Effect of Data Annotation, Feature Selection and Model Choice on Spatial Description Generation in French

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

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

Jessie Levacher **Julie Pain** **Adam Quinquenel**
INSA Rouen
Avenue de l'Université
76801 Saint-Étienne-du-Rouvray Cedex, France
{firstname.lastname}@insa-rouen.fr

Abstract

In this paper, we look at automatic generation of spatial descriptions in French, more particularly, selecting a spatial preposition for a pair of objects in an image. Our focus is on assessing the effect on accuracy of (i) increasing data set size, (ii) removing synonyms from the set of prepositions used for annotation, (iii) optimising feature sets, and (iv) training on best prepositions only vs. training on all acceptable prepositions. We describe a new data set where each object pair in each image is annotated with the best and all acceptable prepositions that describe the spatial relationship between the two objects. We report results for three new methods for this task, and find that the best, 75% Accuracy, is 25 points higher than our previous best result for this task.

1 Introduction

The research in this paper addresses the area of image description generation with applications in automatic image captioning and assistive technologies. An important aspect, and long-standing research topic, is to identify the entities, or objects, in images. However, a good image description will also say something about how entities relate to each other, not just list them. Spatial relations, and prepositions to express them, are particularly important in this context, but until very recently there had been no research directly aimed at this subtask, although some research came close (Mitchell et al., 2012; Kulkarni et al., 2013; Yang et al., 2011). Elliott & Keller (Elliott and Keller, 2013) did address the subtask, but with hardwired rules for just eight preposition. The work reported by Ramisa et al. (2015) is closely related to our work and also uses geometric and label features to predict prepositions.

2 Data

The new data set we have created for the experiments in this paper is a set of photographs in which objects in 20 classes are annotated with bounding boxes and class labels, and each object pair with prepositions that describe the spatial relationship between the objects. The data was derived from the VOC'08 data (Everingham et al., 2010) by selecting images with 2 or 3 bounding boxes, and adding the preposition annotations. The data has twice as many images as in our previous work (Belz et al., 2015), and a smaller set of prepositions (see below).

2.1 Annotation

For each object pair in each image, and for both orderings of the object labels, L_s, L_o and L_o, L_s, three French native speakers selected (i) the best preposition for the given pair (free text entry), and (ii) the possible prepositions for the given pair (from a given list) that accurately described the spatial relationship between the two objects in the pair. As a result, we have a total of 4,140 object pair annotations which fold out into 9,278 training instances.

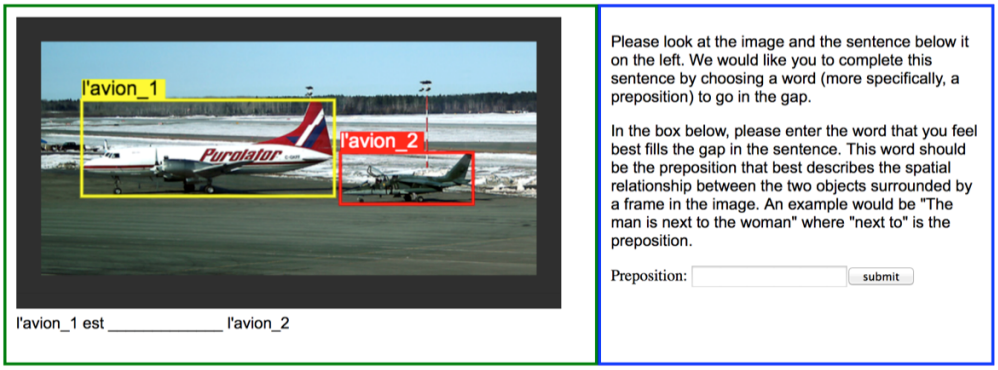

Figure 1: Screen grab of annotation tool, showing first task (free-text entry of single best preposition).

Figure 1 is a screen grab from our annotation tool showing the first annotation task (free-text entry of single best preposition). In the second task, annotators chose from the following set of 17 prepositions:

> *à côté de, a l'éxterieur de, au dessus de, au niveau de, autour de, contre, dans, derrière, devant, en face de, en travers de, le long de, loin de, par delà, près de, sous, sur.*

In our previous work with French data (Belz et al., 2015) we additionally had *en dessous de, en haut de, parmi* and *à l'interieur de*. We removed *parmi*, because it was never used in our previous annotation efforts, and the other three because the preposition set also contains near synonyms for them. Below, we refer to the data annotated with the smaller set as **DS-17** and that with the larger **DS-21**.

