as single case, since even plural words are assigned one box based on the nature of the annotations in this dataset.

## 6 Experiments

### 6.1 Human Subject Evaluation

We propose three questions to evaluate each annotation: (Q1) Is the extracted fact correct (Yes/No)? The purpose of this question is to evaluate errors captured by the first step, which extracts facts by Sedona or Clausie. (Q2) Is the fact located in the image (Yes/No)? In some cases, there might be a fact mentioned in the caption that does not exist in the image and is mistakenly considered as an annotation. (Q3) How accurate is the box assigned to a given fact (a to g)? a (about right), b (a bit big), c (a bit small), d (too small), e (too big), f (totally wrong box), g (fact does not exist or other). Our instructions on these questions to the participants can be found in this url (Eval, 2016).

We evaluate these three questions for the facts that were successfully assigned a box in the image, because the main purpose of this evaluation is to measure the usability of the collected annotations as training data for our model. We created an Amazon Mechanical Turk form to ask these three questions. So far, we collected a total of 10,786 evaluation responses, which are an evaluation of 3,595 ($\mathbf{f}_v, \mathbf{f}_l$) pairs (3 responses/ pair). Table 2 shows the evaluation results, which indicate that the data is useful for training, since≈83.1% of them are correct facts with boxes that are either about right, or a bit big or small (a,b,c). We further some evaluation responses that we collected from volunteer researchers in Table 2 showing similar results.

Fig. 6 shows some successful qualitative results that include four extracted structured facts from MS COCO dataset (e.g., <person, using, phone>, <person, standing>, etc). Fig 7 also show a negative example where there is a wrong fact among the extracted facts (i.e., <house, ski>). The main reason for this failure case is that "how" is mistyped as "house"; see Fig 7. The supplementary materials (Elhoseiny et al., 2016b) includes all the captions of these examples and also additional qualitative examples.

### 6.2 Hardness Evaluation of the collected data

In order to study how the method behave in both easy and hard examples. This section present statistics of the successfully extracted facts and relate it to the hardness of the extraction of these facts. We start by defining hardness of an extracted fact in our case and its dependency on the fact type. Our method collect both second- and third-order facts. We refer to candidate subjects as all instances of the entity in the image that match the subject type of either a second-order fact <S,P> or a third-order fact <S,P,O>. We refer to candidate objects as all instances in the image that match the object type of a third-order fact <S,P,O>. The selection of the candidate subjects and candidate objects is a part of our method that we detailed in Sec 5. We define the hardness for second order facts by the number of candidate subjects and the hardness of third order facts by the number of candidate subjects multiplied by the number of candidate objects.


Figure 6: Several Facts successfully extracted by our method from two MS COCO scenes

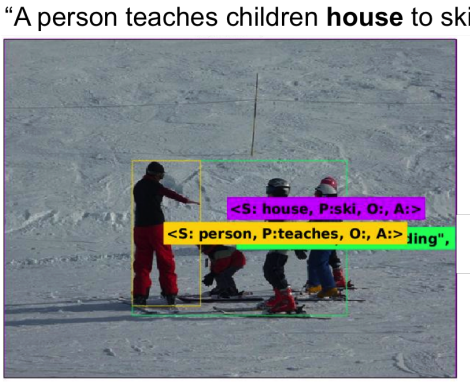



Figure 7: An example where one of the extracted facts are not correct due to a spelling mistake

In Fig 8 and 9, the Y axis is the number of facts for each bin. The X axis shows the bins that correspond to hardness that we defined for both second and third order fats. Figure 8 shows a histogram of the difficulties for all Mturk evaluated examples including both the successful and the failure cases. Figure 9 shows a similar histogram but for but for subset of facts verified by the Turkers with Q3 as (about right). The figures show that the method is able to handle difficulty cases even with more than 150 possibilities for grounding. We show these results broken out for MSCOCO and Flickr30K Entities datasets and for each fact types in the supplementary materials (Elhoseiny et al., 2016b).

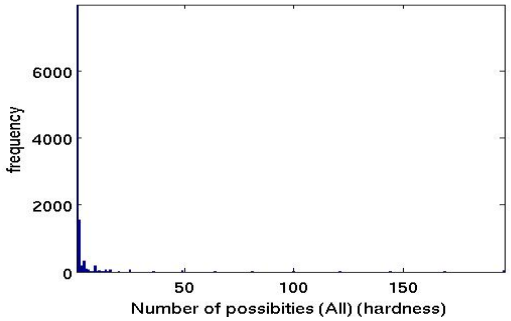



Figure 8: (All MTurk Data) Hardness histogram after candidate box selection using our method

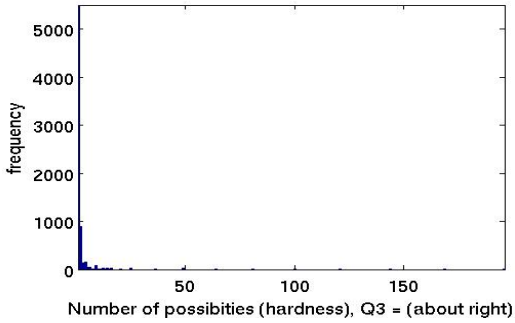



Figure 9: (MTurk Data with Q3=about right)Hardness histogram after our candidate box selection

## 7 Conclusion

We present a new method whose main purpose to collect visual fact annotation by a language approach. The collected data help train visual system systems on the fact level with the diversity of facts captured by any fact described by an image caption. We showed the effectiveness of the proposed methodology by extracting hundreds of thousands of fact-level annotations from

MSCOCO and Flickr30K datasets. We verified and analyzed the collected data and showed that more than 80% of the collected data are good for training visual systems.

## References


Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV*.

Jason Baldridge. 2014. The opennlp project. *URL: http://opennlp.apache.org/index.html,(accessed 2 February 2014)*.

Xinlei Chen, Ashish Shrivastava, and Arpan Gupta. 2013. Neil: Extracting visual knowledge from web data. In *ICCV*.

Jinho D Choi. 2014. Clearnlp.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: clause-based open information extraction. In *WWW*.

Santosh K Divvala, Alireza Farhadi, and Carlos Guestrin. 2014. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*.

Mohamed Elhoseiny, Scott Cohen, Walter Chang, Brian Price, and Ahmed Elgammal. 2016a. Sherlock: Scalable fact learning in images.

Mohamed Elhoseiny, Scott Cohen, Walter Cheng, Brian Price, and Ahmed Elgammal. 2016b. Autoomatic annotation of structured facts in images- supplementary materials. `https://www.dropbox.com/s/22m6jxvtqhhg10q/supplementary.zip?dl=0`. [Online; accessed 19-Nov-2015].

SAFA Eval. 2016. Safa eval instructions. `https://dl.dropboxusercontent.com/u/479679457/Sherlock_SAFA_eval_Instructions.html`. [Online; accessed 02-March-2016].

James Fan, David Ferrucci, David Gondek, and Aditya Kalyanpur. 2010. Prismatic: Inducing knowledge from a large scale lexicalized relation resource. In *NAACL HLT*. Association for Computational Linguistics.

Andrej Karpathy, Armand Joulin, and Fei Fei F Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2015. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*. Springer.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*.

Bryan Plummer, Liwei Wang, Chris Cervantes, Juan Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *NIPS*.

Delia Rusu, Lorand Dali, Blaz Fortuna, Marko Grobelnik, and Dunja Mladenic. 2007. Triplet extraction from sentences. In *International Multiconference" Information Society-IS*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator.

Jianxiong Xiao, James Hays, Krista Ehinger, Aude Oliva, Antonio Torralba, et al. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *ACL*.