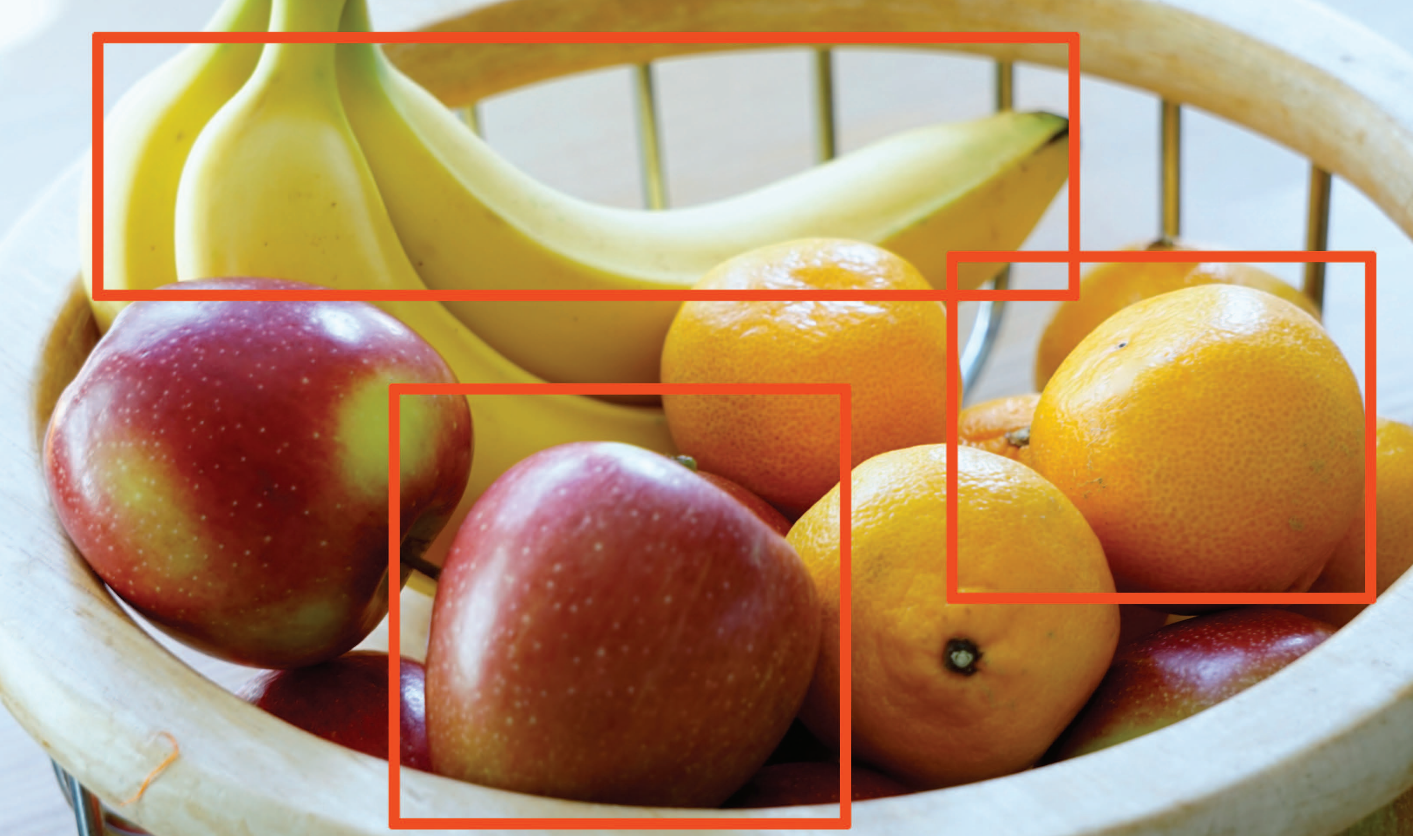


©ISTOCKPHOTO.COM/T_DOWNES

Learning to Generate Descriptions of Visual Data Anchored in Spatial Relations

Adrian Muscat
Department of Communications and Computer Engineering
University of Malta, Msida, Malta

