Describing Spatial Relationships between Objects in Images in English and French

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

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

Maxime Aberton and Sami Benjelloun
INSA Rouen
Avenue de l’Université
76801 Saint-Étienne-du-Rouvray Cedex, France
{maxime.aberton, sami.benjelloun}@insa-rouen.fr

Abstract

The context for the work we report here is the automatic description of spatial relationships between pairs of objects in images. We investigate the task of selecting prepositions for such spatial relationships. We describe the two datasets of object pairs and prepositions we have created for English and French, and report results for predicting prepositions for object pairs in both of these languages, using two methods: (a) an existing approach which manually fixes the mapping from geometrical features to prepositions, and (b) a Naive Bayes classifier trained on the English and French datasets. For the latter we use features based on object class labels and geometrical measurements of object bounding boxes. We evaluate the automatically generated prepositions on unseen data in terms of accuracy against the human-selected prepositions.

1 Introduction

Automatic image description is important not just for assistive technology, but also for applications such as text-based querying of image databases. A good image description will, among other things, refer to the main objects in the image and the relationships between them. Two of the most important types of relationships for image description are activities (e.g. a child riding a bike), and spatial relationships (e.g. a dog in a car).

The task we investigate is predicting the prepositions that can be used to describe spatial relationships between pairs of objects in images. This is an important subtask in image description, 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 subtask is Elliott & Keller (2013) where the mapping is rule-based.

Spatial relations also play a role in referring expression generation (Viethen and Dale, 2008; Golland et al., 2010) where the problem is, however, often framed as a content selection problem from known abstract representations of the objects and scene, and the aim is to enable unique identification of the object referred to.

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 listing, 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 (Section 2.3). We did this in separate experiments for both English and French.

The overall aim is to create models for the mapping from image, bounding boxes and labels to spatial prepositions as indicated in Figure 1. We compare two approaches to modelling the mapping. One is taken from previous work (Elliott and Keller, 2013) and defines manually constructed rules to implement the mapping from image ge-
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).

beside(person(Obj1), person(Obj2));
−→ beside(person(Obj2), dog(Obj3));
in_front_of(dog(Obj3), person(Obj1))

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).

Table 1: Object class label frequencies.

| Object Class | Frequency |
|--------------|-----------|
| person       | 783       |
| dog          | 123       |
| car          | 112       |
| chair        | 92        |
| horse        | 92        |
| cat          | 88        |
| bird         | 86        |
| bicycle      | 79        |
| motorbike    | 77        |
| la moto      | 63        |
| l'écran      | 60        |
| l'avion      | 59        |
| la bouteille | 58        |
| le bateau    | 57        |
| le canapé    | 44        |
| le train     | 43        |
| la plante    | 33        |
| le mouton    | 27        |
| la vache     | 15        |
| la table     | 9         |

Table 1: Object class label frequencies.

Ometries to prepositions (Section 3.1). The other is a Naive Bayes classifier trained on a range of features to represent object pairs, computed from image, bounding boxes and labels (Section 3.2). We report results for English and French, in terms of two measures of accuracy (Section 5).

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 in one of the 20 VOC’08 object classes is annotated with six types of information of which we use the following three:

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.

3. **Occlusion**: a high level of occlusion is present.

Examples of all six types of annotation can be seen in Figure 2. We use the object class labels in predicting prepositions, and for the French experiments we translated them as follows (in the same order as the English labels above):

l’avion, l’oiseau, le vélo, le bateau, la bouteille, le bus, la voiture, le chat, la chaise, la vache, la table, le chien, le cheval, la moto, la personne, la plante, le mouton, le canapé, le train, l’écran

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 even distribution over the 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_E = \{ \text{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 } \}\n
For the list of French prepositions we started by compiling the list of possible translations of the English prepositions, after which we checked the list against 200 example images which resulted in a few additions and deletions. The final list for French has the following 21 prepositions (note there is no 1-to-1 correspondence with the English prepositions):

\[ V_F = \{ \text{à côté de, a l’intérieur de, a l’extérieur de, au dessus de, au niveau de, autour de, contre, dans, derrière, devant, en dessous de, en face de, en haut de, en travers de, le long de, loin de, par delà, parmi, près de, sous, sur } \}\n
\(^1\)http://nlp.stanford.edu/software/lex-parser.shtml

### 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, \( \text{Obj}_1 \) and \( \text{Obj}_2 \), 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 indices (numbered in order of decreasing size of bounding box).

