ENTITY-BASED SCENE UNDERSTANDING

BY

CHRISTOPHER MICHAEL CERVANTES

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Adviser:

Associate Professor Julia Hockenmaier

ABSTRACT

Unifying multiple descriptions to determine the details of an everyday event can be a challenging task for humans. Though incorporating other modalities like images or videos can help humans unify such descriptions, this remains a challenging task for computational systems. We define *entity-based scene understanding* as the task of identifying the entities in a visual scene from multiple descriptions. This task subsumes coreference resolution, bridging resolution, and grounding to produce mutually consistent relations between entity mentions and groundings between mentions and image regions. Using neural classifiers and integer linear program inference, we show that grounding is improved when forced to conform to relation predictions. We introduce the Flickr30k Entities v2 dataset, and show how our methods can be used to automatically generate similarly rich annotations for the MSCOCO dataset.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ILP	Integer Linear Programming
LSTM	Long-Short Term Memory network (see Hochreiter and Schmidhuber (1997))
MASI	Measuring Agreement on Set-valued Items (see Passonneau (2006))
NLP	Natural Language Processing
NN, NNS, NNP, NNPS	noun (singular or mass), noun (plural), proper noun (singular), and proper noun (plural), respectively (see Marcus et al. (1993))
NP	Noun Phrase
nsubj, dobj	nominal subject and direct object, respectively (see De Marneffe and Manning (2008))
POS	Part of Speech
PP	Prepositional Phrase
VP	Verb Phrase

CHAPTER 1: INTRODUCTION

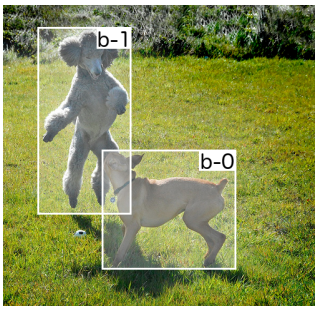
When a person describes something that they saw during their day, it is fair to assume that the description accurately represents what happened. Without any additional information, most people would take the description at face value; what the person saw must be what happened. If other people describe the same event, however, even the most mundane, everyday occurrence can become clouded with uncertainty. Even ignoring the role that memory might play in such a scenario, different people may focus on different aspects of a scene, may describe similar aspects in different ways, and may even disagree on basic details.

Reconciling multiple descriptions of the same scene is a difficult task for humans. When witnesses fundamentally disagree – on what happened during a routine traffic stop, or who was present at a meeting and what was discussed – significant external data may be required, from photos to videos to travel logs and additional eyewitness accounts. Even in a less extreme scenario where each witness is trying to honestly and accurately describe what they saw, the act of unifying those different perspectives, each with differing language, remains a challenging task.

Images can help simplify this task. When multiple people describe an event for which there exists a picture, it becomes much easier for people to understand what was meant by the witnesses, even if it isn't actually what was said. In some ways, images can serve as a compliment to language: where images capture the the full visual context of a scene, descriptions tend to focus on the important aspects – some of which may not be visual – and omit context.

For computational systems, understanding a scene in this setting remains a very challenging task. Such a system would have to extract meaning from both the descriptions and the images, and as such would lie at the intersection of *natural language processing* (NLP) and *computer vision*. Moreover, this extracted meaning would have to be unified to produce a single representation – a single sense of understanding – for the scene, incorporating the entities that were involved, their attributes, relations, and the event that took place.

In this thesis, we take a first step toward this broad goal by considering an entity-centric view of understanding a scene. Given an image and a set of sentences describing it (as in



Two dogs$_1$[b-0, b-1], the gray poodle$_2$[b-1] high in the air$_-$, play on the grass$_3$.
One gray poodle$_2$[b-1] jumping in the air$_-$ in front of another tan dog $_4$[b-0]
A gray "labradoodle"$_2$[b-1] jumps over another large dog$_4$[b-0].
Two dogs$_1$[b-0, b-1] are chasing each other$_1$[b-0, b-1] in a yard$_3$.
Two dogs$_1$[b-0, b-1] playing in grass$_3$.

chain 2 $\subset$ chain 1
chain 4 $\subset$ chain 1

Figure 1.1: A Flickr30k Entities v2 image. Coreferent mentions are color-coded and share subscripts, and groundings are shown with superscripts

Figure 1.1), we define *entity-based scene understanding* as the task of identifying the set of entities in the scene, where an entity is a non-empty set of coreferent mentions (noun-phrase chunks) and a (possibly empty) set of image regions. Since image captions often refer to plural entities (e.g. "Two dogs"), it is also important to identify subset relations between entities (e.g. "One gray poodle" is a subset of "Two dogs").

Entity-based scene understanding is a composite task, incorporating coreference resolution, bridging anaphora resolution, and grounding. Supervised approaches for this task, therefore, require the presence of rich, high-quality image caption data. In support of this thesis and similar vision and language tasks, we helped to develop the Flickr30k Entities (Young et al., 2014; Plummer et al., 2015) and Flickr30k Entities v2 datasets.

Given such data, we are able to use supervised approaches that leverage local classification and global integer linear programming (ILP) inference. We show that while coreference and bridging – combined into a task we refer to as *relation prediction* – and grounding can be performed separately, a joint approach can help grounding significantly and produce mutually consistent relation and grounding predictions.

On its own, entity-based scene understanding may seem like a synthetic task that requires very expensive, specialized data. Our approach, however, can be applied to similar image caption datasets in order to automatically generate these rich annotations. We show that when our approach is applied to the MSCOCO dataset (Lin et al., 2014), we can produce high-quality coreference, bridging, and grounding annotations.

CHAPTER 2: BACKGROUND

The goal of entity-based scene understanding is to correctly identify entities and the set relations between them from text and images. This is a composite task, combining coreference resolution, bridging resolution, and text-to-image grounding. Therefore, we must first review these established tasks, noting the ways in which the tasks and the approaches used to address them differ from our own. Finally, we note that while there are many vision and language datasets, there are comparatively few that contain the rich, high-quality annotations that our approach requires, motiving our need to augment existing data.

2.1 RELATION PREDICTION

Identifying entities from text is crucial to understanding the scene as a whole. One common mechanism to do so requires a) finding entity *mentions* – non-overlapping noun phrase (NP) spans that refer to some entity in the scene – and b) partitioning those mentions into equivalence classes, or *coreference chains*. This process is referred to as *coreference resolution*, and is a well-studied task in the NLP literature.

Typically, coreference resolution links mentions to the intra-document antecedents to which they refer. This was traditionally accomplished by performing global inference over pairwise decisions (Soon et al., 2001; Ng and Cardie, 2002; Punyakanok et al., 2004; Bengtson and Roth, 2008; Chang et al., 2011), but more recent work has added mentions to chains using chain-level features (Lee et al., 2011; Wiseman et al., 2015, 2016; Clark and Manning, 2015, 2016a,b), and even newer end-to-end neural models have yielded state of the art performance (Lee et al., 2017). Though these approaches differ, they make two broad assumptions about coreference resolution: a) that coreference chains can be built up by finding the best previously occurring referent in the document (Martschat and Strube, 2015), and b) that documents are long spans of text containing multiple sentences (e.g. news articles).

In the cross-caption setting, mentions may corefer with others across independent but parallel captions. While similar to cross-document coreference resolution – in which coreference chains may contain mentions from multiple documents (Singh et al., 2011; Dutta and

Weikum, 2015) – image captions are single sentences written about everyday scenes. Though neither can rely on the former assumption, above, cross-document coreference still relies on the latter. In practice, this means that while cross-document coreference can still leverage the anaphoric coreference, named entities, and discourse features (which are common in standard newswire datasets (Singh et al., 2011; Pradhan et al., 2012)), cross-caption coreference cannot. As a result, standard tools and standard approaches to coreference resolution aren't appropriate for the image caption setting, which is a known problem when trying to perform coreference resolution in new domains (Guha et al., 2015).

Previous work on unsupervised coreference resolution in the image caption domain (Hodosh et al., 2010) has also needed to accomodate these domain challenges. Our work differs in two important ways. First, we have labeled data with which we can use supervised approaches to coreference resolution. More importantly, however, we treat coreference resolution as part of a larger task: that of predicting relations between mentions. This relation prediction task is primarily inspired by Gardent et al. (2004), which defines a set of definite description relations that establish entity coherence. We define *relation prediction* as the task of determining coreference and bridging relations between all mentions (rather than simply definite descriptions). In this way, we not only extend coreference resolution to the image caption domain, but we augment the task meaningfully with the inclusion of bridging.

Both coreference and bridging anaphora resolution operate over the set of entity mentions, and both aim to link mentions together: in the former determining if mentions are coreferent, in the latter determining if mentions share set membership or meronymy relations. Approaches to coreference and bridging resolution also tend to be similar, often sharing the procedure of global inference over pairwise decisions (Hou et al., 2013). These similarities, along with the fact that coreference and bridging are mutually exclusive – that is, a mention cannot refer to a subset of another if both mentions refer to the same entity – enables us to fold both into the relation prediction task: for each ordered pair of mentions, we predict coreference, subset, superset, or no relation[1] and unify these predictions at inference time.

[1]Though meronymy is also a bridging relation, we leave this for future work.

2.2 GROUNDING

We frame grounding as a reference resolution task, where we associate entity mentions to the image regions they describe. While this is most similar to work that associates phrases with ground truth image regions (Iida et al., 2011; Krishnamurthy and Kollar, 2013; Kennington and Schlangen, 2015), we work with a much larger, more diverse dataset of everyday scenes. While other grounding systems predict both salient image regions and their associations to phrases (Plummer et al., 2015; Karpathy and Fei-Fei, 2015; Fukui et al., 2016; Zhu et al., 2016; Hu et al., 2016; Plummer et al., 2017), these approaches differ from ours in two important ways. First, these systems typically identify a single image region corresponding to a set of entities ("Two people" is ground to a region around both people), where our approach identifies individual regions and assigns each to the mention describing the set. Second, these approaches consider only individual captions in isolation, rather than leveraging the information present in parallel captions.

Our work is most similar to Kong et al. (2014), which grounds phrases from multiple sentences to 3D image regions. Though their approach performs coreference resolution, they do so over single, multiple-sentence descriptions (paragraphs), rather than multiple, single-sentence descriptions (parallel captions). Further, they ground phrases to cuboids of a set number of object categories, rather than our diverse set of everyday image regions.

We combine relation prediction and grounding with joint inference, which operates over both gold entity mentions and image regions[2]. This joint approach enables us to make predictions about the scene as a whole, taking into account the ways in which coreference, set membership, and grounding interact with one another to produce a single, consistent representation of the entities in a scene.

2.3 DATA

Entity-based scene understanding is a task that requires reasoning over both images and natural language, and thus requires data that pairs images and text. Specifically, our defi-

[2]We use gold mentions and regions because of the inherent difficulty of grounding, leaving their prediction for future work

nition of this task requires data that contains

1. Everyday images accompanied by natural language descriptions
2. Independently-written sentences describing the same scene
3. Object-level annotations, which localize entities in the image

While not technically required for our task, our supervised approach also requires labeled data. Thus, in addition to the criteria above, the ideal data must also contain

4. Grounding annotations, which associate entity mentions to the image regions they describe
5. Coreference annotations, that link entity mentions within and across captions when they refer to the same entity or set of entities

Typically, vision and language datasets take the form of image caption data, where an image is described by one or more captions. Such is the case in the UIUC Sentence dataset (Rashtchian et al., 2010), which pairs 1k images from PASCAL VOC (Everingham et al., 2010) with five captions per image. While this dataset meets the first three criteria – images are associated with parallel captions and object annotations for PASCAL's 20 object categories – it does not meet the last two: there are neither grounding nor coreference annotations.

