Cross-caption coreference resolution for automatic image understanding

Micah Hodosh Peter Young Cyrus Rashtchian Julia Hockenmaier
Department of Computer Science
University of Illinois at Urbana-Champaign
{mhodosh2, pyoung2, crashtc2, juliahmr}@illinois.edu

Abstract

Recent work in computer vision has aimed to associate image regions with keywords describing the depicted entities, but actual image 'understanding' would also require identifying their attributes, relations and activities. Since this information cannot be conveyed by simple keywords, we have collected a corpus of "action" photos each associated with five descriptive captions. In order to obtain a consistent semantic representation for each image, we need to first identify which NPs refer to the same entities. We present three hierarchical Bayesian models for cross-caption coreference resolution. We have also created a simple ontology of entity classes that appear in images and evaluate how well these can be recovered.

1 Introduction

Many photos capture a moment in time, telling a brief story of people, animals and objects, their attributes, and their relationship to each other. Although different people may give different interpretations to the same picture, people can readily interpret photos and describe the entities and events they perceive in complex sentences. This level of image understanding still remains an elusive goal for computer vision: although current methods may be able to identify the overall scene (Quattoni and Torralba, 2009) or some specific classes of entities (Felzenszwalb et al., 2008), they are only starting to be able to identify attributes of entities (Farhadi et al., 2009), and are far from recovering a complete semantic interpretation of the depicted situation. Like natural language processing, computer vision requires suitable training data, and there are currently no publicly available data sets that would enable the development of such systems.

Photo sharing sites such as Flickr allow users to annotate images with keywords and other descriptions, and vision researchers have access to large collections of images annotated with keywords (e.g. the Corel collection). A lot of recent work in computer vision has been aimed at predicting these keywords (Blei et al., 2003; Barnard et al., 2003; Feng and Lapata, 2008; Deschacht and Moens, 2007; Jeon et al., 2003). But keywords alone are not expressive enough to capture relations between entities. Some research has used the text that surrounds an image in a news article as a proxy (Feng and Lapata, 2008; Deschacht and Moens, 2007). However, in many cases, the surrounding text or a user-provided caption does not simply describe what is depicted in the image (since this is usually obvious to the human reader for which this text is intended), but provides additional information. We have collected a corpus of 8108 images associated with several simple descriptive captions. In contrast to the text near an image on the web, the captions in our corpus provide direct, if partial and slightly noisy, descriptions of the image content. Our data set differs from paraphrase corpora (Barzilay and McKeown, 2001; Dolan et al., 2004) in that the different captions of an image are produced independently by different writers. There are many ways of describing the same image, because it is often possible to focus on different aspects of the depicted situation, and because certain aspects of the situation may be unclear to the human viewer.

One of our goals is to use these captions to obtain a semantic representation of each image that is consistent with all of its captions. In order to obtain such a representation, it is necessary to identify the entities that appear in the image, and to perform cross-caption coreference resolution, i.e. to identify all mentions of the same entity in the five captions associated with an image. In this paper, we compare different meth-

Proceedings of the Fourteenth Conference on Computational Natural Language Learning, pages 162–171,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

A golden retriever (ANIMAL) is playing with **a smaller black and brown dog**(ANIMAL) in a **pink collar** (CLOTHING).
A smaller black dog (ANIMAL) is fighting with **a larger brown dog** (ANIMAL) in **a forest** (NAT_BACKGROUND).
A smaller black and brown dog (ANIMAL) is jumping on a **large orange dog** (ANIMAL).
Brown dog (ANIMAL) with **mouth** (BODY_PART) open near **head**(BODY_PART) of **black and tan dog** (ANIMAL).
Two dogs (ANIMAL) playing near **the woods** (NAT_BACKGROUND).

Figure 1: An image with five captions from our corpus. Coreference chains and ontological classes are indicated in color.

ods of cross-caption coreference resolution on our corpus. In order to facilitate further computer vision research, we have also defined a set of coarse-grained ontological classes that we use to automatically categorize the entities in our data set.

2 A corpus of action images and captions

Image collection and sentence annotation We have constructed a corpus consisting of 8108 photographs from Flickr.com, each paired with five one-sentence descriptive captions written by Amazon's Mechanical Turk[1] workers. We downloaded a few thousand images from each of six selected Flickr groups[2]. To facilitate future computer vision research on our data, we filtered out images in black-and-white or sepia, as well as images with watermarks, signatures, borders or other obvious editing. Since our collection focuses on images depicting actions, we then filtered out images of scenery, portraits, and mood photography. This was done independently by two members of our group and adjudicated by a third.

We paid Turk workers $0.10 to write one descriptive sentence for each of five distinct and randomly chosen images that were displayed one at a time. We required a small qualification test that examined the workers' English grammar and spelling and we restricted the task to U.S. workers (see Rashtchian et al. (2010) for more details). Our final corpus contains five sentences for each of our 8108 images, totaling 478,317 word tokens, and an average sentence length of 11.8 words. We first spell-checked[3] these sentences, and used OpenNLP[4] to POS-tag them. We identified NPs using OpenNLP's chunker, followed by a semi-automatic procedure to correct for a number of systematic chunking errors that could easily be corrected. We randomly selected 200 images for further manual annotation, to be used as test and development data in our experiments.