We previously used only images with exactly 2 object bounding boxes; these images are also included (newly annotated) in our new data set. In some of our experiments below we report results for just this subset and refer to it as **DS-17-2o**. The remaining half of the data (containing only images with 3 bounding boxes) is referred to as **DS-17-3o**.

We replaced the VC'08 object class labels with their French equivalents in the annotations, yielding the following set of words (used for the language features, see Section 3.1 below):

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

We used pairwise kappa to assess inter-annotator and intra-annotator agreement for our three annotators (who annotated one third of the data each). For selection of best prepositions this is straightforward; for all prepositions it is less straightforward, because the sets of selected prepositions differ in set size and overlap size. Our approach was to align the preposition sets and to pad out the aligned sets with blank labels if an annotator did not select a preposition selected by another annotator. Calculated in this way on a batch of 40 images, for *best* prepositions, average inter-annotator agreement was 0.67, and average intra-annotator agreement was 0.81. For *all* prepositions, average inter-annotator agreement was 0.63, and average intra-annotator agreement was 0.77.[1]

2.2 Object Class Label and Preposition Counts

The following table shows occurrence counts for the 12 most frequent object class labels in DS-17:

la personne	la voiture	le chien	la chaise	le cheval	le chat	l'oiseau	le bateau	l'écran	la plante	la bouteille	l'avion
3946	606	518	419	349	347	345	321	285	278	273	259

Some prepositions were selected far more frequently than others; the top 12 are:

près de	à côté de	devant	derrière	au niveau de	contre	sous	sur	loin de	en face de	au dessus de	le long de
2183	1483	1084	1031	926	704	434	380	372	271	123	78

3 Methods

The training data contains a separate training instance (L_s, L_o, p) for each preposition p selected

[1] These would have been even higher had it not been for one of the annotators who had much lower kappas than the others.

by human annotators for the template 'L_s est p L_o' (e.g. *le chien est devant la personne*), given an image in which (just) Obj_s and Obj_o are surrounded by bounding boxes labelled with object class labels L_s and L_o. All models are trained and tested with leave-one-out cross-validation.

3.1 Learning Methods

Naive Bayes Model (NB): We use a Naive Bayes model as in our previous work (Belz et al., 2015) which maps our set of language and visual features to prepositions (for details of all features see Section 3.1). The model uses the language features for defining the prior model and the visual features for defining the likelihood model.

SVM Model: Using the same features, we trained a multi-class SVM model employing one-versus-one classification.[2] This involves training $k(k-1)/2$ pairs of binary preposition classifiers for a multi-class prediction task involving k prepositions. The SVM model was trained with an RBF kernel, characterised by a coefficient of $1/(|features|)$.

Decision-Tree Model (DT): Again using the same features, we created a multi-class probabilistic decision-tree model[2] with a maximum tree depth of 4 for the DS-17 data set, and 5 for the DS-21 data set (from training and validation error plots).

Logistic Regression Model (LR): Using the same features, we trained a multi-class logistic regression model employing one-versus-rest classification[1]. The model makes use of L1-norm regularisation with an inverse regularisation strength of 0.9.

3.2 Evaluation methods

To compare results in this paper, we use variants of Accuracy from our previous work (Belz et al., 2015). The dimension along which the variants we use here differ is output rank. Different variants, denoted $Acc(n)$, where $n = 1...4$, return Accuracy rates for the top n outputs produced by systems, such that a system output is considered correct if a target (human-selected) output is among the top n outputs produced by the system (so for $s = 1$ the measure is just standard Accuracy).

[2]Implemented using scikit-learn (http://scikit-learn.org).

	DS-21	DS-17-2o
	$Acc(1)$	$Acc(1)$
NB	50.2	67.0
DT	50.4	66.2
SVM	46.1	59.4
LR	**53.4**	**72.7**

Table 1: $Acc(1)$ results for the data with the larger (DS-21) and smaller (DS-17-2o) preposition sets, for all 4 models.