More recently, the ReferIt dataset (Kazemzadeh et al., 2014) links objects with referring expressions: short descriptions necessary to uniquely identify the object in its image. While this dataset contains images and descriptions (criterion 1), object-level annotations (2), and grounding annotations (4), it does not contain parallel descriptions formed as complete sentences about the image's contents as a whole. Similarly, the Visual Genome dataset (Krishna et al., 2017) contains many everyday images (108k), with grounding annotations, but like ReferIt their descriptions are short phrases, not parallel sentences. Moreover, their grounding annotations are too noisy for the subtle semantics present in a task like entity-based scene understanding (e.g. the phrase "a boy wearing jeans" may refer to different – but colocated – image region annotations than "a little boy").

When this project was undertaken, there did not exist a dataset that met all of the criteria above. The closest were Flickr30k (Young et al., 2014) and MSCOCO (Lin et al.,

2014). Both contain everyday images annotated with multiple, independently-written image captions (criteria 1 and 2), but only MSCOCO contained any object-level annotations (3). Neither contained grounding annotations, and neither contained coreference annotations.

In order to support both entity-based scene understanding and similar vision and language tasks, new datasets had to be created: Flickr30k Entities (Plummer et al., 2015) and Flickr30k Entities v2.

CHAPTER 3: IMAGE CAPTION DATA

The entity-based scene understanding task requires the presence of high-quality, multi-modal data. Specifically, we need images described by multiple captions, where we have annotations linking coreferent entity mentions within and across captions and grounding annotations linking mentions to image regions that tightly bound the scene's visual entities.

The Flickr30k Entities dataset (Plummer et al., 2015) was designed to fulfill all of these requirements. Building on the Flickr30k dataset (Young et al., 2014) , Flickr30k Entities added both coreference and grounding annotations. During the annotation process, however, it became clear that there was significant noise present in the data, precluding its use for a task as complex as entity-based scene understanding. Therefore, we introduce the Flickr30k Entities v2 dataset – a refinement of the original Flickr30k Entities dataset – which includes more accurate chunking, coreference labeling, and box associations.

Finally, we note that an important by-product of our approach to entity-based scene understanding is the ability to produce annotations given a dataset with images described by multiple captions. The MSCOCO dataset (Lin et al., 2014) is a perfect candidate for such annotations, as it contains a wide range of everyday images described by multiple captions, but does not contain any grounding or coreference annotations. To evaluate the degree to which our automatically generated annotations align with human annotations, we also annotate a small subset of MSCOCO to include these Flick30k Entities v2 annotations.

3.1 FLICKR30K ENTITIES

The Flickr30k Entities dataset[1] (Plummer et al., 2015) builds on the Flickr30k dataset (Young et al., 2014) which contains ∼32k images that are each associated with five independently written captions. Flickr30k Entities adds coreference and grounding annotations to those images and captions. To do so, Flickr30k Entities contains two important abstractions: *mentions* and *bounding boxes*.

[1]Much of the Flickr30k Entities dataset design and collection predates the author's involvement in the project; the author acknowledges the main contributions of Bryan Plummer, Liwei Wang, Juan Caidedo, Julia Hockenmaier, and Svetlana Lazebnik

In this data, mentions are minimal NP chunks that describe a (possibly singleton) set of entities in the scene. Thus, "[The man] in [the tan jacket]..." is two mentions, shown in brackets. The exception to these minimal chunks are *XofY constructions*: mentions with an internal "of" that refer to a single set of entities (e.g. "a pile of sand", "a group of people"). Mentions may be *nonvisual* when they do not or cannot refer to a visual entity, like "time", "the background" or "the camera" taking the picture. Typically, however, mentions refer to some visual entity in the scene. These visual entities are categories into eight lexical types: *people*, *animals*, *clothing*, *colors*, *bodyparts*, *vehicle*, *instruments* and *scene*.

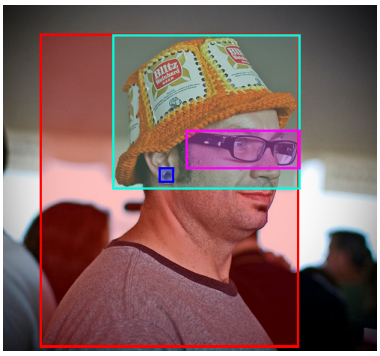
Flickr30k Entities provides coreference annotations for all visual mentions[2]; when two mentions refer to the same visual set of entities, they are assigned to the same coreference chain. In this way, coreference chains are synonymous with sets of entities. Flickr30k Entities leveraged this by annotating chains with grounding annotations in the form of bounding box associations. That is, chains are associated with the bounding boxes – rectangular regions that tightly bound an entity in the image – to which they refer. In cases where a chain refers to a set of entities (e.g. "two dogs"), it is associated with a set of boxes for each individual entity (e.g. two boxes, one for each dog). Sometimes, however, a chain refers to a number of entities large enough that it would have made bounding box annotations unreasonable to collect (e.g. "a crowd of people"). In these cases, a single box appears around the set.

Three example images are shown in Figure 3.1, along with their coreference and grounding annotations.

3.1.1 Flickr30k Entities Annotation Process

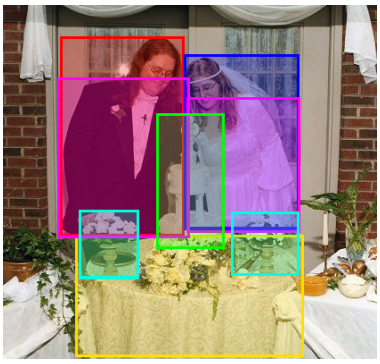
The annotation process for Flickr30k Entities was divided into two main stages, both completed on Amazon Mechanical Turk (AMT). In the first, coreference annotations were produced by collecting binary coreference links between pairs of mentions. In this stage, workers were shown an image, two mentions, and the captions from which the mentions originated. The workers were then asked whether the mentions referred to the same entity or set of entities. To reduce the number of necessary annotations, three simplifications were

[2]While pronouns are visual, Flickr30k Entities did not annotate pronouns with coreference or grounding labels.



A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

During a gay pride parade in an Asian city, some people hold up rainbow flags to show their support.
A group of youths march down a street waving flags showing a color spectrum.
Oriental people with rainbow flags walking down a city street.
A group of people walk down a street waving rainbow flags.
People are outside waving flags.



A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.

Figure 3.1: Example annotations from Flickr30k Entities. For each image's captions, coreference chains and corresponding bounding boxes are color-coded. In the leftmost image, each chain refers to a single entity and a single bounding box. Nonvisual scene or event terms (e.g. "outside" or "parade") have no box. In the middle image, people (red) and flags (blue) are chains referring to multiple entities and thus multiple boxes. In the rightmost image, the blue chain refers to the bride, the red refers to the groom, and the purple chain refers to both.

made.

- Rather than annotate the entire set of possible links between mentions (an $|M| \times |M|$ size set), representative mentions from chains ($m \in c$) are used as chains are built. In this way, the binary annotation between the unannotated mention and the representative mention of the chain holds for all other mentions in the chain.
- Mentions from the same caption cannot be coreferent. While this is not technically true, this assumption held often enough to reduce the necessary number of annotations.
- Mentions of the same lexical type cannot be coreferent.

In order to collect bounding box annotations, the Flickr30k Entities annotation pipeline contained four AMT tasks: (1) Box Requirement, (2) Box Drawing, (3) Box Quality, and

(4) Box Coverage. In each, workers were shown an image, caption, and one mention to represent a coreference chain.

In the Box Requirement task, workers were asked if the representative mention required a box to be drawn. This task should have identified nonvisual mentions[3], though due to the annotation requirements at the time chains referring to the entire scene – that is, chains for which the associated bounding box would surround the whole image – were also marked as not requiring a box.

The Box Drawing task required workers to draw a single bounding box for a given chain. In cases where a chain referred to a set of entities, an image would pass through the Box Drawing task multiple times.

The Box Quality task verified whether a box drawn in the Box Drawing task a) accurately surrounded the entity to which the chain refers, and b) tightly surrounded that entity (that is, did not include spurious image regions).

Finally, the Box Coverage task asked workers to determine if all boxes described by the chain's representative mention had been drawn. If not, the image was sent back to the Box Drawing task.

3.1.2 Flickr30k Entities Statistics

Flickr30k Entities contains 3.2 mentions per caption (16 per image), clustered into 7.7 coreference chains per image. Each chain is composed of an average of 2.1 mentions, and is associated with an average of 1.1 boxes (there are a total of 8.7 boxes per image).

3.2 FLICKR30K ENTITIES V2

The Flickr30k Entities annotation process added rich, complex coreference and grounding annotations to the Flickr30k dataset. This process, however, resulted in significant noise that precluded the dataset's use in complex natural language tasks. To address this noise, we manually refined the Flickr30k Entities dataset using expert annotators. This refinement

[3]Nonvisual mentions were primarily identified with a lexicon in Flickr30k Entities

– Flickr30k Entities v2 – contains more accurate chunking, coreference and grounding annotations for a portion of the training splits and all of the development and testing data (using the same splits as defined in Young et al. (2014)).

3.2.1 Annotation Errors in Flickr30k Entities

Annotation errors in the original Flickr30k Entities manifest in three main ways: chunking, coreference, and grounding.

Chunking errors refer to incorrectly partitioned captions. Here, a chunk is defined as a non-overlapping text span encapsulating a phrase (e.g. noun phrase, verb phrase, prepositional phrase). These errors were primarily caused by the automated pre-processing tools used prior to human annotation. Though all chunking errors are problematic, the most important chunking errors were those around mentions (NP chunks). Mis-chunked mentions took four forms

- Extraneous Word(s): Mentions that included additional word(s) and shouldn't have. This most often occurred in the form '[subject verb]' where the proper chunking is '[subject] verb' (e.g. "[The woman ran]" instead of "[The woman] ran").
- Missing Word(s): Mentions that did not include word(s) and should have. This was usually in the form '[noun] verbal-noun' where the proper chunking is '[noun verbal-noun]' (e.g. "[The ballet] practice" instead of "[The ballet practice]").
- Split Mentions: Mentions that have been chunked separately but should have been combined. These usually occurred with long, descriptive mentions, such that part of the adjective sequence is chunked separately (e.g. '[This long, drawn out], [illustrative description]').
- Merged Mentions: Mentions that have been chunked together and shouldn't have been, often in the form '[mention and mention]' where the correct chunking is '[mention] and [mention]' (e.g. "[A white helmet and green and black jacket]" instead of "[A white helmet] and [green and black jacket]").

These types were not mutually exclusive, as is the case for "a red top and white and brown skirt crosses". Here, this mention was merged (should be: "[a red top] and [white and brown skirt crosses]"), but even after this correction one of the chunks has an extraneous word (should be: "white and brown skirt").

Coreference errors occurred when coreference chains were incorrectly constructed, which can happen in two ways: when two or more mentions have been partitioned into multiple chains when they refer to a single set of entities, and when two or more mentions have been incorrectly clustered into the same chain when they refer to separate sets of entities. While the former was common in cases that required careful attention (e.g. similar or identical mentions like "a man" and "man" referred to separate entities), the latter often occurred in set membership cases (e.g. "a dog" was marked as coreferent of "a group of dogs").

Grounding errors – that is, errors in the association between chains and bounding boxes – appeared in three forms.

- Missing Box(es): A chain referred to an image region that did not appear in the original Flickr30k Entities bounding box annotations, necessitating a new box to be drawn.
- Unassociated Box(es): A chain should have been associated with a bounding box that appeared in the data but wasn't.
- Spurious Box(es): A chain was associated with a bounding box but should not have been.

These errors were not mutually exclusive (e.g. many chains had both unassociated and spurious boxes).

Each of these errors had the potential to impact the others. Mis-chunked mentions may have caused errors in the coreference annotations, and coreference errors almost always caused grounding errors. This combination was relatively common. Consider, for example, two mentions: "two men walk" and "two people". In such a case, the original annotators would have likely marked these as not-coreferent (due to confusion, apathy, or rigid adherence to the instructions, since – technically speaking – "two men walk" and "two people" are not referring to the same thing). Therefore, these mentions would have been partitioned into separate chains, and would have had separate boxes drawn for them. Due to the way that the Flickr30k Entities annotation pipeline merged boxes, it is possible that these two box sets may have not been disjoint. Thus, the Flickr30k Entities v2 refinement effort would have required fixing the chunking – to "two men" and "two people" – and then the coreference errors – putting both mentions into a single chain – and finally the grounding errors – choosing the best set of boxes from the two possibly disjoint sets to describe the two men.