Gold standard coreference annotation We manually annotated NP chunks, ontological classes, and cross-caption coreference chains for each of the 200 images in our test and development data. Each image was annotated independently by two annotators and adjudicated by a third.[5] The development set contains 1604 mentions. On average, each caption has 3.2 mentions, and each image has 5.9 coreference chains (distinct entities).

Ontological annotation of entities In order to understand the role entities mentioned in the sentences play in the image, we have defined a simple ontology of entity classes (Table 1). We distinguish entities that constitute the background of an image from those that appear in the foreground. These entities can be animate (people or animals) or inanimate. For inanimate objects, we distinguish static objects from "movable" objects. We also distinguish man-made from natural objects and backgrounds, since this matters for computer vision algorithms. We have labeled the entity mentions in our test and development data with classes from this ontology. Again, two of us annotated each image's mentions, and adjudication was performed by a single person. Our ontology is similar to, but smaller than the one proposed by Hollink and Worring (2005) for video retrieval, which in turn is based on Hoogs et al. (2003) and Hunter (2001).

3 Predicting image entities from captions

Figure 1 shows an image from our corpus. Different captions use different words to refer to the

[1] https://www.mturk.com

[2] The groups:"strangers!", "Wild-Child (Kids in Action)", "Dogs in Action (Read the Rules)", "Outdoor Activities", "Action Photography", "Flickr-Social (two or more people in the photo)".

[3] We used Unix's `aspell` to generate possible corrections and chose between them based on corpus frequencies.

[4] http://opennlp.sourceforge.net

[5] We used MMAX2 (Müller and Strube, 2006) both for annotation and adjudication.

Ontological Class	Examples
animal	*dog, horse, cow*
background_man-made	*street, pool, carpet*
background_natural	*ocean, field, air*
body_part	*hair, mouth, arms*
clothing	*shirt, hat, sunglasses*
event	*trick, sunset, game*
fixed_object_man-made	*furniture, statue, ramp*
fixed_object_natural	*rock, puddle, bush*
image_attribute	*camera, picture, closeup*
material_man-made	*paint, frosting*
material_natural	*water, snow, dirt*
movable_man-made	*ball, toy, bowl*
movable_natural	*leaves, snowball*
nondepictable	*something, Batman*
orientation	*front, top, [the] distance*
part_of	*edge, side, top, tires*
person	*family, skateboarder*
property_of	*shadow, shade, theme*
vehicle	*surfboard, bike, boat*
writing	*graffiti, map*

Table 1: Our ontology for entities in images.

same entity, or even seemingly contradictory modifiers (*"orange"* vs. *"brown"* dog). In order to predict what entities appear in an image from its captions, we need to identify how many entities each sentence describes, and what role these entities play in the image (e.g. person, animal, background). Because we have five sentences associated with each image, we also need to identify which noun phrases in the different captions of the same image refer to the same entity. Because the captions were generated independently, there are no discourse cues such as anaphora to identify coreference. This creates problems for standard coreference resolution systems trained on regular text. Our data also differs from standard coreference data sets in that entities are rarely referred to by proper nouns.

Our first task is to identify which noun phrases may refer to the same entity. We do this by identifying the set of entity types that each NP may refer to. We use WordNet (Fellbaum, 1998) to identify the possible entity types (WordNet synsets) of each head noun. Since the salient entities in each image are likely to be mentioned by more than one caption writer, we then aim to restrict those types to those that may be shared by some head nouns in the other captions of the same image. This gives us an inventory of entity types for each mention, which we use to identify coreferences, restricted by the constraint that all coreferent mentions refer to an entity of the same type.

4 Using WordNet to identify entity types

WordNet (Fellbaum, 1998) provides a rich ontology of entity types that facilitates our coreference task.[6] We use WordNet to obtain a lexicon of possible entity types for each mention (based on their lexical heads, assumed to be the last word with a nominal POS tag[7]). We first generate a set of candidate synsets based solely on the lexical heads, and then generate lexicon entries based on relations between the candidates.

WordNet synsets provide us with synonyms, and hypernym/hyponym relations. For each mention, we generate a list of candidate synsets. We require that the candidates are one of the first four synsets reported and that their frequency is to be at least one-tenth of the most frequent synset. We limit candidates to ones with "physical_entity#n#1", "event#n#1", or "visual_property#n#1" as a hypernym, in order to ensure that the synset describes something that is depictable. To avoid word senses that refer to a person in a metaphorical fashion, (e.g. *pig* meaning slovenly person or *red* meaning communist), we ignore synsets that refer to people if the word has a synset that is an animal or color.[8]

In general, we would like for mentions to be able to take on more specific word senses. For example, we would like to be able to identify that *"woman"* and *"person"* may refer to the same entity, whereas *"man"* and *"woman"* typically would not. However, we also do not want a type inventory that is too large or too fine-grained.