Anja Belz
School of Computing, Engineering and Mathematics
University of Brighton, Brighton, UK

Abstract—The explosive growth of visual data both online and offline in private and public repositories has led to urgent requirements for better ways to index, search, retrieve, process and manage visual content. Automatic methods for generating image descriptions can help with all these tasks, and also play an important role in assistive technology for the visually impaired. The task we address in this paper is the automatic generation of image descriptions that are anchored in spatial relations. We construe this as a three-step task where the first step is to identify objects in an image, the second step detects spatial relations between object pairs on the basis of language and visual features; and in the third step, the spatial relations are mapped to natural language (NL) descriptions. We describe the data we have created, and compare a range of machine learning methods in terms of the success with which they learn the mapping from features to spatial relations, using

Digital Object Identifier 10.1109/MCI.2017.2708559
Date of publication: 19 July 2017

Corresponding author: Anja Belz (Email: a.s.belz@brighton.ac.uk).

automatic and human-assessed evaluations. We find that a random forest model performs best by a substantial margin. We examine aspects of our approach in more detail, including data annotation and choice of features. We describe six alternative natural language generation (NLG) strategies, and evaluate the generated NL strings using measures of correctness, naturalness and completeness. Finally, we discuss evaluation issues, including the importance of extrinsic context in data creation and evaluation design.

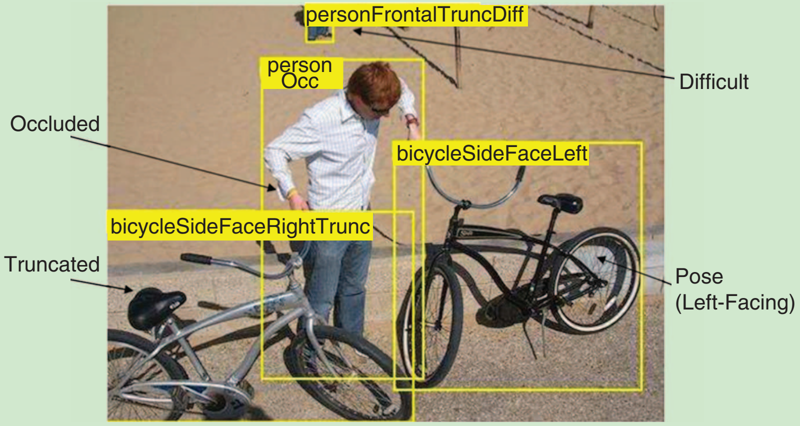
I. Introduction

The motivation for the research presented in this paper is two-fold. On the one hand, there is the now routinely cited explosion of data that is characteristic of the information age: visual data—one of the three main big data categories, along with textual and numerical data—continues to proliferate at an enormous rate, online as well as in privately and publicly held offline repositories. This has led to urgent requirements for better ways to index, search, retrieve, process and manage visual content. On the other hand, substantial proportions of the population are excluded from visual online content and the information age more generally [1], non-observance of accessibility requirements making the internet a frustrating experience for visually impaired people [2]. Blindness and partial sight are increasing, due to changing demographics and greater incidence of diseases such as diabetes, at great financial and human cost (WHO, [3]). Automatic image description plays a role in indexing, search, retrieval and management of visual data as well as in making visual data accessible to the visually impaired, used on the fly or offline e.g. as part of fulfilling accessibility requirements.

Taking even the most cursory look at human image descriptions, it is clear that humans prioritize mention of foregrounded and/or relatively large entities such as people, animals, cars, buildings, etc., and their attributes (color, size, etc.). Similarly important are relationships linking these entities, to each other and to their surroundings, including relationships with a temporal dimension (*a boy* ***riding*** *a bicycle*; *a dog* ***swimming in*** *a lake*), and relationships without (*a boy* ***on*** *a bike*; *a dog* ***in*** *a lake*). While the truth content of descriptions such as these with respect to an image is relatively easy to determine, this is frequently not the case for human-authored image descriptions. E.g., looking at Figure 1, strictly speaking, only the fourth description does not involve any conjecture beyond the evidence in the image.