#### 2.3.1 English data

Next to the image was shown the template sentence “The \( \text{Obj}_1 \) is \_ the \( \text{Obj}_2 \)”, and the list of possible prepositions extracted from VOC 1K (see last section). The option ‘NONE’ was also available in case none of the prepositions was suitable (but participants were discouraged from using it).

Table 1 shows occurrence counts for the 20 object class labels, while the two columns on the left of Table 2 show how many times each preposition was selected by the annotators in the English version of the experiment. The average number of prepositions per object pair chosen by the English annotators was 2.01.

Each pair of objects was presented twice, the template incorporating the objects once in each or-
Table 2: Number of times each preposition was selected by the English and French annotators.

| Preposition     | English | French |
|-----------------|---------|--------|
| next to         | 304     | à côté de |
| beside          | 211     | près de  |
| near            | 156     | devant  |
| close to        | 149     | contre  |
| in front of     | 141     | derrière |
| behind          | 129     | sur     |
| on top of       | 115     | au niveau de |
| underneath      | 90      | au dessus de |
| beneath         | 84      | en face de |
| far from        | 74      | en dessous de |
| under           | 68      | loin de  |
| NONE            | 64      | par delà |
| alongside       | 56      | le long de |
| by              | 50      | dans    |
| upon            | 44      | autour de |
| against         | 26      | en travers de |
| opposite        | 26      | à l’interieur de |
| beyond          | 20      | AUCUN   |
| atop            | 18      | à l’extérieur de |

Table 2: Number of times each preposition was selected by the English and French annotators.

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 image and the objects’ bounding boxes and class labels.

In this section we look at two methods for mapping language and visual image features to prepositions. Each takes as input an image in which two objects in the above object classes have been annotated with rectangular bounding boxes and object class labels, and returns as output preposition(s) that describe the spatial relationship between the two objects in the image.

3.1 Rule-based method

The rule-based method we examine is a direct implementation of the eight geometric relations defined in Visual Dependency Grammar (Elliott and Keller, 2013; Elliott, 2014). An overview is shown in Figure 3, for details see Elliott (2014, p. 13ff).

In order to implement these rules as a classifier, we pair each rule with the preposition referenced in it. In the case of surrounds, we use around instead. Two of the relations are problematic for us to implement, namely behind and in front of, because they make use of manual annotations that in fact encode whether one object is behind or in

---

2Showing objects in both orders is necessary for non-reflexive prepositions such as under, in, on, but also allows for other (unknown) factors that may influence preposition choice such as respective size of first and second object.
front of the other. We do not have this information available to us in our annotations.

What we do have is the ‘occluded’ flag (see list of VOC’08 annotations in Section 2.1 and Figure 2) which encodes whether the object tagged as occluded is partially hidden by another object. The problem is that the occluding object is not necessarily one of the two objects in the pair under consideration, i.e. the occluded object might be behind something else entirely. Nevertheless, the ‘occluded’ flag, in conjunction with bounding box overlap, gives us an angle on the definition of in front of (‘the Z-plane relationship is dominant’); we define the two problematic relations as follows:

- Y is tagged ‘occluded’ and the overlap between X and Y is more than 50% of the bounding box area of Y.
- X is tagged ‘occluded’ and the overlap between X and Y is more than 50% of the bounding box area of X.

In pseudocode, and for English, our implementation looks as follows (a is the centroid angle, P is the output list of prepositions, and ‘overlap’ is the area of the overlap between the bounding boxes of Object 1 and Object 2):

```plaintext
P = {}
if overlap is 100% of Obj2 then
    P = P ∪ {around}    \(> \text{Obj1 surrounds Obj2}\)
end if
if overlap > 50% of Obj1 then
    P = P ∪ {on}         \(> \text{Obj1 on Obj2}\)
end if
if Obj2 occluded and overlap > 50% of Obj2 then
    P = P ∪ {in front of} \(> \text{Obj1 in front of Obj2}\)
else if Obj1 occluded and overlap > 50% of Obj1 then
    P = P ∪ {behind}    \(> \text{Obj1 behind Obj2}\)
end if
if 225 < a < 315 then
    P = P ∪ {above}    \(> \text{Obj1 above Obj2}\)
else if 45 < a < 135 then
    P = P ∪ {below}    \(> \text{Obj1 below Obj2}\)
else if opposite conditions are met then
    P = P ∪ {opposite} \(> \text{Obj1 opposite Obj2}\)
else
    P = P ∪ {beside}    \(> \text{Obj1 beside Obj2}\)
end if
return P
```