Once the candidate synsets are generated, we consider all pairs of nouns (n_1, n_2) that occur in different captions of the same image and examine all corresponding pairs of candidate synsets (s_1, s_2). If s_2 is a synonym or hypernym of s_1, it is possible that two captions have different words describing the same entity, so we add s_1 and s_2[9] to the lexicon of n_1. Adding s_2 to n_1's lexicon allows it to act as an umbrella sense covering other nouns describing the same entity.[10] We add s_2 to

[6] For the prediction of ontological classes, we use our own ontology because WordNet is too fine-grained for this purpose.

[7] If there are two NP chunks that form a "[NP ... group] of [NP...]" construction, we only use the second NP chunk.

[8] An exception list handles cases (*diver, blonde*), where the human sense is more likely than the animal or color sense.

[9] We don't add s_2 if it is "object#n#1" or "clothing#n#1".

[10] This is needed when captions use different aspects of the entity to describe it (for example, *"skier"* and *"a skiing man"*).

the lexicon of n_2 (since if n_1 is using the sense s_1, then n_2 must be using the sense s_2) and if n_1 occurs at least five times in the corpus, we add s_1 to the lexicon of n_2.

5 A heuristic coreference algorithm

Based on WordNet candidate synsets, we define a heuristic algorithm that finds the optimal entity assignment for the mentions associated with each image. This algorithm is based on the principles driving our generative model described below, and on the observation that salient entities will be mentioned in many captions and that captions tend to use similar words to describe the same entity.

Simple heuristic algorithm:

1. For each noun, choose the synset that appears in the most number of captions of an image, and break ties by choosing the synset that covers the fewest distinct lemmatized nouns.
2. Group all of the noun phrase chunks that share a synset into a single coreference chain.

6 Bayesian coreference models

Since we cannot afford to manually annotate our entire data set with coreference information, we follow Haghighi and Klein (2007)'s work on unsupervised coreference resolution, and develop a series of generative Bayesian models for our task.

6.1 Model 0: Simple Mixture Model

In our first model, based on Haghighi and Klein's baseline Dirichlet Process model, each image i corresponds to the set of observed mentions $\mathbf{w}^i$ from across its captions. Image i has a hidden global topic T_i, drawn from a distribution with a GEM prior with hyperparameter γ as explained by Teh et al. (2006). In a Dirichlet process, the GEM distribution is an infinite analog of the Dirichlet distribution, allowing for a potentially infinite number of mixture components. $P(T_i = t)$ is proportional to γ if t is a new component, or to the number of times t has been drawn before otherwise. Given a topic choice $T_i = t$, entity type assignments Z_j for all mentions w_j in image i are in turn drawn from a topic-specific multinomial θ_t over all possible entity types $\mathbf{E}$ that was drawn from a Dirichlet prior with hyperparameter β. Similarly, given an entity type $Z_i = z$, each corresponding (observed) head word w_j is drawn from an entity type-specific multinomial ϕ_z over all possible words $\mathbf{V}$, drawn from a finite Dirichlet prior with hyperparameter α. The set of all images belonging to the same topic is analogous to an individual document in Haghighi and Klein's baseline model.[11] All headwords of the same entity type are assumed to be coreferent, similar to Haghighi and Klein's model. As described in section 4, we use WordNet to identify the subset of types that can actually produce the given words. Therefore, similar to the way Andrzejewski and Zhu (2009) handled a priori knowledge of topics, we will define an indicator variable δ_{ij} that is 1 iff the WordNet information allows word i to be produced from entity set j and 0 otherwise.

6.1.1 Sampling Model 0

We find $arg\,max_{\mathbf{Z},\mathbf{T}} P(\mathbf{Z}, \mathbf{T}|\mathbf{X})$ with Gibbs sampling. Here, $\mathbf{Z}$ and $\mathbf{T}$ are the collection of type and topic assignments, with $\mathbf{Z}^{-j} = \mathbf{Z} - \{Z_j\}$ and $\mathbf{T}^{-i} = \mathbf{T} - \{T_i\}$. This style of notation will be extended analogously to other variables. Let $n_{e,x}$ represent the number of times word x is produced from entity e across all topics and let p_j be the number of images assigned to topic j. Let $m_{t,e}$ represent the number of times entity type e is generated by topic t. Each iteration consists of two steps: first, each Z_i is resampled, fixing $\mathbf{T}$; and then each T_i is resampled based on $\mathbf{Z}$[12].

1. Sampling Z_j:

$$\begin{aligned}
P(Z_j = e|w_j \in \mathbf{w}^i, \mathbf{Z}^{-j}, \mathbf{T}) &\propto P(w_j|Z_j = e)P(Z_j = e|T_i) \\
P(w_j = x|Z_j = e) &\propto \left(\frac{n_{e,x}^{-j} + \alpha}{\sum_{x'} n_{e,x'}^{-j} + \alpha}\right)\delta_{xe} \\
P(Z_j = e|T_i = t) &= \frac{m_{t,e}^{-j} + \beta}{\sum_{e'} m_{t,e'}^{-j} + \beta}
\end{aligned}$$

