Generating Descriptions of Spatial Relations between Objects in Images

Adrian Muscat
Communications & Computer Engineering
University of Malta
Msida MSD 2080, Malta
adrian.muscat@um.edu.mt

Anja Belz
Computing, Engineering & Maths
University of Brighton
Lewes Road, Brighton BN2 4GJ, UK
a.s.belz@brighton.ac.uk

Abstract

We investigate the task of predicting prepositions that can be used to describe the spatial relationships between pairs of objects depicted in images. We explore the extent to which such spatial prepositions can be predicted from (a) language information, (b) visual information, and (c) combinations of the two. In this paper we describe the dataset of object pairs and prepositions we have created, and report first results for predicting prepositions for object pairs, using a Naive Bayes framework. The features we use include object class labels and geometrical features computed from object bounding boxes. We evaluate the results in terms of accuracy against human-selected prepositions.

1 Introduction

The task we investigate is predicting the prepositions that can be used to describe the spatial relationships between pairs of objects in images. This is not the same as inferring the actual 3-D real-world spatial relationships between objects, but has some similarities with that task. This is an important subtask in automatic image description (which is important not just for assistive technology, but also for applications such as text-based querying of image databases), but it is rarely addressed as a subtask in its own right. If an image description method produces spatial prepositions it tends to be as a side-effect of the overall method (Mitchell et al., 2012; Kulkarni et al., 2013), or else relationships are not between objects, but e.g. between objects and the 'scene' (Yang et al., 2011). An example of preposition selection as a separate sub-task is Elliott & Keller (2013) where the mapping is hard-wired manually.

Our main data source is a corpus of images (Everingham et al., 2010) in which objects have been annotated with rectangular bounding boxes and object class labels. For a subset of 1,000 of the images we also have five human-created descriptions of the whole image (Rashtchian et al., 2010).

We collected additional annotations for the images (Section 2.3) which list, for each object pair, a set of prepositions that have been selected by human annotators as correctly describing the spatial relationship between the given object pair.

The aim is to create models for the mapping from image, bounding boxes and labels to spatial prepositions as indicated in Figure 1. In this we use a range of features to represent object pairs, computed from image, bounding boxes and labels. We investigate the predictive power of different types of features within a Naive Bayes framework (Section 3), and report first results in terms of two measures of accuracy (Section 4).

2 Data

2.1 VOC'08

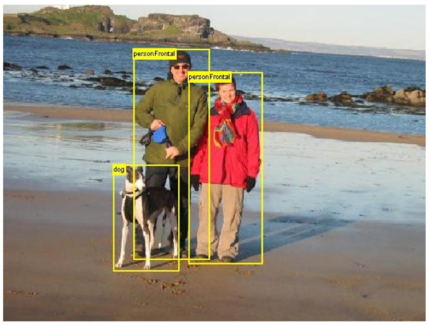
The PASCAL VOC 2008 Shared Task Competition (VOC'08) data consists of 8,776 images and 20,739 objects in 20 object classes (Everingham et al., 2010). In each image, every object belonging to one of the 20 VOC'08 object classes is annotated with its object class label and a bounding box (among other annotations):

1. *class*: one of: aeroplane, bird, bicycle, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, sofa, train, tv/monitor.
2. *bounding box*: an axis-aligned bounding box surrounding the extent of the object visible in the image.

2.2 VOC'08 1K

Using Mechanical Turk, Rashtchian et al. (2010) collected five descriptions each for 1,000 VOC'08 images selected randomly but ensuring there were

Proceedings of the 15th European Workshop on Natural Language Generation (ENLG), pages 100–104,
Brighton, September 2015. ©2015 Association for Computational Linguistics



$\longrightarrow$ beside(person(Obj_1), person(Obj_2));
beside(person(Obj_2), dog(Obj_3));
in_front_of(dog(Obj_3), person(Obj_1))

Figure 1: Image from PASCAL VOC 2008 with annotations, and prepositions representing spatial relationships (objects numbered in descending order of size of area of bounding box).

50 images in each of the 20 VOC'08 object classes. Turkers had to have high hit rates and pass a language competence test before creating descriptions, leading to relatively high quality.

We obtained a set of candidate prepositions from the VOC'08 1K dataset as follows. We parsed the 5,000 descriptions with the Stanford Parser version 3.5.2[1] with the PCFG model, extracted the nmod:*prep* prepositional modifier relations, and manually removed the non-spatial ones. This gave us the following set of 38 prepositions:

> **V** = { *about, above, across, against, along, alongside, around, at, atop, behind, below, beneath, beside, beyond, by, close_to, far_from, in, in_front_of, inside, inside_of, near, next_to, on, on_top_of, opposite, outside, outside_of, over, past, through, toward, towards, under, underneath, up, upon, within* }

2.3 Human-Selected Spatial Prepositions

We are in the process of extending the VOC'08 annotations with human-selected spatial prepositions associated with pairs of objects in images. So far we have collected spatial prepositions for object pairs in images that have exactly two objects annotated (1,020). Annotators were presented with images from the dataset where in each image presentation the two objects, Obj_1 and Obj_2, were shown with their bounding boxes and labels. If there was more than one object of the same class, then the labels were shown with subscript indices (where objects are numbered in order of decreasing size of area of bounding box).