This algorithm returns between 1 and 4 prepositions. The counts for multiple outputs are as follows (no different for English and French):

| | Returned in n cases |
|---|---|
| 1 | 580 |
| 2 | 159 |
| 3 | 247 |
| 4 | 14 |

For evaluating the rule-based classifier against the French human-selected prepositions we translated the eight English prepositions as follows (listed in the same order as in Figure 3):

- sur, autour de, à côté de, en face de,
- au dessus de, en dessous de, devant,
- derrière

### 3.2 Naive Bayes Classifier

Our second preposition selection method is a Naive Bayes Classifier. Below we describe how we model the prior and likelihood terms, before describing the whole model. The terms come together as follows under Naive Bayes:

\[
P(v_j|F) \propto P(v_j)P(F|v_j)
\] (1)
where \( v_j \in V \) are the possible prepositions, and \( F \) is the feature vector.

### 3.2.1 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.

In order to test this model separately, we simply construe it as a classifier to give us the most likely preposition \( v_{OL} \):

\[
v_{OL} = \arg\max_{v \in V} P(v_j | L_s, L_o)
\]

(2)

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

### 3.2.2 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.

For separate evaluation, a 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:

\[
v_{ML} = \arg\max_{v \in V} \prod_{i=1}^{6} P(F_i | v_j)
\]

(3)

### 3.2.3 Complete 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:

\[
v_{NB} = \arg\max_{v \in V} P(v_j | F_1, ..., F_6, L_s, L_o)
\]

\[
= \arg\max_{v \in V} P(v_j | L_s, L_o) \prod_{i=1}^{6} P(F_i | v_j)
\]

(4)

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

### 4 Evaluation Measures
We use two methods (\( Acc_A \) and \( Acc_B \)) of calculating accuracy (the percentage of instances for
| Preposition         | $v_{NB}$ |             | $v_{RB}$ |             | $Acc_B(1..n)$ | $Acc_{B}^{Syn}(1..n)$ | $Acc_B(1..n)$ | $Acc_{B}^{Syn}(1..n)$ |
|---------------------|----------|-------------|----------|-------------|--------------|------------------------|--------------|------------------------|
|                     | $n=1$    | $n=2$       | $n=3$    | $n=4$       | $n=1$        | $n=2$       | $n=3$    | $n=4$       | $n=1$        | $n=2$       | $n=3$    | $n=4$   |
| next to              | 23.0     | 77.0        | 89.8     | 93.1        | 73.7         | 94.7        | 70.1     | 76.3       | 78.7        | 78.7       | 100      | 100     |
| beside               | 58.3     | 81.5        | 85.8     | 91.9        | 75.8         | 96.2        | 11.3     | 22.0       | 26.2        | 26.2       | 10.6     | 26.2    |
| near                | 43.6     | 55.1        | 74.4     | 82.7        | 44.2         | 96.8        | 8.5      | 14.0       | 22.5        | 24.0       | 8.5      | 24.0    |
| close to             | 4.7      | 14.8        | 51.7     | 87.9        | 16.1         | 94.0        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |
| in front of          | 29.1     | 39.7        | 48.2     | 52.5        | 29.1         | 52.5        | 5.7      | 8.0        | 10.0        | 23.0       | 9.8      | 26.2    |
| behind              | 31.0     | 38.0        | 50.4     | 73.6        | 31.0         | 73.6        | 8.5      | 14.0       | 22.5        | 24.0       | 8.5      | 24.0    |
| on                  | 72.2     | 83.5        | 85.2     | 86.1        | 80.0         | 86.1        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |
| on top of            | 10.7     | 76.7        | 81.6     | 82.5        | 80.6         | 84.5        | 5.7      | 8.0        | 10.0        | 23.0       | 9.8      | 26.2    |
| underneath           | 53.3     | 68.9        | 84.4     | 86.7        | 68.9         | 90.0        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |
| beneath             | 15.5     | 73.8        | 79.8     | 85.7        | 15.5         | 85.7        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |
| far from            | 44.6     | 62.2        | 66.2     | 68.9        | 44.6         | 68.9        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |
| under               | 22.1     | 27.9        | 82.4     | 83.8        | 67.6         | 83.8        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |
| NONE                | 34.4     | 53.1        | 67.2     | 73.4        | 34.4         | 73.4        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |
| alongside           | 0.0      | 5.4         | 8.9      | 12.5        | 0.0          | 10.7        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |
| by                  | 4.0      | 8.0         | 10.0     | 38.0        | 72.0         | 86.0        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |
| upon                | 0.0      | 4.5         | 75.0     | 77.3        | 81.8         | 86.4        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |
| against             | 7.7      | 11.5        | 19.2     | 26.9        | 7.7          | 26.9        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |
| opposite            | 19.2     | 34.6        | 42.3     | 50.0        | 19.2         | 46.2        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |
| beyond              | 15.0     | 25.0        | 25.0     | 30.0        | 15.0         | 30.0        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |
| around              | 33.3     | 33.3        | 50.0     | 66.7        | 33.3         | 66.7        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |
| above               | 14.3     | 14.3        | 14.3     | 57.1        | 14.3         | 57.1        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |
| below               | 0.0      | 25.0        | 75.0     | 75.0        | 0.0          | 75.0        | 20.9     | 55.7       | 77.4        | 78.3       | 35.4     | 85.2    |