2. Sampling T_i:

$$\begin{aligned}
P(T_i = j|\mathbf{w}, \mathbf{Z}, \mathbf{T}^{-i}) &\propto P(T_i = j|\mathbf{T}^{-i})P(\mathbf{Z}|T_i = j, \mathbf{T}^{-i}) \\
&\propto P(T_i = j|\mathbf{T}^{-i})\prod_{k \in \mathbf{w}^i} P(Z_k|T_i = j) \\
&= P(T_i = j|\mathbf{T}^{-i})\prod_{k \in \mathbf{w}^i} \frac{m_{j,Z_k}^{-i} + \beta}{\sum_{e'} m_{j,e'}^{-i} + \beta} \\
P(T_i = j|\mathbf{T}^{-i}) &\propto \begin{cases} \gamma, & \text{If its a new topic} \\ p_j & \text{Otherwise} \end{cases}
\end{aligned}$$

[11] Since we do not have multiple images of the same well-known people or places, referred to by their names, we do not perform any cross-image coreference

[12] Sampling on the exponentiated posterior to find the mode as Haghighi and Klein (2007) did was found to not significantly affect results on our tasks

Image 21: Caption 1: {$x_{21,1}$:*a golden* **retriever**; $x_{21,2}$:*a smaller black and brown* **dog**; $x_{21,3}$:*a pink* **collar**}
Caption 2: {$x_{21,4}$:*a smaller black* **dog**; $x_{21,5}$:*a larger brown* **dog**; $x_{21,6}$:*a* **forest**}
Caption 3: {$x_{21,7}$:*small black and brown* **dog**; $x_{21,8}$:*a large orange* **dog**}
Caption 4: {$x_{21,9}$:*brown* **dog**; $x_{21,10}$:**mouth**; $x_{21,11}$:**head**; $x_{21,12}$:*black and tan* **dog**}
Caption 5: {$x_{21,13}$:*two* **dogs**; $x_{21,14}$:*the* **woods**}

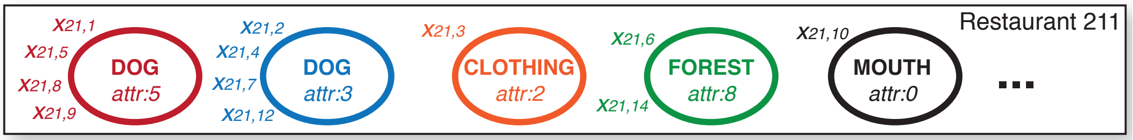

Figure 2: Models 1 and 2 as Chinese restaurant franchises: each image topic is a franchise, each image is a restaurant, each entity is a table, each mention is a customer. Model 2 adds attributes (in italics).

6.2 Model 1: Explicit Entities

Model 0 does not have an explicit representation of entities beyond their type and thus cannot distinguish multiple entities of the same type in an image. Although Model 1 also represents mentions only by their head words (and thus cannot distinguish *black dog* from *brown dog*), it creates explicit entities based on the Chinese restaurant franchise interpretation of the hierarchical Dirichlet Process model (Teh et al., 2006). Figure 2 (ignoring the modifiers / attributes for now) illustrates the Chinese restaurant franchise interpretation of our model. Using this metaphor, there are a series of restaurants (= images), each consisting of a potentially infinite number of tables (= entities), that are frequented by customers (= entity mentions) who will be seated at the tables. Restaurants belong to franchises (= image topics). Each table is served one dish (= entity type, e.g. DOG, CLOTHING) shared by all the customers. The head word of a mention $x_{i,j}$ is generated in the following manner: customer j enters restaurant i (belonging to franchise T_i) and sits down at one of the existing tables with probability proportional to the number of other customers there, or sits at a new table with probability proportional to a constant. A dish e^i_a (DOG) from a potentially infinite menu is served to each new table a, with probability proportional to the total number of tables it is served at in the franchise T_i (or to a constant if it is a new dish). The (observed) head word of the mention $x_{j,i}$ (*dog, retriever*) is then drawn from the multinomial distribution over words defined by the entity type (DOG) at the table. The menu (set of dishes) available to each restaurant and table is restricted by our lexicon of WordNet synsets for each mention. More formally, each image topic t defines a distribution over entities drawn from a global GEM prior with hyperparameter κ. Therefore, the probability of an entity a is proportional to the number of its existing mentions in images of the same topic, or to κ, if it is previoulsy unmentioned. The type of each entity, e_a, is drawn from a topic-dependent multinomial with global Dirichlet prior. The head words of mentions are generated by their entity type as in Model 0. Mentions assigned to the same entity are considered to be coreferent. Based on the nature of our corpus, we again assume that two words cannot be coreferent within a sentence, restrict the distribution to not allow inter-sentence coreference and renormalize the values accordingly.

6.2.1 Sampling Model 1

There are three parts to our resampling procedure: resampling the entity assignment for each word, resampling the entity type for each entity, and resampling the topic of each image. The kth word of image i, sentence j, will now be $w^i_{j,k}$; e^i_a is the entity type of entity a in image i; $a^i_{j,k}$ is the entity that word k of sentence j is produced from in image i, and $Z^i_{j,k}$ represents that entity's type. $\mathbf{a}$ is the set of all current entity assignments and $\mathbf{e}$ are the type assignments for entities. m is now defined as the number of entities of a certain type being drawn for an image, n is defined as before and $c_{i,a}$ is the number of times entity a is expressed in image i. Topics are resampled as in Model 0.