Next to the image was shown the template sentence "The *Obj*$_1$ is ___ the *Obj*$_2$", and the list of possible prepositions extracted from VOC 1K (see preceding section). The option 'NONE' was also available in case none of the prepositions was suitable (participants were discouraged from using it).

Each template sentence was presented twice, with the objects once in each order, "The *Obj*$_1$ is ___ the *Obj*$_2$" and "The *Obj*$_2$ is ___ the *Obj*$_1$".[2] Participants were asked to select all correct prepositions for each pair.

The following table shows occurrence counts for the 10 most frequent object labels:

person	dog	car	chair	horse	cat	bird	bicycle	motorbike	tv/monitor
783	123	112	92	92	88	86	79	77	63

Some prepositions were selected far more frequently than others; the top nine are:

next_to	beside	near	close_to	in_front_of	behind	on	on_top_of	underneath
304	211	156	149	141	129	115	103	90

3 Predicting Prepositions

When looking at a 2-D image, people infer all kinds of information not present in the pixel grid on the basis of their practice mapping 2-D information to 3-D spaces, and their real-world knowledge about the properties of different types of objects. In our research we are interested in the extent to which prepositions can be predicted without any real-world knowledge, using just features that can be computed from the objects' bounding boxes and labels. In this section we explore the predictive power of language and visual features within a Naive Bayes framework:

[1] http://nlp.stanford.edu/software/lex-parser.shtml

[2] Showing objects in both orders is necessary to capture non-reflexive prepositions such as *under, in, on* etc.

$$P(v_j|\mathbf{F}) \propto P(v_j)P(\mathbf{F}|v_j) \tag{1}$$

where $v_j \in \mathbf{V}$ are the possible prepositions, and $\mathbf{F}$ is the feature vector. Below we look at the predictive power of the prior model and the likelihood model as well as the complete model.

3.1 Prior Model

The prior model captures the probabilities of prepositions given ordered pairs of object labels L_s, L_o, where the normalised probabilities are obtained through a frequency count on the training set, using add-one smoothing. We then simply construe the model as a classifier to give us the most likely preposition v_{OL}:

$$v_{OL} = \begin{matrix} argmax \\ v \in \mathbf{V} \end{matrix} \; P(v_j|L_s, L_o) \tag{2}$$

where v_j is a preposition in the set of prepositions $\mathbf{V}$, and L_s and L_o are the object class labels of the first and second objects.

3.2 Likelihood Model

The likelihood model is based on a set of six geometric features computed from the image size and bounding boxes:

F_1: Area of Obj_1 (Bounding Box 1) normalized by Image size.
F_2: Area of Obj_2 (Bounding Box 2) normalized by Image Size.
F_3: Ratio of area of Obj_1 to area of Obj_2.
F_4: Distance between bounding box centroids normalized by object sizes.
F_5: Area of overlap of bounding boxes normalized by the smaller bounding box.
F_6: Position of Obj_1 relative to Obj_2.

F_1 to F_5 are real valued features, whereas F_6 is a categorical variable over four values (N, S, E, W). For each preposition, the probability distributions for each feature is estimated from the training set. The distributions for F_1 to F_4 are modelled with a Gaussian function, F_5 with a clipped polynomial function, and F_6 with a discrete distribution. The maximum likelihood model, which can also be derived from the naive Bayes model described in the next section by choosing a uniform $P(v)$ function, is given by:

$$v_{ML} = \begin{matrix} argmax \\ v \in \mathbf{V} \end{matrix} \; \prod_{i=1}^{6} P(F_i|v_j) \tag{3}$$

3.3 Naive Bayes Model

The naive Bayes classifier is derived from the maximum-a-posteriori Bayesian model, with the assumption that the features are conditionally independent. A direct application of Bayes' rule gives the classifier based on the posterior probability distribution as follows:

$$\begin{aligned} v_{NB} &= \begin{matrix} argmax \\ v \in \mathbf{V} \end{matrix} \; P(v_j|F_1, ...F_6, L_s, L_o) \\ &= \begin{matrix} argmax \\ v \in \mathbf{V} \end{matrix} \; P(v_j|L_s, L_o) \prod_{i=1}^{6} P(F_i|v_j) \end{aligned} \tag{4}$$

Intuitively, $P(v_j|L_s, L_o)$ weights the likelihood with the prior or *state of nature* probabilities.

4 Results

The current data set comprises 1,000 images, each labelled with one or more prepositions. The average prepositions per image over the whole dataset is 2.01. For training purposes, we create a separate training instance (Obj_s, Obj_o, v) for each preposition v selected by our human annotators for the given object pair Obj_s, Obj_o.

The models are evaluated with leave-one-out cross-validation, and two methods (Acc_A and Acc_B) of calculating accuracy (the percentage of instances for which a correct output is returned). The notation e.g. $Acc_A(1..n)$ is used to indicate that in this version of the evaluation method at least one of the top n most likely outputs (prepositions) returned by the model needs to match the (set of) human-selected reference preposition(s) for the model output to count as correct.