3.3 Features

The four methods described in the following section all use the following feature set (described in more detail in Belz et al., 2015):

- $F0$: Object label L_s.
- $F1$: Object label L_o.
- $F2$: Area of bounding box of Obj_s normalised by image size.
- $F3$: Area of bounding box of Obj_o normalised by image size.
- $F4$: Ratio of Obj_s bounding box area to that of Obj_o.
- $F5$: Distance between bounding box centroids.
- $F6$: Area of overlap of bounding boxes normalised by the smaller bounding box.
- $F7$: Distance between centroids divided by approximated average width of bounding boxes.
- $F8$: Position of Obj_s relative to Obj_o (N, E, S, W).

Note that to make the categorial features (F0, F1, F8) work for the logistic regression model we map them to 1-hot encodings (n bits for n feature values).

4 Experiments and Results

4.1 Preposition Set

In this set of experiments, we wanted to see what the effect on learning is of removing synonyms from the set of prepositions and re-annotating the data with the reduced set. We compared results for our previous French data (DS-21) with the corresponding subset of our new data (DS-17-2o), both with similar numbers of training instances. Note that because the annotations differ, we are testing on slightly different sets of target outputs. Table 1 shows the Accuracy results for the four models from Section 3.1.

Numbers clearly demonstrate a very substantial benefit from removing synonyms for all tested methods, improvement ranging from 13.3 points to 19.3. The benefit is biggest for LR, smallest for SVM.

	DS-17-2o		DS-17	
	$Acc(1)$	$Acc(2)$	$Acc(1)$	$Acc(2)$
NB	67.0	82.0	64.7	80.9
DT	66.2	80.5	67.7	81.4
SVM	59.4	78.5	-	-
LR	**72.7**	**86.8**	**74.9**	**89.2**

Table 2: $Acc(1)$ and $Acc(2)$ results for the smaller (DS-17-2o) and larger (DS-17) data sets with 17 prepositions.

4.2 Data Set Size

Here we look at the effect of adding more data to the training set, comparing results for DS-17-2o (1,020 images; 4,426 training instances) with results for the whole of DS-17 (2,070 images; 9,278 training instances). Table 2 shows the results: there are some improvements from the size increase for all methods except NB, but the only sizeable one is for LR.

4.3 Different Models

Tables 1 and 2 provide an overview of results for the four models above on DS-21, DS-17-2o and DS-17. Of the new methods (SVM, DT, LR), SVM does much worse than the others (we therefore leave it out of the remaining experiments below). The LR model achieves the best results across all data sets.

Looking at Acc(1) vs. Acc(2) results (Table 2), differences are very similar (around 14-15 points) for all methods except for SVM for which it is much bigger, implying that SVM more often has a target preposition in second place.

4.4 Feature Optimisation

We start with the results on DS-17 for the three best models as a baseline and try to improve over them using greedy lasso as a simple feature optimisation method which starts by selecting the single best feature and then keeps adding the next feature that achieves the best result in combination with previously selected feature(s). Table 3 shows $Acc(1)$, $Acc(2)$ and $Acc(3)$ results for DS-17, before and after feature optimisation. Feature optimisation does not make a difference to LR, but improves the results for DT slightly, and for NB substantially, by leaving out features 5, 6 and 8, and 6 and 7, respectively.

4.5 Best vs. All Annotations

Unlike in our previous work, our new data contains information about which preposition annotators thought was best out of the ones they considered possible (see Section 2), so we can now compare results for training on best prepositions only vs. all possible prepositions for object pairs.

There are more than twice the number of training instances for all possible prepositions (9,278) than for best prepositions only (4,140), so it is not a like-for-like comparison. We therefore also report (under the heading 'all-sub' in Table 4) results for a randomly selected subset of the all-prepositions data of the same size as the best-prepositions-only data (averaged over 4 different runs).

The results in Table 4 show very clearly the benefit of training on all possible prepositions compared to best only, although the benefit is less marked for the NB method. While results for 'all-sub' are lower than for 'all', and some of the improvement in the 'all' results is likely due to larger data set size, the 'all-sub' results nevertheless show clearly that the largest part of the improvement is due to training on all possible prepositions (that being the only difference between the 'best' and 'all-sub' data).