Entity Assignment Resampling Entity assignments for words are resampled one sentence at a time in the order the headwords appear in the sentence. For each word in the sentence, entity assignments are defined by the distribution of Figure 3. The headword is assigned to an existing entity with probability proportional to the number of entities already assigned to that entity and the probability that the entity emits that word. The word is assigned to a new entity with a newly

Model 1:

$$P(e^i_a = e|\mathbf{a}, \mathbf{w}, T_i = t, \mathbf{e}^{-\mathbf{i},\mathbf{a}}) \propto (m^{-\mathbf{e}^{\mathbf{i},\mathbf{a}}}_{t,e} + \beta) \prod_{\{w^i_{j,k}=x|a^i_{j,k}=a\}} \frac{n^{-\mathbf{e}^{\mathbf{i},\mathbf{a}}}_{e,x} + \alpha}{\sum_y (n^{-\mathbf{e}^{\mathbf{i},\mathbf{a}}}_{e,y} + \alpha)} \delta_{x,e}$$

$$P(a^i_{j,k} = a|\mathbf{a}^{-(i,j,k'|k' \geq k)}, \mathbf{e}, \mathbf{T}) \propto \begin{cases} c^{-j,k'}_{i,a} P(w^i_{j,k}|Z^i_{j,k} = e^i_a, \mathbf{a}^{-(i,j,k'|k' \geq k)}) \rho^i_{j,a}, \text{ if } a \text{ is not new} \\ \kappa P(e^i_a|\mathbf{e}^{-\mathbf{i},\mathbf{a}}, T_i) P(w^i_{j,k}|Z^i_{j,k} = e^i_a, \mathbf{a}^{-(i,j,k'|k' \geq k)}), \text{o/w} \end{cases}$$

With:

$$P(w^i_{j,k} = x|Z^i_{j,k} = e, \mathbf{a}^{-(i,j,k'|k' \geq k)}) \propto \frac{n^{-(i,j,k'|k' \geq k)}_{e,x} + \alpha}{\sum_y (n^{-(i,j,k'|k' \geq k)}_{e,y} + \alpha)} \delta_{x,e}$$

$$P(e^i_a = e|\mathbf{e}^{-\mathbf{i},\mathbf{a}}, T_i = t) \propto \frac{m^{-\mathbf{e}^{\mathbf{i},\mathbf{a}}}_{t,e} + \beta}{\sum_{e'} (m^{-\mathbf{e}^{\mathbf{i},\mathbf{a}}}_{t,e'} + \beta)}$$

Model 2:

$$P(a^i_{j,k} = a|\mathbf{a}^{-(i,j,k'|k' \geq k)}, \mathbf{e}, \mathbf{T}, \mathbf{b}) \propto \begin{cases} c^{-j,k'}_{i,a} P(w^i_{j,k}|Z^i_{j,k} = e^i_a, \mathbf{a}^{-(i,j,k'|k' \geq k)}) \rho^i_{j,a} P(\mathbf{d}^i_{j,k}|b^i_a), \text{ if } a \text{ is not new} \\ \kappa P(e^i_a|\mathbf{e}^{-\mathbf{i},\mathbf{a}}, T_i) P(w^i_{j,k}|Z^i_{j,k} = e^i_a, \mathbf{a}^{-(i,j,k'|k' \geq k)}) P(b^i_a) P(\mathbf{d}^i_a|b^i_a), \text{o/w} \end{cases}$$

$$P(b^i_a = b|\mathbf{D}^i_a, \mathbf{b}^{-\mathbf{i},\mathbf{a}}) \propto P(b^i_a = b|\mathbf{b}^{-i,a}) \prod_{d \in \mathbf{D}^i_a} \frac{(s^{-i,a}_{b,d} + \zeta)}{\sum_{d'} (s^{-i,a}_{b,d'} + \zeta)}$$

Figure 3: Sampling equations for Models 1 and 2

drawn entity type with probability proportional κ, the probability that the entity type is for an image of the given topic (normalized over WordNet's possible entities for the word), and the probability the drawn type produces the word. $\rho^i_{j,a} = 1$ iff entity a of image i does not appear in sentence j and $\rho^i_{j,a} = 0$ otherwise. $\mathbf{a}^{-(i,j,k'|k' \geq k)}$ represents removing the kth or later words in sentence j of image i

Entity Type Resampling Fixing the assignments, the type of each entity is redrawn based on the distribution in Figure 3. It is proportional to the probability that a certain entity type is in an image of a given topic and, independently for each of the words, the probability that the given word expresses the type. $n^{-\mathbf{e}^{\mathbf{i},\mathbf{a}}}_{e,x}$ is the number of times entity type e is expressed as word x not counting the words attached to the currently entity being resampled and $m^{-\mathbf{e}^{\mathbf{i},\mathbf{a}}}_{t,e}$ is the number of times an entity of type e appears in an image of topic t not counting the current entity being resampled. The probability of a given image belonging to a topic is proportional the number of images already in the topic (or γ) followed by the probability that each of the entities in the image were drawn from that topic.