4.1 Accuracy method A

$Acc_A(1..n)$ returns the proportion of times that at least one of the top n prepositions returned by a model for an ordered object pair is in the complete set of human-selected prepositions for the same object pair. Acc_A can be seen as a system-level Precision measure. The table below shows $Acc_A(1)$ and $Acc_A(1..2)$ results for the three models:

Model	$Acc_A(1)$	$Acc_A^{Syn}(1)$	$Acc_A(1..2)$
v_{OL}	34.4%	43.9%	46.1%
v_{ML}	30.9%	35.6%	46.2%
v_{NB}	51.0%	57.2%	64.5%

Table 1: $Acc_B(1..n)$ for v_{NB} model and $n \leq 4$.

Preposition	$n = 1$	$n = 2$	$n = 3$	$n = 4$
next to	23.0%	77.0%	89.8%	93.1%
beside	58.3%	81.5%	85.8%	91.9%
near	43.6%	55.1%	74.4%	82.7%
close to	4.7%	14.8%	51.7%	87.9%
in front of	29.1%	39.7%	48.2%	52.5%
behind	31.0%	38.0%	50.4%	73.6%
on	72.2%	83.5%	85.2%	86.1%
on top of	10.7%	76.7%	81.6%	82.5%
underneath	53.3%	68.9%	84.4%	86.7%
beneath	15.5%	73.8%	79.8%	85.7%
far from	44.6%	62.2%	66.2%	68.9%
under	22.1%	27.9%	82.4%	83.8%
NONE	34.4%	53.1%	67.2%	73.4%
Mean	34.0%	57.9%	72.8%	80.7%
Mean Acc_B^{Syn}	50.9%	66.4%	77.9%	83.1%

In addition, the middle column above shows $Acc_A(1)$ results when sets of synonymous prepositions are considered identical. The synonym sets we chose for this purpose are: {*above, over*}, {*along, alongside*}, {*atop, upon, on, on_top_of*}, {*below, beneath*}, {*beside, by, next_to*}, {*beyond, past*}, {*close_to, near*}, {*in, inside, inside_of, within*} {*outside, outside_of*}, {*toward, towards*}, {*under, underneath*}.

4.2 Accuracy method B

$Acc_B(1..n)$ computes the mean of preposition-level accuracies. Accuracy for each preposition v is the proportion of times that v is returned as one of the top n prepositions out of those cases when v is in the human-selected set of reference prepositions. Acc_B can be seen as a preposition-level Recall measure.

Table 1 lists the $Acc_B(1..n)$ values for the v_{NB} model for each n up to 4; values are shown for the 13 most frequent prepositions (in order of frequency) and for the mean of all preposition-level accuracies. The last row shows the means for a version of Acc_B that takes synonyms into account as described in the last section.

5 Discussion

Looking at the naive Bayes results in Table 1, accuracy for some prepositions (e.g. *close to*) improves dramatically from $Acc_B(1)$ to $Acc_B(1..4)$. This implies that where the target preposition is not ranked first, it is often ranked second, third or fourth. There are synonym effects at work as shown by the Acc^{Syn} results; but there also is competition between prepositions that are not near synonyms, as shown by the fact that $Acc_A(1..2)$ results are better than $Acc_A^{Syn}(1)$ results.

For some prepositions, accuracy remains low even at n=4. This may reflect the general issue that human annotators use two different perspectives in selecting prepositions: (i) that of a viewer looking at the image, and (ii) that of one or both of the objects involved in the spatial relationship being described. Regarding (i), e.g. in the image in Figure 1, the dog is 'in front of' the person because it is between the viewer and the person. Regarding (ii), in other examples, a person can be 'in front of' a monitor, or one chair 'opposite' another, even when the viewer sees them both from the side.

The naive Bayes framework we have investigated here is a simple approach which is likely to be outperformed by more sophisticated ML methods. E.g. in calculating the likelihood term $P(F|v)$, our approach assumes the features to be independent; feature weighting per preposition was not carried out; and the data set is small relative to what we are using it for.

6 Conclusion

We have described (i) a dataset we are developing in which object pairs are annotated with prepositions that describe their spatial relationship, and (ii) methods for automatically predicting such prepositions on the basis of features computed from image and object geometry and object class labels. We have found that on the basis of language information (object class labels) alone we can predict prepositions with 34.4% accuracy, rising to 43.9% if we count near synonyms as correct. Using both language and visual information we can predict prepositions with 51% accuracy, rising to 57.2% with near synonyms. We have also found that where the target preposition is not ranked top, it is often ranked very near the top, as can be seen from the Acc_B results.

The next step in this research will be to increase our dataset and to apply machine learning methods such as support vector machines and neural networks to our learning task.

References

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *EMNLP'13*, pages 1292–1302.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.

Gaurav Kulkarni, Visruth Premraj, Vicente Ordonez, Sudipta Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2891–2903.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daum Iii. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of EACL'12*.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. 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.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 444–454. Association for Computational Linguistics.