Correcting these kinds of annotation errors in Flickr30k Entities was a non-trivial task, requiring that careful attention be paid to the content of the images, the text in the captions, and the way in which the chains and boxes were associated. As a result, the v2 refinement effort was a lengthy, manual process that could not have been accomplished with automated tools or AMT workers. Instead, we used automated tools to make coarse judgements about which images needed refinement and experts to refine the actual annotations.

3.2.2 Flickr30k Entities v2 Annotation Queue

Correcting the kinds of annotation errors present in Flickr30k Entities was a costly, time consuming process that would have been unreasonable to perform over the entire dataset. Therefore, we used a set of heuristics to create an annotation queue: that is, a list of images arranged in order of most likely to be in need of refinement to least. In this way, we were able to ensure that our efforts were focused on removing the most amount of error from the dataset, given limited resources.

Though there are three broad categories of annotation errors in Flickr30k Entities, the heuristic filters only took the first two into account (chunking and coreference), because while there are cues in captions to help detect the presence of these kinds of errors, it is often not possible to detect grounding errors without looking at the image and available boxes.

In order to detect chunking errors, we looked at the following

- Atypical POS: Mentions for which the last word has a part of speech other than a noun (NN, NNS, NNP, NNPS).
- Atypical Dependencies: Mentions for which the Stanford Dependency Parser (De Marneffe et al., 2006) produces arcs originating in the mention that are not subject or object dependencies (e.g. nsubj, dobj).
- Long mentions: Mentions longer than 40 characters.

The first two rules were intended to identify extraneous and missing words, the first more directly than the second. It was often the case that, when a mention had a last word that wasn't a noun, the mention was mis-chunked (either due to an extraneous or missing word). The second heuristic was based on the same intuition, though the false positive rate was

higher, both because mentions could have dependencies other than subjects and objects and because the Stanford Dependency Parser was often in error when used on image captions. Finally, the third heuristic was a naive way to detect merged mentions (our preliminary analysis suggested that mentions longer than 40 characters were typically in error).

To detect coreference errors, we looked at the following

- Non-coreferent first mentions which shared a lexical type and plurality.
- Heterogeneously typed chains (coreference chains with mentions with different lexical types).

In Flickr30k Entities, first mentions – that is, the first mention in a caption – are typically the main subject of their caption and, given the relative simplicity of most captions describing the same image, are usually coreferent. Thus, non-coreferent first mentions – particularly those that share a lexical type – were likely candidates for chains that needed to be merged. To identify chains that needed to be split, we looked at chains that had mentions of different lexical types[4], which was a strong indication that a chain contained spurious mentions.

Combining these heuristics naively (that is, taking the number of instances of these cases over the total number of mentions or chains) enabled us to assign a [0-1] score to each image which we treated as a confidence that the image was in need of correction. While none of these heuristics were particularly accurate, their total provided a reasonable mechanism to identify error-prone images.

In order to ensure that experiments using Flickr30k Entities v2 could be evaluated accurately, we chose to re-annotate all of the development and test splits of Flickr30k Entities. In addition, we used the above heuristics to identify the ∼3k training images most in need of correction. This number was both reasonable, given the resources available, and the inflection point at which the images in the annotation queue began to be approximately equal in their degree of error (according to our analysis, the vast majority of images needed some kind of correction).

[4]It is worth noting that the annotation procedure defined in Section 3.1.1 specified that mentions of different types could not be coreferent; nevertheless, this was a notable problem in the final Flickr30k Entities dataset

3.2.3 Flickr30k Entities v2 Annotation Process

The Flickr30k Entities v2 refinement effort took place in two phases. In the first, spelling errors were corrected by automatic filters and manual annotation. In the second, human annotators [5] were trained to refine the original annotations, where the training focused on the kinds of errors described in Section 3.2.1. In addition, annotators were instructed to annotate all pronouns with coreference labels (which were omitted in Flickr30k Entities) and pay special attention to nonvisual mentions (which were not to take any coreference annotations).

The refinement was completed using a web interface that enabled users to change chunk boundaries, chunk types, coreference assignments, and chain / box associations. No new boxes were drawn as part of the v2 refinement, largely because of the additional effort that such an undertaking would have required.

The annotation web interface showed annotators an image and its captions with emphasized mentions and color-coded coreference chain and bounding box information. Annotators assigned mentions to new or existing coreference chains, assigned tokens to new or existing chunks, and added associations between bounding boxes and new coreference chains. Annotators did not draw new boxes, nor did they change the associations between bounding boxes and existing coreference chains. Where a preexisting association was in error, annotators created a new chain and assigned boxes to it.

To ensure quality, about half of the refinements made by these annotators were reviewed by the author, and about half of the total number of refinements were completed solely by the author (the most difficult cases required expert annotation that extended beyond the training, revolving around particularly complex images or difficult semantic distinctions). Furthermore, a random set of ∼800 images were reviewed by multiple annotators to give some measure of overall dataset quality, with respect to the refinement.

Comparing these annotations using the MASI distance metric (Passonneau, 2006) with standard inter-annotator agreement, our annotators strongly agreed with one another with

[5] Undergraduate students recruited for this purpose

scores of $\kappa = 0.85$[6] and $\alpha = 0.84$ [7]. For completeness, we also computed agreement with a binary distance metric, rather than MASI, which resulted in scores of $\kappa = 0.69$ and $\alpha = 0.68$, showing that our annotators agree even in a setting less appropriate for measuring coreference annotations.

3.2.4 Flickr30k Entities v2 Discussion

The v2 refinement contains more accurate chunking, coreference labeling, and box associations for all images in the development and test sets, along with ∼3k training images. Of these ∼9k reviewed images, 35% required some chunking change, 38% required changing bounding box associations, and 90% required a change to the coreference annotations.

In aggregate, these new annotations provide cleaner, more reliable data. As an example of the kinds of changes that were made, consider Figure 3.2.

Original Annotations:
A performance going on **that consists** of **a woman** and **two men** standing by **trains** and talking .
Two men and **a woman** are performing on **stage** in a **“ Thomas the Tank Engine** ” play .
A group of actors is performing **a Thomas the Tank** themed play .

v2 Annotations:
A performance going on that consists of **a woman** and **two men** standing by **trains** and talking .
Two men and **a woman** are performing on **stage** in **a “ Thomas the Tank Engine ” play** .
A group of actors is performing **a Thomas the Tank themed play** .

Figure 3.2: Flickr30k Entities original and v2 annotations, for three of the five captions

This image highlights multiple types of commonly occurring errors in the original Flickr30k Entities annotations. Chunking errors were a significant problem, both around proper nouns

[6]Cohen's κ is 1 when annotators are in complete agreement (Cohen, 1960)
[7]Krippendorf's α is 1 when item ratings are perfectly reliable (Krippendorff, 1970)

("a Thomas the Tank themed play") and verbs immediately following nouns ("that consists"). Though not all coreference chains are incorrect, only one non-singleton chain remains unchanged ("two men"). The changes to the "woman" chain (red) exemplifies a common issue: mentions with a subset relationship ("a woman" is a subset of "A group of actors") are not coreferent and thus must be assigned to separate chains.

In addition to these kinds of issues, Flickr30k Entities v2 tackles the oftentimes subtle distinction between visual and nonvisual mentions. In the original Flickr30k Entities annotations, nonvisual mentions were identified using a lexicon. It is often the case, however, that the distinction between visual and nonvisual is context-specific (e.g. "[the middle] of a jump" (nonvisual) versus "[the middle] of the street" (visual)). Therefore, nonvisual mentions were identified manually for each image reviewed during the v2 refinement.

At a high level, Flickr30k Entities v2 doesn't differ that much from Flickr30k Entities: v2 contains 3.6 mentions per caption (+0.4 over Entities), or 18.1 mentions per image (+2.1), clustered into 8.6 coreference chains per image (+0.9). Each chain is composed of an average of 2.2 mentions (+0.1), associated with 1.2 boxes (+0.1), where there are 8.6 boxes per image (no change from Entities). These kinds of statistics, however, obfuscate that while the number of mentions or chains haven't changed much, the quality of those mentions, chains, and box associations has changed meaningfully for around a third of the dataset. While the focus on the development and test splits means that there remains some noise and inconsistencies in the training split, Flickr30k Entities v2 can be used for nuanced natural language tasks and – most importantly – can be used to accurately evaluate such tasks.

3.2.5 Synthetic Labels for Flickr30k Entities v2

Though careful effort went into the v2 refinement effort, the task undertaken in this thesis work required two additional sets of labels: for subsets and unreviewed pronouns. While the design of the v2 refinement effort never included subset labels, training pronouns were simply not annotated due to resource constraints (recall that all images reviewed as part of the v2 refinement annotated pronouns with coreference labels). We produce these labels using heuristics, defined below.

Subset Labels Different captions sometimes partition multi-element entities differently, which makes it crucial to understand when an entity is a subset of another. In Flickr30k Entities v2, we automatically generate subset labels between pairs of mentions[8] using bounding box data and syntactic cues.

Mentions may only be in a subset relation if they are not coreferent, and – if both mentions are non-pronominal – they must be of the same lexical type. We consider an ordered pair of mentions (m_i, m_j) meeting these criteria to be in a subset relation if the associated bounding boxes $\mathbf{b_i}$ are a proper subset of $\mathbf{b_j}$ ($\mathbf{b_i} \subset \mathbf{b_j}$). Since sometimes one entity has multiple overlapping bounding boxes, we also consider overlapping boxes with an intersection-over-union score over 90% as equal for the purposes of determining subsets. That is, we also label $m_i \subset m_j$ when $|\mathbf{b_i}| < |\mathbf{b_j}|$ and $\forall b_i \in \mathbf{b_i}\ \exists b_j \in \mathbf{b_j}$ such that $\text{iou}(b_i, b_j) > 0.9$ (where $\text{iou}(b_i, b_i) = 1$). 98% of all subset relations were found by this method.

Certain syntactic structures identify subset relations more reliably than the box data. We therefore also consider m_i to be a subset of the caption's first mention m_0 (typically the main subject) when a) m_i appears in an appositive construction preceding the first verb phrase, as in "[Two people], [a man] and [a woman], walk...", or b) m_i appears as X in a partitive XofY construction where Y is coreferent with m_0, as in "[Two dogs], [one] of [which]...". 5% of all subset relations are found by this method (3% overlap with the box method).

We then enforce transitive closure of the subset relation, such that if $m_i \subset m_j$ and $m_j \subset m_k$, then $m_i \subset m_k$. This identifies a very small number of additional subset relations (fewer than 1% of the total). Finally, we set $m_j \supset m_i$ for any $m_i \subset m_j$.

Unreviewed Pronouns Given that pronouns were omitted in Flickr30k Entities and most training images were not reviewed during the v2 refinement, the majority of the training data still lacks annotations for pronouns. In order to provide supervision for our models, then, we deterministically produce coreference labels[9] between pronouns and intrasentential, non-pronominal mentions in each training image that was not reviewed during the v2 refinement.

[8]Though subset relations hold between entities, they also hold for any pair of mentions that describe those entities

[9]This procedure also associates bounding boxes with pronouns, as assigning a mention to a coreference chain also associates all the chain's boxes with the mention

To do so, we rely on rule-based anaphora resolution heuristics similar to Mitkov (1998) and Harabagiu and Maiorano (1999). These rules are inspired by traditional binding theory, where subject and object pronouns must refer to an antecedent subject outside their current clause and reflexive and reciprocal pronouns must refer to an antecedent subject within their clause (Chomsky, 1993). We approximate these rules with the following heuristics, where a pronoun m_{pro} may only have a coreference link with candidate mention m_i if m_i is non-pronominal, has matching plurality, and has matching gender[10].