6.3 Model 2: Explicit Entities and Modifiers

Certain entities cannot be distinguished simply by head word alone, such in the example in Figure 2. Model 2 augments Model 1 with the ability to generate modifiers. In addition to an entity type, each entity draws an attribute from a global distribution drawn from a GEM distribution with hyperparameter η. An attribute is a multinomial distribution, on possible modifier words, drawn from a Dirichlet prior with parameter ζ. From the attribute, each modifier word is drawn independently. Therefore given an attribute b and a set of modifiers $\mathbf{d}$: $P(\mathbf{d}|b) \propto \prod_{d \in \mathbf{d}} (s_d + \zeta)$ where s_d is number of times modifier d is produced by attribute b. In addition, the probability of a certain attribute b given all other assignments is given by:

$$P(b^i_a = b|\mathbf{b}^{-i,a}) \propto \begin{cases} \eta, & \text{If its a new attribute} \\ r_b, & \text{Otherwise} \end{cases}$$

where r_b is the number of entities with attribute b. As in Model 1, mentions assigned to the same entity are considered coreferent. Consider the *"smaller black dog"* mention in Figure 2. When the mention is being resampled, the attribute choice for each table will bias the probability distribution towards the table whose attribute is more likely to produce *"smaller"* and *"black"*. Therefore, the model can now better distinguish the two dogs in the image.

6.3.1 Sampling Model 2

The addition of modifiers only directly effects the distribution when resampling entity assignments since attributes are independent of entity types, image topics, and headwords of noun phrases. The sampling distribution are again shown in Figure 3. In a separate sampling step, it is now necessary to resample the attribute assigned to each entity: The probability of drawing a certain attribute is illustrated in Figure 3 with $\mathbf{D}^i_a$ as the set of all the modifiers of all the noun phrases currently assigned to entity a of image i, and $s_{b,d}^{-i,a}$ as the number of times attribute b produces modifier d without the current assignment of entity a of image i.

6.4 Implementation

The topic assignments for each image are initialized to correspond to the Flickr groups the images were taken from. Each mention was initialized as its own entity with type and attribute sampled from a uniform distribution.

As our training is unsupervised, each of the models were ran on the entire dataset. For Model 0, after burn-in, 20 samples of Z were taken spaced 200 iterations apart, while for Model 1 samples were taken spaced 100 apart, and 25 apart for Model 2. The implementation of Model 2 ran too slow to effectively judge when burn in occurred, impacting the results.

The values of parameters α, β, γ, κ, η, ζ, and the number of initial attributes were hand-tuned based on the average performance on our annotated development subset of 100 images.[13]

7 Evaluation of coreference resolution

We evaluate each of the generative models and the heuristic coreference algorithm on the annotated test subset of our corpus consisting of 100 images with both the OpenNLP chunking and the gold standard chunking. We report our scores based on the MUC evaluation metric. The results are reported in Table 2 as the average scores across all the samples of two independent runs of each model. We also present results on Model 0 without using WordNet where every word can be an expression of one of 200 fake entity sets. The same table also shows the performance of a baseline model and the upper bound on performance imposed by WordNet.

A baseline model: In our baseline model, two noun phrases in captions of the same image are coreferent if they share the same head noun.

Upper bound on performance: Although WordNet synsets provide a good indication of whether two mentions can refer to the same entity or not, they may also be overly restrictive in other cases. We measure the upper bound on performance that our reliance on WordNet imposes by finding the best-scoring coreference assignment that is consistent with our lexicon. This achieves an F-score of 90.2 on the test data with gold chunks.

Performance increases in each subsequent model. The heuristic beats each of the models, but in some sense it is an extreme version of Model 1. Both it and Model 1 attempt to produce entity sets that cover as many captions as possible, while minimizing the number of distinct words involved. The heuristic locally forces this case, at the expense of no longer being a generative model.

8 Ontological Class Prediction

As a further step towards understanding the semantics of images, we develop a model that labels each entity with one of the ontological classes defined in section 2. The immediate difficulty of this task is that our ontology includes not only semantic distinctions, but also spatial and visual ones. While it may be easy to tell which words are animals and which are people, there is only a fine distinction at the language level whether an object is movable, fixed, or part of the background.[14]

8.1 Model and Features

We define our task as a classification problem, where each entity must be assigned to one of twenty classes defined by our ontology. We use a Maximum Entropy classifier, implemented in the MALLET toolkit (McCallum, 2002), and define the following text features:

NP Chunk: We include all the words in the NP chunk, unfiltered.

WordNet Synsets and Hypernyms: The most likely synset is either the first one that appears in WordNet or one of the ones predicted by our coreference system. For each of these possibilities, we include all of that synset's hypernyms.

Syntactic Role: We parsed our captions with the C&C parser (Clark and Curran, 2007), and record whether the word appears as a direct object of a verb, as the object of a preposition, as the subject of the sentence, or as a modifier. If it is a modifier, we also add the head word of the phrase being modified.