| Mean                | 24.3     | 41.6        | 57.6     | 67.4        | 41.1         | 71.2        | 24.5     | 33.7       | 40.4        | 40.7       | 31.8     | 46.0    |

Table 4: English $Acc_B$ results: $Acc_B(1..n)$, $n \leq 4$; $Acc_B^{Syn}(1)$; and $Acc_B^{Syn}(1..4)$ for $v_{NB}$ and $v_{RB}$ models. Shown: all prepositions of frequency 20 and above, in order of frequency. Also included are less frequent words if they are in the set of eight prepositions produced by the $v_{RB}$ method.

which a correct output is returned). The notation $Acc_A(1..n)$ or $Acc_A(1..n)$ 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 one of the human-selected reference prepositions for the model output to count as correct.

Furthermore, we use the notation $Acc_A^{Syn}(1..n)$ or $Acc_A^{Syn}(1..n)$ to indicate that in this version, at least one of the top $n$ most likely outputs (prepositions) returned by the model, or one of its near synonyms, needs to match one of the human-selected reference prepositions for the model output to count as correct.

The near synonym sets used for English 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, plus 11 singleton sets.

For French we used: a l’intérieur de, dans, au dessus de, en haut de, en dessous de, sous, plus 15 singleton sets. This gives us 18 sets for French, and 22 for English.