- Subject / Object pronouns link to the furthest candidate.
- Reflexive / Reciprocal pronouns link to the nearest candidate.
- If a relative pronoun is X in an 'X [to be/ like] Y' construction, X links with Y (e.g. "[what] appears to be [a park]").
- Other relative pronouns link to the nearest candidate, excluding X if the relative pronoun is Y in an XofY construction (e.g. "[Two dogs], [one] of [which]").
- "both", "all", and "it" link to the nearest candidate.

Comparing the predictions made by these heuristics to the intra-caption links between mentions (where at least one mention is a pronoun) yields an accuracy of 88.81% on the development data.

3.3 MSCOCO

The MSCOCO dataset contains ∼300k images of everyday scenes, object segmentations for 80 object categories, and five captions per image (Lin et al., 2014). Though both MSCOCO and Flickr30k Entities v2 contain images depicting everyday scenes, ∼30% of MSCOCO images do not contain people or animals, instead depicting static objects (e.g. the contents of a room). In addition, MSCOCO does not annotate coreference between entity mentions, nor associates mentions with image regions.

In order to evaluate how well our methods perform on MSCOCO, we manually annotated 200 training and 200 development images[11] with the same coreference and grounding annotations as Flickr30k Entities v2, treating MSCOCO's object segmentations as bounding

[10]Subject, object, and reflexive pronouns also prefer attachments to candidates of lexical type *people* or *animals* when such attachments are available

[11]Test data for MSCOCO is not publicly available

boxes for consistency. Though the annotation process was similar to that defined in Section 3.2.3, the grounding annotations were more strict. In Flickr30k Entities v2, a person and their clothing may be ground to the same box, but since the segmentations in MSCOCO are categorized, this is not possible. Our annotated MSCOCO training images contain ∼2.6 mentions per chain, ∼6.2 chains per image and ∼8.8 segmentations per image.

CHAPTER 4: ENTITY-BASED SCENE UNDERSTANDING

In this chapter we describe our approach to entity-based scene understanding: the task of identifying entities and set relations between them. Our approach subdivides the problem into several subtasks. *Relation prediction* is the task of identifying coreference and set membership bridging relations between mentions, such that the resulting graph of mentions is consistent, with respect to the relations. *Grounding* is the task of finding the best set of image regions (including the empty set) for a given mention. For both, classifiers are used as scoring functions to make local decisions (e.g. over a pair of mentions, over a mention and bounding box) and global inference is used to ensure consistency.

The following sections define the features, classifiers, and inference procedures used for these tasks. We also define our joint inference procedure that produces mutually consistent relation and grounding graphs. Finally, we detail some of the specific implementation decisions we made, as well as our attempts to include nonvisual prediction and why this was ultimately not useful for entity-based scene understanding.

4.1 FEATURES

Feature engineering for our tasks – specifically for relation prediction – was a careful process involving significant analysis, tuning, and ablation using the Flickr30k Entities v2 development split. The features we used were initially inspired by Bengtson and Roth (2008). To these, we added various lexical features that we found to be useful in capturing phenomena closely associated with coreference, bridging, and grounding.

The sets of features are split into two groups. *Pairwise features* are used for relation classification, and are extracted from ordered pairs of mentions (m_i, m_j). *Singleton features* are used for cardinality, affinity, and nonvisual classification, and are extracted from single mentions m. Both rely on lists generated from the training data for determiners, mass nouns, collective nouns, portions, quantifiers, articles, prepositions, lexical types, and pronoun types. These lists can be found in Appendix 7. Where singleton features rely on singular lists – that is, lists of single head words, etc. – pairwise features rely on lists

Pairwise Features (One-Hot)	
Known Quantity (i/j)	Explicit text in the mention that can be mapped to 1 through 6 (all values are 0 if no such text appears in the mention
Head Pair	The ordered pair of mention head words
Lemma Pair	The ordered pair of lemmatized mention head words
Subject-Of-Verbs	The ordered pair of subject-of-verbs
Object-Of-Verbs	The ordered pair of object-of-verbs
First Word (i/j)	The first word of the mention
Numeric Modifier Pair	The ordered pair of numerical terms in the mentions (e.g. "two | one")
Modifier Pair	The ordered pair of modifiers (that is, extent text that is not included in the Numerical Pair feature)
Adjacent Preposition (left/right)	The ordered pair of prepositions immediately adjacent (left or right) to the mentions (all entries are 0 if mentions do not both have adjacent prepositions)
Distance	The distance (in number of mentions) between m_i and m_j; only applies when $c_i == c_j$ and m_i precedes m_j; distances over 10 mentions are binned together
Chunk Type Pair (left-/right)	The chunk types adjacent (to the left or right) to the mention pair (e.g. "PP | VP")
Pronoun Type (i/j)	The mention's pronoun type (see Section A.1.2), if the mention is a pronoun

Table 4.1: One-hot pairwise features used for relation prediction models

Pairwise Features (Real)	
Lexical Type Match	m_i and m_j share a lexical type (1.0 if exact match; 0.5 if lexical types overlap; 0.0 if m_i and m_j share no lexical type
Lexical Type Match (other)	m_i and m_j are of type *other* (1.0 if both are strictly *other*; 0.5 if either has multiple types; 0.0 otherwise)
Category Match	m_i and m_j have matching MSCOCO categories (1.0 if exact match; 0.5 if category overlap; 0.0 otherwise)

Table 4.2: Real-valued pairwise features used for relation prediction models

containing ordered pairs of items (e.g. "clothing | colors", "a | the").

Pairwise features are shown in Tables 4.3, 4.2, and 4.1. Note that where possible the feature descriptions are collapsed. If there exists a feature for m_i and another for m_j, there will be a single entry labeled "[name] (i/j)". Note also that the captions from which the mentions originate are denoted as c_i and c_j, respectively.

One-Hot Features Many features are expressed as one-hot vectors: that is, vectors that take value 1 when the mention or mention pair matches the item at a certain index in a list, 0 in all other positions. In cases where the list frequencies have a long tail in the training data (e.g. head words, modifiers), only items that appear more than once are used in the one-hot vector construction. In cases where the items originate from a closed vocabulary (e.g. adjacent chunk types, lexical types), only items that appear more than 1000 times in the training data are used.

N-hot Features Two features are expressed as *n-hot vectors*: Lexical Type (i/j) and Category (i/j). In these cases, when a mention has a single type or category (e.g. "other" or "pizza"), the features are a one-hot vector. When a mention has multiple types or categories (e.g. "people | other" or "broccoli | pizza"), the vector takes value 1 at each corresponding index.

Patterns Analysis of coreference and bridging cases suggested two important syntactic structures: appositives and lists. We identify these by using regular expressions over chunk type string representations of captions (e.g. "NP VP PP NP ..."). If a mention appears in a caption with one of these patterns (and is among the first mentions), we consider the mention to belong to one of these constructions.
Appositive: `^NP , (NP (VP |ADJP |PP |and )*)+,.*$`
Lists: `^NP , (NP ,?)* and NP.*$`

Governing Verbs In cases where the Stanford Dependency Parser (De Marneffe et al., 2006) returns a valid dependency tree, it is possible for mentions to be governed by a verb via a subject or object arc (e.g. nsubj, dobj). In these cases, we refer to the verb for which a mention is the subject or the object as a 'subject-of-verb' or 'object-of-verb', respectively.

Collectives Several features rely on the notion of collective nouns, particularly the distinction between collective count nouns, mass nouns, quantities, and portions. This distinction is motivated in large part by Grimshaw (2007): count nouns can be modified by quantifiers and refer to a countable number of entities (e.g. "some boxes"), while mass nouns refer to a

(typically) uncountable number of entities that may be modified by portions that may make the construction countable (e.g. "sand" is a mass, "a pile of sand" is a single unit; note that here "pile" is a portion of sand, not a quantity).

4.1.1 Singleton Features

The majority of singleton features are single-mention versions of pairwise features. This includes boolean features – Contains Article, Contains Mass, Contains Collective, Contains Portion, Is Singular Noun, Is Plural Noun, Deictic Pronoun – one hot features – Head Word, Numeric Modifier, Modifier, Subject-Of-Verb, Object-Of-Verb, Chunk Type (left / right), Pronoun Type, Pronoun, Adjacent Preposition, Known Quantity – and n-hot features –Lexical Type, MSCOCO category. The only singleton features that are not versions of pairwise features were added for nonvisual prediction. These features – Is Nonvisual (boolean) and Nonvisual Lemma (one-hot) – rely on a list of frequent nonvisual head words generated from Flickr30k Entities v2 training data.

4.1.2 Neural Features

While the brunt of the predictive power in our neural models come from the implicit feature representation of mentions produced by the LSTM, we concatenate this with explicit features for the mention or mention pair, using a subset of the features defined in Section 4.1. Specifically, this subset includes all but the very high-dimensional one-hots; for pairwise features, this means excluding Head Pair, Lemma Pair, Subject-Of-Verbs, Object-Of-Verbs, First Word, Numeric Modifier Pair, Modifier Pair, and Preposition Pair.

In this rest of this work, we refer to these features as ϕ_i or ϕ_{ij} for singleton and pairwise features encoding m_i or (m_i, m_j), respectively.

4.2 CLASSIFIERS

Both relation prediction and grounding operate on the notion that local classification is passed to global ILP inference. As a result, it is necessary for our classifiers to produce

Pairwise Features (Boolean)	
Caption Match	$c_i == c_j$
i Precedes j	m_i precedes m_j in their caption; can only be true if $c_i == c_j$
Head Match	The head word (last word) in m_i matches that of m_j
Head POS Match	The part-of-speech for the head word of m_i matches that of m_j
Lemma Match	The lemmatized version of the head word for m_i matches that of m_j
Substring Match	The lemmatized head word of m_i is a substring of the lemmatized head word of m_j (or vice versa)
Extent Match	m_i and m_j are both non-empty and match when their head words are removed
Personal Pronouns Match	m_i and m_j have matching personal pronouns (Appendix A.1.3: Personal Pronouns)
Lexical Type Match - Only	m_i and m_j have the same lexical type (Lexical Type Match == 1) and they are the only mention of that type in their originating caption
Chunk Match (left/right)	Mentions' adjacent (left or right) chunks have matching types (e.g. PP, VP)
Out Dependency Match	Both c_i and c_j have dependency parses and the dependency arcs originating in m_i and m_j have matching types (e.g. both have out dependencies of type nsubj)
Determiner Plurality Match	Both m_i and m_j start with a determiner and those determiners have matching plurality (Appendix A.1.3: Determiners)
Is Subject (i/j)	The mention is a subject (has a subject-of-verb)
Is Object (i/j)	The mention is an object (has an object-of-verb)
Is Subject Match	Both m_i and m_j are subjects
Is Object Match	Both m_i and m_j are objects
Subject Of Match	Both m_i and m_j are subjects and if their subject-of-verbs match
Object Of Match	Both m_i and m_j are objects and if their object-of-verbs match
Deictic Pronoun (i/j)	The given mention is a deictic pronoun (Appendix A.1.2: Deictic Pronouns)
XofY (i/j)	The mention is X in an XofY construction (the text between the mention and the next mention in its caption is "of")
Appositive (i/j)	The mention is in an appositive construction
In List (i/j)	The mention is in a simple list construction
Is Animate (i/j)	The mention is of lexical type *people* or *animals*
Is That (i/j)	The mention string equals "that"
i Identity j	m_i and m_j are separated by a single VP chunk which contains "to be", "is", "are", or "like" (e.g. 'm_i looks like m_j', 'm_i appears to be m_j')
i Of j	m_i and m_j appear in an XofY construction (m_i of m_j)
First in Caption (i/j)	The mention is the first in its originating caption
Adjacent	$c_i == c_j$ and m_i is immediately preceding m_j such that m_i and m_j are adjacent
Contains Article (i/j)	The mention contains an article ("a", "the", "an")
Contains Mass (i/j)	The mention contains a mass noun (See Mass Nouns – Section A.1.1)
Contains Collective (i/j)	The mention contains a collective noun (See Collective Nouns – Appendix A.1.1)
Contains Portion (i/j)	The mention contains a portion noun (See Portion Nouns – Section A.1.1)
Is Singular Noun (i/j)	The mention's head word has a POS tag of NN or NNP
Is Plural Noun (i/j)	The mention's head word has a POS tag of NNS or NNPS
Lemma Not Head Match	True *iff* Lemma Match is True but Head Match is False

Table 4.3: Boolean pairwise features used for relation prediction models

scores over possible labels. For our linear baselines, we use logistic regression or multinomial logistic regression to do so. For our neural models, we apply a softmax over the final layer to produce probabilities.