[13] $(0.1, 0.1, 100, 0.001875, 100, 0.0002, 20)$ respectively.

[14] For example, we deem bowls and silverware to be movable objects; furniture, fixed; and carpets, background. Moreover, in all three cases, we must correctly distinguish that these objects are man-made and not found in nature.

Model	OpenNLP chunks			Gold chunks		
	Rec.	Prec.	F1	Rec.	Prec.	F1
Baseline	57.3	89.5	69.9	64.1	96.2	77.0
Upper bound				82.1	100	90.2
WN Heuristic	70.6	84.8	77.0	80.4	90.6	85.2
Model 0 w/o WN	79.7	59.8	68.4	85.1	62.7	72.2
Model 0	66.8	83.1	74.1	75.9	90.3	82.5
Model 1	69.6	83.8	76.0	78.0	90.8	83.9
Model 2	69.2	84.4	76.1	77.9	91.5	84.1

Table 2: Coreference resolution results (MUC scores; Models 0-2 are averaged over all samples)

8.2 Experiments

We use two baselines. The naive baseline categorizes words by selecting the most frequent class of the word. If no instances of the word have occurred, it uses the overall most frequent class. The WordNet baseline works by finding the most frequent class amongst the most relevant synsets for a word. It calculates the class frequency for each synset by assuming each word has the sense of its first synset and incrementing the frequency of the first synset and its hypernyms. When categorizing a word, it finds the set of closest hypernyms of the word that have a non-zero frequency, and chooses the class with the greatest sum of frequency counts amongst those hypernyms.

We train the MaxEnt classifier using semi-supervised learning. Initially, we train a classifier using the 500 sentence gold standard development set. For each class, we use the top 5%[15] of the labels to label the unlabeled data and provide additional training data. We then retrain the classifier on the newly labeled examples and the development set, and run it on the test set. For each coreference chain in the test set, we relabel all of the mentions in the chain to use the majority class, if a clear majority exists. If no such majority exists, we leave the labels as is. The MaxEnt classifier experiments were conducted by varying the source of the synset assigned to each word. For each of our coreference systems, we report two scores (Table 3). The first is the average accuracy when using the output from two runs of each model with about 20 samples per run, and the second uses the output that performs best on the coreference task when scored on the development data.

Discussion Although we use WordNet to classify our entity mentions, we designed our ontology by considering only the images and their captions, with no particular mapping to WordNet in mind. Therefore, these experiments provide of a proof of concept for the semi-supervised labeling of a corpus using any semantic/visual ontology.

Classifier (synset prediction)	Accuracy (gold chunks)		
Naive Baseline	72.0		
WordNet Baseline	81.0		
MaxEnt (1st-synset)	84.4		
MaxEnt (WN heuristic)	82.7		
	Avg.	σ	**Best-Coref**
MaxEnt (Model 1)	83.9	0.5	84.5
MaxEnt (Model 2)	84.1	0.4	85.3

Table 3: Prediction of ontological classes

Overall, Model 2 had the best performance for this task. This demonstrates that the additional features of Model 2 force synset selections that are consistent across the entire corpus, and are sensitive to the modifiers appearing with them. The WordNet heuristic selects synsets in a fairly arbitrary manner - all other things being equal, the synsets are chosen without reference to what other synsets are chosen by similar clusters of nouns.

9 Evaluating entity prediction

Together, the coreference resolution algorithm and ontology classification model provide us with a set of distinct, ontologically-categorized entities appearing in each image. We perform a final experiment to evaluate how well our models can recover the mentioned entities and their ontological types for each image. We now represent each entity as a tuple (L, c), where L is its coreference chain, and c is the ontological class of these mentions. [16]

We compute the precision and recall between the predicted and gold standard tuples for each image. We consider a tuple (L', c') correctly predicted only when a copy of (L', c') occurs both in the set of predicted tuples and the set of gold standard tuples.[17] Then, as usual, for precision we

[15] This was tuned using 10-fold cross-validation of the development set.

[16] Note that for each image, the tuples of all entities correspond to a partition of the set of the head-word mentions in an image.

[17] We assign no partial credit because incorrect typing or

Model	Recall	Precision	F-score
Baseline	28.4	20.6	23.9
WordNet Heuristic	48.3	43.9	46.0
Model 1 (avg)	51.7	42.8	46.8
Model 1 (best-coref)	50.9	45.4	48.0
Model 2 (avg)	52.2	42.7	47.0
Model 2 (best-coref)	52.3	46.0	49.0

Table 4: Overall entity recovery. We measure how many entities we identify correctly (requiring complete recovery of their coreference chains and correct prediction of their ontological class.

normalize the number of overlapping tuples by the number of predicted tuples, and for recall, by the number of gold standard tuples. We report average precision and recall over all images in our test set.

We report scores for four different pairs of ontological class and coreference chain predictions. As a baseline, we use the ontological classes predicted by the our naive baseline and the chains predicted by the “same-words-are-coreferent” coreference resolution baseline.