For the rule-based selection method we do not have the ranked outputs needed to compute $Acc_A$ and $Acc_B$. Interpreting the output set $P$ directly as ranked would mean preserving the order in which prepositions are selected by rules which is likely to be unfair to this method. Instead we randomly shuffle $P$ and then interpret it as ranked, with the first in this shuffled list giving the highest ranked output $v_{RB}$. To be on the safe side we average all results over 10 different random shuffles. Note that from $n = 4$ upwards, it makes no difference whether the outputs are truly ranked or not.
| Preposition  | \(v_{NB} \) | \(v_{RB} \) |
|--------------|-------------|-------------|
|              | \( Acc_B(1..n) \) | \( Acc_B^{syn}(1..n) \) | \( Acc_B(1..n) \) | \( Acc_B^{syn}(1..n) \) |
|              | \( n=1 \) | \( n=2 \) | \( n=3 \) | \( n=4 \) | \( n=1 \) | \( n=2 \) | \( n=3 \) | \( n=4 \) | \( n=1 \) | \( n=2 \) | \( n=3 \) | \( n=4 \) |
| à côté de    | 40.1  | 65.0  | 80.7  | 91.2  | 40.1  | 91.2  | 66.7  | 72.6  | 74.1  | 74.1  | 65.1  | 74.1  |
| près de      | 23.5  | 49.2  | 75.4  | 83.6  | 23.5  | 83.6  | 5.2   | 13.1  | 15.8  | 15.8  | 6.2   | 15.8  |
| devant       | 23.2  | 38.4  | 46.9  | 53.7  | 23.2  | 53.7  | 4.3   | 11.2  | 16.1  | 16.8  | 7.1   | 16.8  |
| contre       | 41.0  | 63.4  | 78.3  | 83.2  | 41.0  | 83.2  | 27.2  | 60.7  | 77.8  | 77.8  | 28.1  | 77.8  |
| derrière     | 16.1  | 29.8  | 46.0  | 70.8  | 16.1  | 70.8  | 16.7  | 53.7  | 16.1  | 16.8  | 7.1   | 16.8  |
| sur          | 53.0  | 70.9  | 85.5  | 88.9  | 53.0  | 88.9  | 19.5  | 63.7  | 78.3  | 83.2  | 53.0  | 88.9  |
| au niveau de | 28.2  | 59.1  | 71.8  | 78.2  | 28.2  | 78.2  | 38.6  | 70.3  | 81.1  | 81.1  | 38.6  | 70.1  |
| sous         | 78.9  | 90.5  | 90.5  | 92.6  | 89.5  | 95.8  | 30.1  | 43.7  | 48.6  | 48.6  | 59.4  | 100   |
| au dessus de | 19.5  | 56.1  | 62.2  | 69.5  | 19.5  | 69.5  | 24.4  | 39.2  | 52.1  | 52.4  | 28.2  | 52.4  |
| en face de   | 20.3  | 34.2  | 48.1  | 54.4  | 20.3  | 54.4  | 35.4  | 35.4  | 35.4  | 35.4  | 35.4  | 35.4  |
| en dessous de| 12.2  | 59.5  | 70.3  | 81.1  | 56.8  | 81.1  | 30.1  | 43.7  | 48.6  | 48.6  | 59.4  | 100   |
| loin de      | 38.6  | 56.1  | 63.2  | 66.7  | 38.6  | 66.7  | 24.4  | 42.3  | 57.1  | 57.1  | 23.6  | 57.1  |
| par delà     | 16.7  | 35.7  | 40.5  | 45.2  | 16.7  | 45.2  | 19.5  | 69.5  | 45.2  | 45.2  | 19.5  | 69.5  |
| le long de   | 7.5   | 20.0  | 22.5  | 22.5  | 7.5   | 22.5  | 7.5   | 22.5  | 7.5   | 22.5  | 7.5   | 22.5  |
| dans         | 56.5  | 78.3  | 82.6  | 91.3  | 56.5  | 91.3  | 29.8  | 42.9  | 57.1  | 57.1  | 28.6  | 57.1  |
| autour de    | 28.6  | 28.6  | 42.9  | 42.9  | 28.6  | 42.9  | 24.4  | 42.3  | 57.1  | 57.1  | 23.6  | 57.1  |
| en travers de| 28.6  | 42.9  | 50.0  | 57.1  | 28.5  | 57.1  | 28.6  | 42.9  | 50.0  | 57.1  | 28.6  | 57.1  |
| à l’intérieur de | 20.0 | 80.0  | 90.0  | 90.0  | 80.0  | 100   | 27.2  | 39.8  | 47.1  | 47.3  | 31.6  | 53.7  |

| Mean         | 30.7  | 53.2  | 63.7  | 70.1  | 37.1  | 70.9  | 27.2  | 39.8  | 47.1  | 47.3  | 31.6  | 53.7  |

Table 5: French \( Acc_B \) results: \( Acc_B(1..n) \), \( n \leq 4 \); \( Acc_B^{syn}(1) \); and \( Acc_B^{syn}(1..4) \) for \( v_{NB} \) and \( v_{RB} \) models. Shown: all prepositions of frequency 10 and above, in order of frequency. Also included are less frequent words if they are in the set of eight prepositions produced by the \( v_{RB} \) method.

**Accuracy measure 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 set of all human-selected prepositions for the same object pair. \( Acc_A \) can be seen as a system-level Precision measure.

**Accuracy measure 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 all cases where \( v \) is in the human-selected set of reference prepositions. \( Acc_B \) can be seen as a preposition-level Recall measure.

5 Results

The current French and English data sets each comprise 1,000 images/object-pair items, each of which is labelled with one or more prepositions. 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 context ‘The \( Obj_s \) is \( v \) the \( Obj_o \)’ (or the French equivalent). The models are trained and tested with leave-one-out cross-validation.