4.2.1 Neural Architecture

Our neural models share a common architecture. Captions are represented as sequences of pre-trained Word2Vec embeddings (Mikolov et al., 2013) passed to bidirectional LSTMs (Hochreiter and Schmidhuber, 1997). We concatenate the LSTM outputs of the forward and backward directions of the mentions' first and last words to encode context in the mention representation (Lee et al., 2017):

$$\mathbf{x}_i^* = [x_{i(0)}^{\text{fw}}, x_{i(0)}^{\text{bw}}, x_{i(n)}^{\text{fw}}, x_{i(n)}^{\text{bw}}]$$

where $x_{i(0)}^{\text{fw}}$ refers to the LSTM's forward direction output corresponding to the first word of mention m_i. We add explicit feature representation ϕ (see Section 4.1) to $\mathbf{x}^*$ to form the intermediate representation, which is then passed to fully connected hidden layers, and the softmax function is applied over possible labels. This architecture is shown in Figure 4.1.

4.3 RELATION PREDICTION

Relation prediction is the task of finding the best graph over mentions $m \in M$ such that each directed edge (m_i, m_j) takes one of four labels: null, coreference, subset, superset ($y \in \{n, c, b, p\}$). Edge weights are produced by multiclass classifiers ρ^{intra} and ρ^{cross} which produce a distribution over labels y. Since intra-caption and cross-caption relations tend to behave differently, we train separate classifiers for intra-caption and cross-caption examples, combining the results into classifiers ρ for notational simplicity.

In the case of our neural model for relation prediction, ordered mention pair (m_i, m_j) is represented by concatenating LSTM outputs with pairwise features: $[\mathbf{x}_i^*, \mathbf{x}_j^*, \phi_{ij}]$. It is also

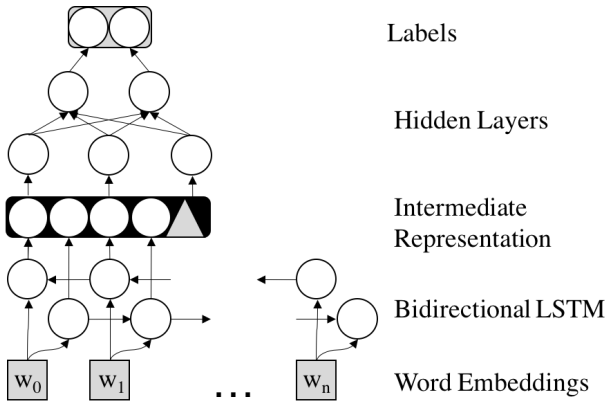


Figure 4.1: Neural architecture: sequences of Word2Vec embeddings are passed to a bidirectional LSTM; outputs are concatenated with task-specific features to form an intermediate representation, which is passed to fully connected hidden layers; softmax is applied over possible labels

important to note that ρ^{intra} is trained on pairs (m_i, m_j) where m_i precedes m_j, and predictions for (m_j, m_i) are based on those for (m_i, m_j): $\rho_{ij}^{intra}(c) = \rho_{ji}^{intra}(c)$, $\rho_{ij}^{intra}(b) = \rho_{ji}^{intra}(p)$, etc. This enables the intra-caption LSTM to capture important ordering information between mentions. There is no ordering across captions, so ρ^{cross} is trained on all cross-caption (m_i, m_j) and (m_j, m_i). Captions are passed separately to the LSTM but $[\mathbf{x}_i^*, \mathbf{x}_j^*, \phi_{ij}]$ is otherwise unchanged.

4.3.1 Relation Inference

Given the set of mentions M and scoring function ρ, the relation inference ILP maximizes Equation 4.1, where $\mathbf{r}$ is a vector of indicator variables such that $r_{ij}^y = 1$ *iff* the directed edge (m_i, m_j) takes label y.

$$\begin{aligned} &\underset{\mathbf{r}}{\text{argmax}} \sum_{i}^{M} \sum_{j \neq i}^{M} \sum_{y} r_{ij}^y \rho_{ij}(y) \\ &\text{s.t.} \quad r_{ij}^y \in \{0, 1\}; \quad \sum_{y} r_{ij}^y = 1 \end{aligned} \tag{4.1}$$

To enforce consistency, we impose the following constraints for all $m_i, m_j, m_k \in M$.

Coreference Symmetry $r_{ij}^c = r_{ji}^c$

For any pair $\{m_i, m_j\}$, both directed links must indicate coreference or neither link may indicate coreference.

Set Consistency $r^b_{ij} = r^p_{ji}$

If a pair of mentions has one directed subset relation, it must also have a directed superset relation.

Subset Transitivity $r^b_{ij} + r^b_{jk} - 1 \leq r^b_{ik}$

If $m_i \subset m_j$ and $m_j \subset m_k$, then $m_i \subset m_k$.

Relation Consistency $r^c_{ij} + r^y_{ik} - 1 \leq r^y_{jk}$

Mentions of the same entity must all have the same relations to other mentions. This also enforces transitive closure for coreference (all mentions in a coreference chain must corefer with all others).

4.4 GROUNDING

We frame grounding as the task of finding the best set of image regions (bounding boxes) for each entity mention, which we divide into the subtasks of a) determining whether a mention describes a box (affinity), and b) determining to how many boxes a mention should ground (cardinality). These predictions are then resolved using ILP inference.

4.4.1 Box / Mention Affinity

The central challenge of the grounding task is determining whether mention m_i describes image region (bounding box) b_o in isolation of other mentions and boxes. We train affinity classifier γ to predict a $[0, 1]$ confidence score which can then be combined with cardinality during inference. Our approach to affinity represents boxes using the fc7 layer of the Fast RCNN network (Girshick, 2015), and the representation of mentions is different for our baseline and neural models.

Baseline Affinity Our baseline affinity model projects image and text features into a shared semantic space with normalized canonical correlation analysis (CCA) (Gong et al., 2014). Here, mentions are represented using Fisher vectors derived from hybrid Gaussian-Laplacian mixture models (Klein et al., 2015) which are built on top of Word2Vec (Mikolov et al., 2013). Given this shared semantic space, we are able to compute the cosine distance between mentions and boxes. Like Plummer et al. (2015), we randomly subsample a maximum of 10 gold bounding boxes for each unique mention string during training. Where their method merges image regions when a mention refers to multiple boxes, however, we associate mentions with individual boxes.

Informed by traditional object detection, we use the CCA scores on development box/-mention pairs to train a logistic regression classifier for each lexical type. In this way, we can convert the CCA scores for the test data into $[0, 1]$ affinity predictions.

Neural Affinity Our neural affinity model uses the architecture detailed in Section 4.2.1, where the intermediate representation for each (m_i, b_o) pair is represented as $[\mathbf{x}_i^*, \phi_i, \text{RCNN}(b_o)]$, where $\text{RCNN}(b_o)$ is the box representation.

4.4.2 Mention Cardinality

We approach grounding as a matching problem, in which we not only must identify whether a mention describes a box, but we must determine the *set* of boxes that a mention describes. This requires a notion of the number of entities a mention describes, or its cardinality. We predict cardinality using multiclass classifier δ, which predicts a distribution over possible labels $n \in \{0, 1, ...10, 11+\}$ [1] such that $\sum_n \delta(n) = 1$.

Our baseline cardinality model is a simple logistic regression model. Our neural model leverages the architecture detailed in Section 4.2.1 where each mention m_i is encoded as $[\mathbf{x}_i^*, \phi_i]$.

[1] Cardinalities over 10 are binned together

4.4.3 Grounding Inference

Given the set of mentions M, the set of bounding boxes B, and scoring functions γ and δ, the grounding inference ILP maximizes Equation 4.2, where γ_{io} is the confidence of affinity between m_i and b_o, $\delta_i(n)$ is the confidence that m_i is associated with n boxes, and $\mathbf{g}$ and $\mathbf{z}$ are vectors of indicator variables where $g_{io} = 1$ when m_i is ground to b_o and $z_i^n = 1$ when m_i is ground to n boxes

$$\begin{aligned} &\underset{\mathbf{g}}{\text{argmax}} \sum_{i}^{M} \sum_{n=0}^{|B|} z_i^n \delta_i(n) + \frac{1}{|B|} \sum_{o}^{B} \left[g_{io}\gamma_{io} + g'(1-\gamma_{io}) \right] \\ &\text{s.t.} \quad g_{io}, g'_{io}, z_i^n \in \{0,1\}; \quad g_{io} + g'_{io} = 1 \end{aligned} \tag{4.2}$$

Equation 4.2 finds the best set of boxes to which each mention should be ground, weighting γ and δ equally. To enforce consistency, we impose the following constraints $m_i \in M$ and $b_o \in B$:

Mention Cardinality

$$\begin{aligned} &0 \leq \beta a_i^n + \sum_o g_{io} - n \leq \beta - 1 \\ &0 \leq \beta b_i^n - \sum_o g_{io} + n \leq \beta - 1 \\ &0 \leq 2 - a_i^n - b_i^n - 2z_i^n \leq 1 \\ &\text{s.t.} \quad a_i^n, b_i^n \in \{0,1\}; \quad \beta = 2|B| + 1 \end{aligned} \tag{4.3}$$

In order to add the cardinality score[2] $\delta_i(n)$ to the objective *iff* m_i is ground to n regions, we define variables z_i^n which take value 1 when $\sum_o g_{io} = n$: the first constraint defines $a_i^n = 1$ for $\sum_o g_{io} < n$, the second defines $b_i^n = 1$ for $\sum_o g_{io} > n$, and the third requires that $z_i^n = 1$ only when $a_i^n = b_i^n = 0$.

Box Minimum $\sum_i g_{io} \geq 1$

In datasets where gold boxes must be described to be present in the data, we enforce that each box must be ground to some mention.

[2]We split the confidence of $\delta_i(11+)$ equally among all n in $10 < n \leq |B|$

4.5 JOINT INFERENCE

Relation prediction and grounding are interrelated: if two mentions corefer, they must ground to the same boxes; if a set relation holds between mentions, it must also hold between their boxes. Given M, B, ρ, γ, and δ, our joint inference procedure maximizes Equation 4.4, where $\mathbf{r}$, $\mathbf{g}$, $\mathbf{z}$ are as defined in Sections 4.3 and 4.4):

$$\underset{\mathbf{r},\mathbf{g}}{\text{argmax}} \sum_{i}^{M} \left[\frac{1}{|M|} \sum_{j \neq i}^{M} \sum_{y} r_{ij}^{y} \rho_{ij}(y) + \frac{1}{2} \left[\sum_{n=0}^{|B|} z_i^n \delta_i(n) + \frac{1}{|B|} \sum_{o}^{B} \left(g_{io}\gamma_{io} + g'_{io}(1-\gamma_{io}) \right) \right] \right] \tag{4.4}$$

This finds the best mutually consistent relations and groundings for an image, weighing relation prediction and grounding equally. To enforce this mutual consistency, we impose the constraints defined in Sections 4.3 and 4.4 in addition to the following.

Grounded Coreference $\quad r_{ij}^{c} + g_{io} - 1 \leq g_{jo}; \quad r_{ij}^{c} + g_{jo} - 1 \leq g_{io}$

Coreferent mentions must ground to the same boxes.