5 Discussion

It is worth recalling that the task we are trying to solve is to guess the actual 3D spatial relationship between two objects in a photograph, from just the object types and various geometric properties of the objects' bounding boxes which give just a rough idea even of the object's size and 2D dimensions in the image. Nevertheless this rudimentary information is enough to predict a correct 3D preposition 75% of the time in the case of our best method, LR, moreover across a variety of large and small, animate and inanimate objects, in indoors and outdoors scenes. The most closely related existing work (Ramisa et al., 2015) reported slightly higher accuracy rates, but for different data sets. Our own previous results (Belz et al., 2015) were considerably worse at around 50%.

The Acc(n) results for $n > 1$ are interesting. E.g. LR places a target preposition in the top two almost 90% of the time. At the same time, our annotators chose on average 2.2 prepositions per (ordered) object pair, with a kappa agreement of 0.63, indicating that there may be more than two good prepositions for an object pair. In future work we will evaluate

	DS-17			DS-17, optimised			
	$Acc(1)$	$Acc(2)$	$Acc(3)$	$Acc(1)$	$Acc(2)$	$Acc(3)$	Best feature set
DT	67.7	81.4	91.0	68.4	82.3	90.7	{0,1,2,3,4,7}
NB	64.7	80.9	90.4	71.6	86.3	93.1	{0,1,2,3,4,5,8}
LR	**74.9**	89.2	94.2	**74.9**	89.2	94.2	{0,1,2,3,4,5,6,7,8}

Table 3: $Acc(1)$, $Acc(2)$ and $Acc(3)$ results for DS-17, before and after feature optimisation, for the three best models.

	DS-17								
	best			all			all-sub		
	$Acc(1)$	$Acc(2)$	$Acc(3)$	$Acc(1)$	$Acc(2)$	$Acc(3)$	$Acc(1)$	$Acc(2)$	$Acc(3)$
DT	51.6	71.8	83.1	67.7	81.4	91.0	64.7	80.9	88.8
NB	57.6	74.8	84.0	64.7	80.9	90.4	61.2	78.8	88.3
LR	**59.3**	78.8	88.8	**74.9**	89.2	94.2	**73.6**	88.4	93.9

Table 4: $Acc(1)$ and $Acc(2)$ results for DS-17, using only best prepositions ('best'), using all prepositions ('all'), and using all prepositions but only a randomly selected subset ('all-sub') of instances from 'all' of size equal to that of the best preposition data.

the acceptability by human evaluators of the top n results. If it turns out, as seems likely, that the top two prepositions are acceptable to human evaluators, then the real accuracy would be closer to 90%.

6 Conclusion

In this paper, we have reported new results for automatic generation of spatial descriptions in French. We described a new data set where object pairs in images are annotated with the best preposition, as well as all possible prepositions, that describe the spatial relationship between the objects. We reported results for three new methods for this task, and found that (i) increasing the size of the data set on its own only has a small beneficial effect on results; (ii) removing synonyms from the annotations results in dramatically improved results for all methods tested, and (iii) training on all possible prepositions for an object pair instead of training on the single best preposition only is of substantial benefit for all methods tested. The best result for our task was achieved with the LR classifier, on the preposition set without synonyms, using all possible prepositions for object pairs. That result, 75% Accuracy, is an entire 25 points higher than our previous best result for this task.

Acknowledgments

This work originated during a Short-term Scientific Mission under European COST Action IC1307 (European Network on Integrating Vision & Language).

We are indebted to the INLG'06 reviewers for comprehensive and helpful comments; we have implemented as many as we could in the available time.

References

A. Belz, A. Muscat, M. Aberton, and S. Benjelloun. 2015. Describing spatial relationships between objects in images in English and French. In *Proceedings of VL'15*.

D. Elliott and F. Keller. 2013. Image description using visual dependency representations. In *Proceedings of EMNLP'13*, pages 1292–1302.

M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. 2010. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338.

G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2891–2903.

M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of EACL'12*.

A. Ramisa, J. Wang, Y. Lu, E. Dellandrea, F. Moreno-Noguer, and R. Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of EMNLP'15*, pages 214–220.

Y. Yang, C. Teo, H. Daumé III, and Y. Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of EMNLP'11*, pages 444–454.