Table 3 shows English and French \( Acc_A \) and \( Acc_A^{syn} \) results for the rule-based method \( (v_{RB}) \), the prior model \( (v_{OL}) \), the likelihood model \( (v_{ML}) \), and the Naive Bayes model \( (v_{NB}) \). The main results are the \( Acc_A(1) \) results, because after all a method needs to select a single preposition in order to be usable, e.g. in image description.

\( Acc_A^{syn}(1) \) gives an idea of how much greater a proportion of a method’s outputs would be considered correct by human evaluators.

The remaining measures give various perspectives on the proportion of times a method came close to getting it right, for four degrees of ‘close’. E.g. \( Acc_A^{syn}(1..4) \) shows what proportion of times one of the top 4 prepositions generated by a method, or one of their near synonyms, was in the reference set.

It is clear that the English results are more affected by synonym effects. E.g. \( Acc_A(1..n) \) for English is nearly 10 percentage points lower than for French for all \( n \), whereas this difference all but
disappears for $\text{Acc}^{\text{Syn}}_{\text{A}}(1..n)$.

Overall, the $v_{\text{NB}}$ method always achieves the best result, as expected. The $v_{\text{ML}}$ model seems to be better at English than French, whereas for $v_{\text{OL}}$ it is the other way around.

Generally, once synonyms are taken into account, the results are strikingly similar for English and French, with the exception of the $V_{\text{ML}}$ model which does worse for French.

Tables 4 and 5 list the $\text{Acc}_{\text{R}}(1..n), n \leq 4$ and $\text{Acc}_{\text{A}}^{\text{Syn}}(1..n), n \in \{1, 4\}$ results for the $v_{\text{NB}}$ and $v_{\text{RB}}$ models; values are shown for the most frequent prepositions (in order of frequency) and for the mean of all preposition-level accuracies. We are not showing all prepositions partly for reasons of space, but also because for the low frequency prepositions, the models tend to underfit or overfit noticeably.

Note that here too we consider the $\text{Acc}_{\text{A}}(1)$ and $\text{Acc}_{\text{A}}^{\text{Syn}}(1)$ figures to be the main results. Among the English prepositions that $v_{\text{NB}}$ does well with (considered under the main $\text{Acc}_{\text{R}}(1)$ measure) are $\text{beside, near, underneath, far from}$, and results for $\text{on}$ are particularly good; $v_{\text{RB}}$ does well for $\text{beside}$.

As for French, $v_{\text{NB}}$ does well with $\text{à côté de, contre, sur, loin de}$, while results for $\text{sous}$ are particularly good. $v_{\text{RB}}$ does well for $\text{à côté de}$. Apart from $\text{near, underneath and contre}$, these are the same prepositions, semantically, as the English ones the methods do well with.

6 Conclusion

We have described (i) English and French datasets 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 (visual information) and from object class labels (language information).

The main method we tested, a Naive Bayes classifier which takes both language and vision information into account, does best in terms of all evaluation methods we used, and it does better on English than on French. When evaluated separately, the prior model which is based on language information only, outperforms the likelihood model which is based on visual information only, in terms of the main evaluation measures $\text{Acc}_{\text{A}}(1)$ and $\text{Acc}_{\text{A}}^{\text{Syn}}(1)$.

Main results in the region of 50% leave room for improvement; the fact that these go up to around 70% when the top 4 results are taken into account indicates that the method gets it nearly right a lot of the time and that for a smaller set of prepositions, and with more sophisticated machine learning methods, better results will be obtained.

It seems clear from the results, and intuitively obvious, that a greater presence of near synonyms in the data makes for a harder modelling task. We had a principled reason for using this particular set of English prepositions: it is the set observed in the human-authored descriptions we used (see Section 2.2). In our future work we will also work with the single best prepositions chosen by annotators to describe spatial relationships. This seems likely to result in a smaller list of prepositions overall and an easier modelling task. In order to get a truer impression of the quality of results we will also carry out human evaluation.

Acknowledgments

The research reported in this paper was supported by a Short-term Scientific Mission grant under European COST Action IC1307 (The European Network on Integrating Vision and Language).

References

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In Proceedings of the 15th Conference on Empirical Methods in Natural Language Processing (EMNLP’10), pages 410–419. Association for Computational Linguistics.

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 Daumé 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.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In Proceedings of the Fifth International Natural Language Generation Conference (INLG’08), pages 59–67. 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 16th Conference on Empirical Methods in Natural Language Processing (EMNLP ’11), pages 444–454. Association for Computational Linguistics.