Grounded Subsets $\quad r_{ij}^{b} + g_{io} - 1 \leq g_{jo}$

If one mention refers to a subset of another, the superset must be ground to all of the subset's boxes.

$$2 \leq z_i^0 + z_j^0 + 2u_{ij} \leq 3$$

$$0 \leq \sum_{o} g_{io} - \sum_{o} g_{jo} + \beta w_{ij} \leq \beta - 1$$

$$r_{ij}^{b} + u_{ij} - 1 \leq w_{ij}; \quad r_{ij}^{b} + w_{ij} - 1 \leq u_{ij}$$

$$\text{s.t.} \quad u_{ij}, w_{ij} \in \{0,1\}; \quad \beta = 2|B| + 1$$

If either the subset or superset are ground to boxes, the superset must be ground to more boxes than the subset: u_{ij} stores whether m_i or m_j are ground to any boxes, w_{ij} stores whether m_i is ground to fewer boxes than m_j. We also require that in order for $m_i \subset m_j$, m_i must be ground to fewer boxes than m_j if m_i or m_j are ground to any boxes.

4.5.1 Sequential Inference

Since our primary goal is the production of mutually consistent relation and grounding graphs, simultaneous relation and grounding inference is not necessary so long as the graphs adhere to joint constraints. Therefore, we also introduce two sequential variants: *Relation then Grounding* and *Grounding then Relation*. Here, an individual inference scheme is used to make predictions for one graph and then joint inference is performed over these fixed predictions and the other graph. In this way, grounding decisions must adhere to relations or relation decisions must adhere to grounding links, respectively.

4.6 IMPLEMENTATION DETAILS

Our quantitative results are based on models trained on the Flickr30k Entities v2 training and development data (25381 plus 3000 images) and evaluated on the test portion of that dataset (3000 images). The example shown in Figure 5.1 is a development image with predictions based on models trained only with the training data. Baseline classifiers are implemented in Scikit-Learn (Pedregosa et al., 2011), neural models in Tensorflow (Abadi et al., 2015), and ILP problems are solved with Gurobi (Gurobi Optimization, 2015). Parameters were tuned on the development data; we use batch sizes of 512, LSTM hidden sizes of 200, 50% dropout on all nodes, and two fully-connected hidden layers (of size 512 and 256) after the LSTM.

4.7 IDENTIFYING VISUAL MENTIONS

In order to completely understand a scene through its visual entities, one possible first step is to determine which mentions refer to something that is pictured in the scene; in essence, identifying visual mentions. While this intuition is theoretically useful, it turned out to be practically unimportant. In our experiments, visual classification was a difficult task and the benefits it provided – excluding nonvisual mentions from taking relations or grounding – did not meaningfully outweigh its detriments – excluding visual mentions or including nonvisual ones.

In this section, we detail the our approach for identifying visual mentions as it would fit in our system for entity-based scene understanding. While this did not prove to be useful for the Flickr30k Entities v2 or MSCOCO data – the results in Chapters 5 and 6 assume all mentions are visual – it is possible that other datasets and similar tasks may require a better understanding of the visual / nonvisual distinction.

4.7.1 Classification

Identifying visual mentions should occur in parallel with relation prediction and grounding, as only visual mentions may take relations and groundings. Therefore, our approach for identifying visual mentions follows the same classification-then-inference scheme.

Most similar to the affinity classifier, we would train visual classifier η to predict a $[0, 1]$ confidence that a mention is visual. Since this task operates over single mentions, the mention representation would be the same as defined in Section 4.1.1 for the logistic regression baseline, or would be the same as defined in Section 4.1.2 for the neural model: $[\mathbf{x}_\mathbf{i}^*, \phi_\mathbf{i}]$.

4.7.2 Visual Inference

As a task in isolation, visual inference simply assumes that each mention m_i with $\eta_i \geq 0.5$ is visual. This is equivalent to the following ILP formulation, where $\mathbf{v}$ is a vector of indicator variables such that $v_i = 1$ *iff* mention m_i is visual.

$$\begin{aligned} &\underset{\mathbf{v}}{\text{argmax}} \sum_{i}^{M} v_i \eta_i + v_i'(1 - \eta_i) \\ &\text{s.t.} \quad v_i, v_i', \in \{0, 1\}; \quad v_i + v_i' = 1 \end{aligned} \tag{4.5}$$

Identifying visual mentions, relation prediction, and grounding are interrelated tasks. In our labeling scheme, only visual mentions may take relations, and only visual mentions may be ground to boxes. In a fully joint inference scheme, we would leverage this interrelatedness by inferring visual mentions, relations, and groundings jointly.

We frame this joint inference as an augmentation of Equation 4.4, where we seek not only to find the best relation and grounding graphs, but to find the best set of visual mentions

as well. Thus, given Given M, B, η, ρ, γ, and δ, this ILP maximizes Equation 4.6, where $\mathbf{v}$, $\mathbf{r}$, $\mathbf{g}$, and $\mathbf{z}$ are vectors of indicator variables as defined in Sections 4.3.1 and 4.4.3.

$$\begin{aligned}\underset{\mathbf{v},\mathbf{r},\mathbf{g}}{\operatorname{argmax}} \sum_{i}^{M} \Bigg[v_i \eta_i + v'_i(1-\eta_i) + \frac{2}{|M|} \sum_{j \neq i}^{M} \sum_{y} r_{ij}^{y} \rho_{ij}(y) + \\ \frac{1}{2} \Bigg[\sum_{n=0}^{|B|} z_i^n \delta_i(n) + \frac{1}{|B|} \sum_{o}^{B} \left(g_{io}\gamma_{io} + g'_{io}(1-\gamma_{io}) \right) \Bigg] \Bigg]\end{aligned} \tag{4.6}$$

For each mention, Equation 4.6 determines if its visual and, if so, finds the best relations and groundings such that the contributions of the visual mention identification, relation prediction, and grounding prediction are weighed equally. This ILP incorporates the constraints in Sections 4.3.1, 4.4.3, 4.5, and the following, which incorporates visual prediction with joint relation and grounding inference.

Visual Relation Constraint $\quad v_i + v_j \geq 2r_{ij}^{y'} \quad \forall y' \in \{c, b, p\}$

Only visual mentions can hold coreference or set relations. In order for mention m_i to hold such a relation with m_j, both m_i and m_j must be visual.

Visual Entity Constraint $\quad v'_i \leq z_i^0$

Only visual mentions can be ground to boxes. If a mention is nonvisual, it must be ground to 0 boxes.

4.7.3 Visual Mention Identification Discussion

While visual mention identification is conceptually important, it did not have practical significance in our experiments. The classifiers (both baseline and neural) provided no meaningful improvement over simply predicting 'visual' for all mentions, a result that was confirmed during joint visual/relation/grounding inference, where the ILP found it easier to assign each mention as visual and thus effectively ignore the *Visual Relation* and *Visual Entity* constraints.

There may be many reasons for this phenomena, but the simplest is that our visual classifier η was simply too weak to be of much use (on its own or as part of inference), which

ultimately ties back to the data. Though the distinction between visual and nonvisual mentions was present in Flickr30k Entities, significant nuance was introduced in the v2 refinement regarding what constitutes a nonvisual mention. Thus, as is the case with much of the v2 refinement, the new annotations are cleaner and more thorough, but the overall state of the training split is somewhat inconsistent.

That said, however, the distinction between visual and nonvisual mentions is important in the everyday domain, particularly as it relates to the entity-based scene understanding task, and future work in this direction may find this inference formulation to be useful.

CHAPTER 5: RESULTS ON FLICKR30K ENTITIES V2

We evaluate each component of our system for entity-based scene understanding separately, comparing baseline classifiers, neural models, task-specific inference, and joint inference. We do not compare our methods to off-the-shelf tools for related tasks, since our tasks' uniqueness makes these comparisons inappropriate[1].

5.1 RELATION PREDICTION

To our knowledge, there is no metric that simultaneously evaluates coreference clustering with asymmetric set relationships between clusters. We therefore evaluate relation prediction in three ways: by mention-pair, overall, and as coreference.

Mention-Pair Evaluation Table 5.1 shows precision, recall, and F1 for each relation type, counting a prediction for the unordered pair $\{m_i, m_j\}$ as correct only when the ordered (m_i, m_j) and (m_j, m_i) edge labels both match the gold. For symmetric relations (coreference, null) both edge labels must agree, and for the asymmetric sub/superset relations, both directions must be labeled correctly.

When measuring by mention pairs, relation prediction conforms to our expectations, with neural classifiers performing better than our baseline, and inference performing better still.

Overall Evaluation We also report each model's performance based on the number of images for which every link is predicted correctly (correct images) and the number of chains that include a) exactly the same mentions as the gold, and b) for which each relation to/from the chain's mentions are the same as the gold (correct chains). As shown in Table 5.2, relation inference performs best, producing a larger number of correct chains than even joint inference, suggesting that even by this metric grounding hurts relation performance.

[1]The statistical Stanford coreference system (Manning et al., 2014) has a B^3 F1 = 28.33% on Flickr30k Entities v2 test data

Relation Prediction – by mention pair (Flickr30k Entities v2 test)			
Relation	**P**	**R**	**F1**
Baseline			
null	95.57	98.23	96.88
coref.	88.59	78.77	83.39
subset	71.30	43.09	53.72
Neural classifiers $\rho = \rho^{intra}$ *and* ρ^{cross} (all pairs)			
null	**96.31**	98.09	97.19
coref.	90.69	79.82	84.91
subset	74.92	50.25	60.16
Relation Inference			
null	95.96	98.61	**97.27**
coref.	91.60	**80.15**	**85.49**
subset	75.28	**52.56**	**61.90**
Joint Inference			
null	95.28	98.90	97.05
coref.	93.22	76.20	83.86
subset	77.24	50.29	60.92
Grounding then Relation Inference			
null	92.56	**99.24**	95.78
coref.	**94.52**	58.11	71.98
subset	**78.13**	44.84	56.98

Table 5.1: Relation prediction results by mention pair for Flickr30k Entities v2 test data (null, coreference, and subset link pairs comprise 84.40%, 13.39%, and 2.21% of the link pairs between mentions)

Coreference Evaluation We also evaluate relation prediction as a standard coreference resolution task, using the B^3 metric (Bagga and Baldwin, 1998). Here, null and set labels are treated as not coreferent (Table 5.3). These results reinforce the conclusions given by the other evaluations: relation inference may produce the best results, but joint inference doesn't perform significantly worse.

Sequential Inference Like joint inference, the sequential inference schemes *Relation then Grounding* and *Grounding then Relation* produce mutually consistent graphs. The former results are identical to those of relation inference (since relation inference is run prior to grounding) and the latter results are shown in Tables 5.1, 5.2, and 5.3. All metrics indicate that requiring relations to conform to decisions made by grounding hurts performance significantly.

Relation Prediction overall (Flickr30k Entities v2 test)		% Correct	
	Acc.	**Chains**	**Images**
Baseline	94.41	60.84	11.87
Neural ρ	94.58	59.84	11.70
Relation Inference	**95.12**	**65.99**	**18.40**
Joint Inference	94.25	65.40	14.87
Grnd. then Rel. Inf.	92.53	58.70	5.90

Table 5.2: Relation prediction performance by link accuracy, correct images and coreference chains for Flickr30k Entities v2 test data

Coreference Resolution – B^3 (Flickr30k Entities v2 test)			
	P	**R**	**F1**
Baseline	85.51	90.55	87.24
Neural ρ	85.14	**91.81**	87.71
Relation Inference	89.78	88.62	**88.69**
Joint Inference	90.87	85.77	87.74
Grnd. then Rel. Inf.	**91.84**	74.90	81.62

Table 5.3: Coreference resolution performance for Flickr30k Entities v2 test data

5.2 GROUNDING

Though grounding is an established task, the common framing is to find the best region for a phrase from a set of proposals, rather than finding the best *set* of regions for a phrase as we do. This renders standard metrics (e.g. Recall@K) inappropriate for our purposes. We therefore evaluate with two schemes (affinity and overall) in Table 5.4.