We also report results using the classes and chains predicted by Model 1, Model 2, and the WordNet Heuristic Algorithm. The influence of the different coreference algorithms comes from the entity types that are used to determine coreference chains, and that also correspond to WordNet candidate synsets. In other words, although the final coreference chain may be predicted by two different models, the synsets they use to do so may differ, affecting the synset and hypernym features used for ontological prediction. We present results in Table 4 for these four different set-ups.

The synsets chosen by the different coreference algorithms clearly have different applicability when it comes to ontological class prediction. Although Model 2 performs comparably to Model 1 and does worse than the WordNet heuristic algorithm for coreference chain prediction, it certainly does better on this task. Since our end goal is creating a unified semantic representation, this final task judges the effectiveness of our models to capture the most detailed entity information. The success of Model 2 means that the incorporation of adjectives informs the proper choice of synsets that are useful in predicting ontological classes.

10 Conclusion

As a first step towards automatic image understanding, we have collected a corpus of images associated with several simple descriptive captions, which provide more detailed information about the image than simple keywords. We plan to make this data set available for further research in computer vision and natural language processing. In order to enable the creation of a semantic representation of the image content that is consistent with the captions in our data set, we use WordNet and a series of Bayesian models to perform cross-caption coreference resolution. Similar to Haghighi and Klein (2009), who find that linguistic heuristics can provide very strong baselines for standard coreference resultion, relatively simple heuristics based on WordNet alone perform surprisingly well on our task, although they are outperformed by our Bayesian models for overall entity prediction. Since our generative models are based on Dirichlet Process priors, they are designed to favor a small number of unique entities per image. In the heuristic algorithm, this bias is built in explicitly, resulting in slightly higher performance on the coreference resolution task. However, while the generative models can use global information to learn what entity type each word is likely to represent, the heuristic is unable to capture any non-local information about the entities, and thus provides less useful input for the prediction of ontological classes.

Future work will aim to improve on these results by overcoming the upper bound on performance imposed by WordNet, and through a more sophisticated model of modifiers. We will also investigate how image features can be incorporated into our model to improve performance on entity detection. Ultimately, identifying the depicted entities from multiple image captions will require novel ways to correctly handle the semantics of plural NPs (i.e. that one caption's *“two dogs”* consist of another's *“golden retreiver”* and *“smaller black dog”*). We foresee similar challenges when dealing with verbs and events.

The creation of an actual semantic representation of the image content is a challenging problem in itself, since the different captions often focus on different aspects of the depicted situation, or provide different interpretation of ambiguous situations. We believe that this poses many interesting challenges for natural language processing, and will ultimately require ways to integrate the information conveyed in the caption with visual features extracted from the image.

incomplete coreference chaining both completely change the semantics of an image.

Acknowledgements

This research was funded by NSF grant IIS 08-03603 INT2-Medium: Understanding the Meaning of Images. We are grateful for David Forsyth and Dan Roth's advice, and for Alex Sorokins support with MTurk.

References

David Andrzejewski and Xiaojin Zhu. 2009. Latent Dirichlet allocation with topic-in-set knowledge. In *NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 43–48.

Kobus Barnard, Pinar Duygulu, David Forsyth, Nando De Freitas, David M. Blei, and Michael I. Jordan. 2003. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse, France, July.

David M. Blei, Michael I, David M. Blei, and Michael I. 2003. Modeling annotated data. In *Proceedings of the 26th International ACM SIGIR Conference*, pages 127–134.

Stephen Clark and James R. Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

Koen Deschacht and Marie-Francine Moens. 2007. Text analysis for automatic image annotation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of Coling 2004*, pages 350–356, Geneva, Switzerland, August. COLING.

A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. 2009. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785, June.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

P. Felzenszwalb, D. McAllester, and D. Ramanan. 2008. A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June.

Yansong Feng and Mirella Lapata. 2008. Automatic image annotation using auxiliary text information. In *Proceedings of ACL-08: HLT*, pages 272–280, Columbus, Ohio, June.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric Bayesian model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 848–855, Prague, Czech Republic.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1152–1161, Singapore, August. Association for Computational Linguistics.

L. Hollink and M. Worring. 2005. Building a visual ontology for video retrieval. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 479–482, New York, NY, USA. ACM.

A. Hoogs, J. Rittscher, G. Stein, and J. Schmiederer. 2003. Video content annotation using visual analysis and a large semantic knowledgebase. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–327 – II–334 vol.2, June.

Jane Hunter. 2001. Adding multimedia to the semantic web - building an mpeg-7 ontology. In *In International Semantic Web Working Symposium (SWWS*, pages 261–281.

Lavrenko Manmatha Jeon, V. Lavrenko, R. Manmatha, and J. Jeon. 2003. A model for learning the semantics of pictures. In *Seventeenth Annual Conference on Neural Information Processing Systems (NIPS)*. MIT Press.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Christoph Müller and Michael Strube. 2006. Multilevel annotation of linguistic data with MMAX2. In Sabine Braun et al, editor, *Corpus Technology and Language Pedagogy*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Ariadna Quattoni and Antonio B. Torralba. 2009. Recognizing indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420. IEEE.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazons mechanical turk. In *NAACL Workshop Creating Speech and Language Data With Amazons Mechanical Turk*.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.