Affinity Evaluation We measure precision, recall, and F1 of affinity links (predictions associating mention m with box b) along with overall link accuracy. In general, grounding performance behaves as expected: the neural classifier performs better than the CCA baseline. Cardinality (grounding inference) and relations (joint inference) help further.

Overall Evaluation We also report the number of correct images (every grounding link is correct) and the number of correct mentions (all of m's grounding links are correct). Joint inference outperforms grounding inference, which significantly outperforms the classifiers. In particular, the number of correct images almost doubles with the incorporation of relations during joint inference.

Grounding (Flickr30k Entities v2 test)					
by (mention, box) pair			by link	%Correct	
P	**R**	**F1**	**Acc.**	**Mnts**	**Imgs**
CCA Baseline					
71.01	39.04	50.38	88.08	39.37	0.60
Neural Affinity Classifier γ					
73.91	53.95	62.37	89.91	46.34	1.57
Grounding Inference					
70.09	59.86	64.57	89.81	60.36	4.77
Joint Inference					
72.07	59.62	65.26	90.15	62.73	9.17
Relation then Grounding Inference					
71.16	**61.74**	**66.12**	**90.19**	**63.01**	**10.10**

Table 5.4: Grounding performance on Flickr30k Entities v2 test data; 15.51% of the gold links between mentions and boxes are positive

Sequential Inference Since *Grounding then Relation* runs grounding inference first, the grounding results are identical to grounding inference alone. The best grounding performance is from *Relation then Grounding*: first performing relation prediction and requiring grounding to conform to those decisions produces the best results for both. Given the relative strengths of these systems, this makes sense. Relation inference can accurately identify entities, and provides high quality signal to grounding.

Gold

Two women$_1$[b-2, b-5] in **shorts**$_2$[b-8, b-9] and **protective equipment**$_3$[b-1, b-3] bumping into **each other**$_1$[b-2, b-5].
Two women$_1$[b-2, b-5] wearing **helmets**$_4$[b-0, b-10] and **safety pads**$_5$[b-4, b-6] appear to be fighting.
Two players$_1$[b-2, b-5] collide during **a recent roller derby match**$_6$.
A woman$_7$[b-2] in **a green shirt**$_8$[b-7] pushes past in **roller derby**$_6$.
Two young women$_1$[b-2, b-5] tackling **each other**$_1$[b-2, b-5] while skating.

chain 7 ⊂ **chain 1**

Predicted

Two women$_1$[b-5, b-6] in **shorts**$_2$[b-7, b-8] and **protective equipment**$_3$[b-1, b-9] bumping into **each other**$_1$[b-5, b-6].
Two women$_1$[b-5, b-6] wearing **helmets**$_4$[b-3, b-10] and **safety pads**$_5$[b-0, b-2] appear to be fighting.
Two players$_1$[b-5, b-6] collide during **a recent roller derby match**$_6$.
A woman$_7$[b-6] in **a green shirt**$_8$[b-4] pushes past in **roller derby**$_6$.
Two young women$_1$[b-5, b-6] tackling **each other**$_1$[b-5, b-6] while skating.

chain 7 ⊂ **chain 1**

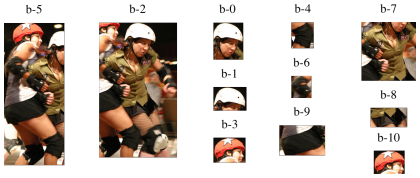


Figure 5.1: Gold annotations and predictions for a Flickr30k Entities v2 dev image; coreference chains are shown with subscripts and color coding, groundings with superscripts, referenced boxes with identifiers

5.3 DISCUSSION

Figure 5.1 shows an example image with predictions made by *Relation then Grounding* inference. The predicted relations are perfect: "Two women", "Two players", "each other" are coreferent, "a woman" is a subset of the "Two women", etc. Grounding performs more poorly: "Two women" is associated to the box for one of the women (b-2) and a small box for an article of clothing (b-6). Some grounding mistakes are more subtle: "helmets" is correctly ground to boxes of white and orange helmets, but the system chooses the wrong box for the white helmet.

In general, our approach to relation prediction works very well. Relation inference has the best results, and, although grounding decisions don't help relations, joint inference only slightly hinders relation performance while benefitting grounding performance (particularly the number of correct images). In fact, each of our steps improves grounding performance,

which is likely due to the inherent difficulty in our framing of affinity. Assigning a probability that a mention describes a box that is interpretable across other boxes and mentions is difficult on its own, and expecting our classifier to assign high confidence to multiple boxes in cases where a mention refers to a set of entities (e.g. "Two women") complicates the task significantly. Despite this difficulty, our method for producing mutually consistent relation and grounding graphs yields good results for both, showing that even a weak grounding system is improved by relations.

CHAPTER 6: TOWARDS MSCOCO ENTITIES

Though we view the identification of entities as a first step toward understanding a scene from multiple descriptions, it may appear as though we've addressed a synthetic problem of our own invention. Our methods assume the presence of multiple descriptions for a single scene, and – as stated in Chapter 3 – we require expensive, high-quality, labeled data on which to train our supervised models. Our approach, however, extends beyond Flickr30k Entities v2, and can be used as a mechanism to automatically generate these rich annotations for similar image caption datasets. In this section, we detail the process for generating such annotations on the MSCOCO dataset (Lin et al., 2014).

Like Flickr30k Entities v2, MSCOCO contains images of everyday scenes, each described by five captions. Unlike our data, however, MSCOCO neither provides coreference labels nor box associations. For a more complete description of MSCOCO, see Section 3.3.

6.1 A PRELIMINARY MSCOCO ENTITIES

We evaluate the performance of our methods on the 200 MSCOCO development images that we annotated (see Section 3.3), where each model is trained using Flickr30k Entities v2 training and development data along with the 200 annotated MSCOCO training images. Since the results in Chapter 5 indicate that requiring grounding to conform to relations yields the best performance[1], we report the results of relation prediction inference in Table 6.1 and the results of grounding on its own and as part of sequential inference in Table 6.2.

6.1.1 Relation Prediction for MSCOCO

In order to perform relation prediction for MSCOCO, we first identify mentions using the same preprocessing steps as for Flickr30k on the raw MSCOCO captions. These mentions are then used in the pipeline detailed in Chapter 4. Relation inference performs about as well on MSCOCO as it does on Flickr30k Entities v2, suggesting that while their domains

[1]Additional experiments on MSCOCO confirm that *Relation then Grounding* performs better than joint inference

Relation Prediction (MSCOCO dev)

Coreference – B^3			% Correct		
P	**R**	**F1**	**Acc.**	**Chains**	**Images**
90.36	86.20	87.53	94.37	69.42	20.50

Relation	**P**	**R**	**F1**
null	94.30	99.19	96.68
coref.	94.72	77.49	85.24
subset	95.59	43.82	60.09

Table 6.1: Relation Prediction performance on MSCOCO dev data; null, coreference, and subset link pairs comprise 80.68%, 17.45%, and 1.86% of the link pairs between mentions, respectively

differ, the language used in both datasets is similar enough that relation models primarily trained on Flickr30k Entities v2 are useful for MSCOCO.

6.1.2 Grounding for MSCOCO

Grounding in MSCOCO is challenging. In the 200 training images we annotated, 45.7% of mentions are not ground to any region[2] and 41.8% of regions are not described by any mention (which does not happen at all in Flickr30k Entities). Since only objects belonging to the 80 categories are segmented, we created a lexicon combining frequent MSCOCO head words and the Flickr30k Entities lexicon to identify the categories to which mentions may belong. According to our lexicon, 57.2% of the training mentions have a category; of these, 85.2% are ground to an image region, accounting for 95% of all grounded mentions. Thus, category information, leveraged by a lexicon with good coverage, can provide meaningful grounding signal.

We leverage this signal in two ways. First, we introduce a heuristic baseline, which assigns confidence $1-\epsilon$ when a mention and a box share a category, and ϵ otherwise[3] We then modify the grounding inference procedure in Section 4.4.3 by removing the *Box Minimum* constraint (which is only useful for datasets where each box is described by a mention) and by adding the constraint that mentions may only ground to boxes of the same category (making the

[2]In Flickr30k Entities v2, 11.5% of mentions are not ground to any box
[3]ϵ ($:= 2^{-1074}$) is used to prevent 0 label confidence

Grounding (MSCOCO dev)					
by (mention, box) pair			by link	%Correct	
P	**R**	**F1**	**Acc.**	**Mnts**	**Imgs**
CCA Baseline					
76.13	40.06	52.50	89.34	61.01	5.50
Heuristic Baseline					
72.47	**86.10**	78.70	93.15	78.63	23.00
Neural Affinity Classifier γ					
75.30	54.71	63.37	90.70	63.59	7.00
Grounding Inference with γ					
91.53	46.98	62.09	91.56	78.92	16.00
Grounding Inference with Heuristic					
87.14	73.25	79.59	94.48	79.08	18.50
Grounding Inference with Heuristic γ *Average*					
91.99	62.25	74.25	93.65	80.63	17.50
Relation then Grounding Inference with Heuristic γ *Average*					
91.59	44.70	60.08	91.26	74.94	19.50
Relation then Grounding Inference with Heuristic* γ *Average*					
88.43	75.26	**81.32**	**94.91**	**82.29**	27.50

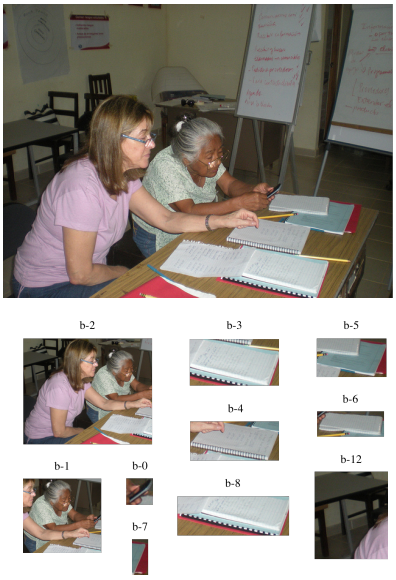
Table 6.2: Grounding performance on MSCOCO dev data; in the bipartite graph between mentions and boxes, 14.71% of the links indicate positive affinity

heuristic baseline recall the upper bound).

Category information is extremely useful for grounding: the heuristic significantly outperforms both the CCA baseline and the affinity classifier. We show the performance of grounding inference with three scoring functions: the classifier γ, the heuristic, and the average of the two.[4]. The heuristic has much better recall, but the average has the highest precision, and the highest number of correct mentions. Since our goal is the production of high-precision annotations, we use this average when combining grounding and relations.

Relation then Grounding inference as defined in Section 4.5.1 produces mutually consistent graphs, but the performance is poor for MSCOCO. Regardless of which affinity scoring function is used, recall significantly decreases. This suggests that it is easier for inference to omit grounding links than adjust assignments when trying to conform to relations. Since we know that grounding inference alone has very high precision, however, we introduce an additional sequential inference scheme, *Relation then Grounding**, in which we run relation

[4]The average of the classifier and heuristic is $\frac{1}{2}(1 - \epsilon + \gamma_{io})$ in cases where m_i and b_o share a category, $\frac{1}{2}(\epsilon + \gamma_{io})$ otherwise



Gold

A woman$_1$[b-2] sitting at a desk$_2$ with an older woman$_3$[b-1].
Two women$_4$[b-1, b-2] are woking on an assignment— together.
Two women$_4$[b-1, b-2] are sitting at a wood desk$_2$.
A woman$_1$[b-2] and an elderly woman$_3$[b-1] sitting at a desk$_2$ together in a classroom$_5$ setting with notebooks$_6$[b-3, b-4, b-5, b-6, b-7, b-8, b-9] and pencils$_7$ on the desk$_2$.
Two women$_4$[b-1, b-2] with glasses$_8$ on sitting at a table$_2$ with notebooks$_6$[b-3, b-4, b-5, b-6, b-7, b-8, b-9] , pencils$_7$ and a cellphone$_9$[b-0].

chain 1 ⊂ chain 4; chain 3 ⊂ chain 4

Predicted

A woman$_1$[b-1, b-2] sitting at a desk$_2$[b-12] with an older woman$_3$[b-1, b-2].
Two women$_4$[b-1, b-2] are woking on an assignment$_5$ together.
Two women$_4$[b-1, b-2] are sitting at a wood desk$_2$[b-12].
A woman$_3$[b-1, b-2] and an elderly woman$_1$[b-1, b-2] sitting at a desk$_2$[b-12] together in a classroom$_6$ setting with notebooks$_7$ and pencils$_8$ on the desk$_2$[b-12].
Two women$_4$[b-1, b-2] with glasses$_9$ on sitting at a table$_2$[b-12] with notebooks$_7$, pencils$_8$ and a cellphone$_{10}$[b-0].

chain 1 ⊂ chain 4; chain 3 ⊂ chain 4

Figure 6.1: Predicted annotations for MSCOCO dev image compared against human annotations; coreference chains are shown with subscripts and color coding; groundings shown with supercripts; referenced boxes are shown individually, where boxes b-9 to b-11 are not ground to any mention in the gold or predicted

inference and grounding inference separately before propagating grounding links along predicted relations. In this scheme, m_i is ground to every box that m_j is ground to if m_i is coreference or a superset[5] of m_j. This provides our best results, showing that relations provide meaningful signal to identify entities.

6.2 DISCUSSION

An example image with predictions made by *Relation then Grounding** inference is shown in Figure 6.1. Our predicted relations are very good; we correctly predict that "an older woman" and "A woman" are both subsets of "Two women", and we correctly associate each instance of the desk with the table. We make two mistakes: we assign a coreference label to a non-visual mention[6], and we mistakenly assign coreference between "A woman"

[5] *Relation then Grounding** inference therefore allows subsets with the same number of boxes
[6] Flickr30k Entities and the 400 annotated MSCOCO images contain some nonvisual mentions (entities that cannot be pictured in the scene), but our methods ignore this distinction, and assume that all mentions can be ground to boxes

and “an elderly woman” (rather than “an older woman”). While it isn’t clear why our model doesn’t capture the high similarity between “elderly woman” and “older woman”, this mistake is understandable given the lack of document-level information about the total number of women in the scene.

While the grounding predictions are good in general, they highlight the weakness of our heuristic. For example, “Two women” is ground correctly, but given our relaxation of proper subsets, each individual woman is also ground to both boxes of category *person*. In the case of the “desk” chain, our method is limited by the gold data; b-12 is the wrong table, but is the only region of category *dining table* (which often includes desks) in the image. Our method is also limited by our lexicon; we do not ground any of the notebooks because our lexicon does not include “notebook” as an entry for the *book* category, to which regions b-3 to b-8 belong.

MSCOCO is a much noisier dataset than Flickr30k Entities v2; many mentions describe un-annotated image regions, and many image regions aren’t described by any mention. Despite this, our methods produce good results on MSCOCO, confirming that relations provide meaningful signal to grounding, even with a strong grounding system.

CHAPTER 7: CONCLUSION

In this thesis, we've introduced the entity-based scene understanding task, which combined coreference, set membership bridging, and grounding. In support of this and similar tasks, we've also discussed the creation of the Flickr30k Entities and Flickr30k Entities v2 datasets. Our primary contribution is the approach to this new task. We show that our individual approaches to the subtasks of relation prediction and grounding produce strong results, and that when these approaches are combined, particularly in sequential inference schemes where groundings must adhere to relations, grounding performance can be significantly improved.

Our approach to entity-based scene understanding may not only be useful in other cases where multiple descriptions refer to the same scene – as in the case of multiple premise entailment (Lai et al., 2017) – but is also useful on its own as a mechanism for automatically generating rich, high-quality annotations for similar image caption datasets like MSCOCO.

The most direct extensions of this work are to improve the individual relation prediction and grounding components. Both operate over gold sets of mentions or mentions and boxes, respectively, but modern approaches often incorporate the prediction of these sets into their task. Thus, a more sophisticated approach to relation prediction would also incorporate mention detection as is common in the coreference resolution literature, and a more sophisticated grounding approach would find the salient image region as it common in the phrase localization literature. Both of these would complicate the entity-based scene understanding task significantly, but would enable a system to work with less structured data (e.g. images and sentences without any object-level or mention-level annotations).

One important conceptual omission that we have made in the definition of this task is that of meronymy. In defining the entities in a scene through their entity mentions – which for our purposes we can equate with definite descriptions; meronymy, or the relation held when an entity is a constituent or part of another, can be seen as the missing piece of relation prediction as we've defined it. Where mention detection and phrase localization would extend our approach at a technical level, the inclusion of meronymy as another pair of directed relations would extend our approach at a conceptual level to include all definite description relations.

These direct extensions are merely the next steps that can be taken within the existing task and approach presented here. Entity-based scene understanding, however, is merely a first step to the much broader goal of understanding the scene in its entirety: defining entities, their relations, their attributes, and the activities in which they are engaged.

This thesis is a foundation toward understanding the everyday world through its entities. Our approach identifies the *what* and *who* in a scene. Building on this foundation, future work may incorporate more fine grained attributes of the entities and the event (the *how*) and temporal and causal aspects to understanding the scene (the *when* and *why*). With a general understanding of the everyday world – a notion of what happened in an everyday scene – such systems may explore more nuanced and complex phenomena, like what the event means, or what is implied by what is and is not described.

A.1 LISTS

A.1.1 COLLECTIVE NOUNS

Collective Nouns
amount, arrangement, array, assortment, band, bunch, bundle, collection, community, congregation, contemporaries, council, crew, crowd, ensemble, family, flight, flock, forest, gang, group, herd, litter, load, lot, mob, number, pack, parade, personnel, series, set, squad, stack, team, throng, troop, troupe, vegetation

Mass Nouns
sand, snow, tea, water, beer, coffee, dirt, corn, liquid, wine

Portion Nouns
pile, sheet, puddle, mound, spray, loaf, cloud, drink, sea, handful, bale, line, row

A.1.2 PRONOUNS

Subject Singular	he, she, it
Subject Plural	they
Object Singular	him, her, it
Object Plural	them
Reflexive Singular	himself, herself, itself, oneself
Reflexive Plural	themselves
Reciprocal	each other, one another, each
Relative	that, which, who, whose, whom, where, when, what
Demonstrative	this, that, these, those, there
Indefinite	anything, anybody, anyone, something, somebody, someone, nothing, nobody, noone, no one
Deictic	another, other, others, one, two, three, four, some
Other	both, all

A.1.3 MISCELLANEOUS

Personal Pronouns
his, hers, its, their

Singular Determiners
a, the, an, another, this, no

Plural Determiners
some, each, all, both, these

REFERENCES

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

A. Bagga and B. Baldwin. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Conference workshop of the International Conference on Language Resources and Evaluation (LREC)*, volume 1, pages 563–566, 1998.

E. Bengtson and D. Roth. Understanding the value of features for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 294–303, 2008.

K.-W. Chang, R. Samdani, A. Rozovskaya, N. Rizzolo, M. Sammons, and D. Roth. Inference protocols for coreference resolution. In *Proceedings of the Conference on Computational Natural Language Learning (CONLL)*, pages 40–44, 2011.

N. Chomsky. *Lectures on government and binding: The Pisa lectures*. 1993.

K. Clark and C. D. Manning. Entity-centric coreference resolution with model stacking. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 1405–1415, 2015.

K. Clark and C. D. Manning. Deep reinforcement learning for mention-ranking coreference models. 2016a.

K. Clark and C. D. Manning. Improving coreference resolution by learning entity-level distributed representations. 2016b.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

M.-C. De Marneffe and C. D. Manning. Stanford typed dependencies manual. Technical report, Technical report, Stanford University, 2008.

M.-C. De Marneffe, B. MacCartney, C. D. Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454. Genoa Italy, 2006.

S. Dutta and G. Weikum. Cross-document co-reference resolution using sample-based clustering with knowledge enrichment. *Transactions of the Association for Computational Linguistics (TACL)*, 3:15–28, 2015.

M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 457–468, 2016.

C. Gardent, H. Manuélian, K. Striegnitz, and M. Amoia. Generating definite descriptions, non-incrementality, inference, and data. *Trends In Linguistics and Studies in Monographs*, 157:53–86, 2004.

R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision (IJCV)*, 106(2):210–233, 2014.

J. Grimshaw. Boxes and piles and what?s in them: Two extended projections or one. *Architectures, Rules, and Preferences: Variations on Themes by Joan Bresnan, Center for the Study of Language and Information Publications*, pages 199–206, 2007.

A. Guha, M. Iyyer, D. Bouman, J. Boyd-Graber, and J. Boyd. Removing the training wheels: A coreference dataset that entertains humans and challenges computers. In *Proceedings of the Conference of the North American Association for Computational Linguistics (NAACL)*, 2015.

I. Gurobi Optimization. Gurobi optimizer reference manual, 2015. URL `http://www.gurobi.com`.

S. Harabagiu and S. Maiorano. Knowledge-lean coreference resolution and its relation to textual cohesion and coherence. In *Proceedings of the workshop on the relation of discourse/-dialogue structure and reference at the Conference of the Association for Computational Linguistics (ACL)*, pages 29–38, 1999.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

M. Hodosh, P. Young, C. Rashtchian, and J. Hockenmaier. Cross-caption coreference resolution for automatic image understanding. In *Proceedings of the Conference on Computational Natural Language Learning (CONLL)*, pages 162–171, 2010.

Y. Hou, K. Markert, and M. Strube. Global inference for bridging anaphora resolution. In *HLT-NAACL*, pages 907–917, 2013.

R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4555–4564, 2016.

R. Iida, M. Yasuhara, and T. Tokunaga. Multi-modal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 84–92, 2011.

A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2015.

S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

C. Kennington and D. Schlangen. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, 2015.

B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3565, 2014.

K. Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70, 1970.

R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123 (1):32–73, 2017.

J. Krishnamurthy and T. Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics (TACL)*, 1:193–206, 2013.

A. Lai, Y. Bisk, and J. Hockenmaier. Natural language inference from multiple premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP)*, pages 100–109. Asian Federation of Natural Language Processing, 2017.

H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Conference on Computational Natural Language Learning (CONLL)*, pages 28–34, 2011.

K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 188–197, 2017.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. 2014.

C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`.

M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

S. Martschat and M. Strube. Latent structures for coreference resolution. *Transactions of the Association for Computational Linguistics (TACL)*, 3:405–418, 2015.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119. 2013.

R. Mitkov. Robust pronoun resolution with limited knowledge. In *Proceedings of the Joint Conference of the Association for Computational Linguistics and International Conference on Computational Linguistics*, pages 869–875, 1998.

V. Ng and C. Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 104–111, 2002.

R. Passonneau. Measuring agreement on set-valued items (masi) for semantic and pragmatic annotation. *Language Resources and Evaluation*, 2006.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

B. A. Plummer, L. Wang, C. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649, 2015.

B. A. Plummer, A. Mallya, C. Cervantes, J. Hockenmaier, and S. Lazebnik. Phrase localization and visual relationship detection with comprehensive linguistic cues. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Conference on Computational Natural Language Learning (CONLL)*, 2012.

V. Punyakanok, D. Roth, W.-t. Yih, and D. Zimak. Semantic role labeling via integer linear programming inference. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, page 1346, 2004.

C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics, 2010.

S. Singh, A. Subramanya, F. Pereira, and A. McCallum. Large-scale cross-document coreference using distributed inference and hierarchical models. In *Proceedings of the Conference on Human Language Technology and the North American Association for Computational Linguistics (HLT-NAACL)*, pages 793–803, 2011.

W. M. Soon, H. T. Ng, and D. C. Y. Lim. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544, 2001.

S. Wiseman, A. M. Rush, and S. M. Shieber. Learning global features for coreference resolution. *arXiv preprint arXiv:1604.03035*, 2016.

S. J. Wiseman, A. M. Rush, S. M. Shieber, and J. Weston. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, 2015.

P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2:67–78, 2014.

Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004, 2016.