In the work reported in this paper we focus on entities and spatial relations as aspects of image description that require minimal conjecture. The questions we wish to address are (1) to what level of accuracy can spatial relations be determined by machine learning (ML) methods; (2) to what extent do human authors agree when determining spatial relations from still images; and (3) what level of quality can be achieved with image descriptions anchored in automatically detected spatial relations. The first and third questions are important because we need good performance for descriptions to be practically useful; the second because annotations provide a poor basis for ML if agreement is low.

We construe automatic generation of image descriptions anchored in spatial relations as a three-step task: (1) entity identification (which we do not address); (2) identifying the spatial relations between pairs of entities on the basis of geometric and language features (Sections V to VII); and (3) generating natural language (NL) descriptions from sets of spatial relations (Section IX). The image data we use is described in Section IV-A, the annotations we collected for them in Section IV-B and IV-C. Recognizing the importance of human evaluation [5], we report two such sets of results (Sections VII-C and IX-F) as well as automatic metric scores. We discuss the role of application purpose in evaluations in some depth in the discussion section (Section X) which brings the paper to a close.



A main holds two bikes near a beach.
A young man wearing a striped shirt is holding two bicycles.
Man with two bicycles at the beach, looking perplexed.
Red haired man holding two bicycles.
Young redheaded man holding two bicycles near beach.

FIGURE 1 Image 2008_008320 from PASCAL VOC 2008 with annotations and image descriptions obtained by Rashtchian et al. [4] (original spelling errors).

II. Language and Space

Psycholinguistics and cognitive linguistics research on language and space has a long history including influential work by Leech [6], Bennett [7], Jackendoff [8], Talmy [9], Herskovits [10] and others who have studied how language marks spatial distinctions and patterns, and ascribes structure to space. In this section we relate our research to some of this literature, in order to provide a framework for conceptualizing and reporting our experiments. A detailed review of spatial preposition models is provided by Kelleher and Costello [11].

It is generally agreed that speakers do not take into account a scene's full, complex details, but instead access radically simplified representations. A process known as *schematization* "reduces a physical scene with all its richness of detail, to a sparse and sketchy semantic content" [10, p. 169]. For Herskovits [10], [12], schematization is a set of object geometry selection functions, e.g. idealizing objects to a point, line, plane, etc., or selecting parts of objects. Importantly, this reduction in complexity is a matter of degree and depends on factors such as distance between objects (e.g. objects are not always, as assumed e.g. by others such as Landau and Jackendoff [13], reduced to points). Other spatial representations accessed by language include[1] the 2D plane of view and axis-based componential representations of objects.

In order "[t]o compute a spatial relation between two objects, one must: 1. Configure them together—that is, select them for attention in a way that makes it possible to apprehend the applicable spatial relations; and 2. Categorize the configuration. […] Several lexically expressible relations can generally categorize a given configuration of two objects" [10, p. 199].

The literature is not very explicit about how linguistic spatial relations formally map to prepositions. Herskovits is clear [10, p. 160ff.] that prepositions have multiple senses each of which has multiple use types and is moreover subject to fuzzy interpretation due to convention-based sense shifting and pragmatic processes of tolerance [12, pp. 78ff. and 86ff.]. However, is there a single linguistic spatial relation for every sense and use type? Are they in some way parametrized to account for fuzzy interpretation? It seems clear at least that the set of configurations that match a preposition's one sense will not be identical to that of another.

From Scene to Description: Building on Herskovits's framework, we construe the processes underlying spatial image description to be as follows (schematization and categorization as above):

> [scene as it is] — *schematize*→ [schematic configuration with two object sets O_t, O_l] — *categorize*→ [spatial relation $p^s_u(O_t, O_l)$] — *realize*→ [description with matching syntactic structure including a prepositional phrase headed by a preposition denoting p]

Schematization is parametrized by linguistic goal and is language specific among other things [10, p. 159]; it produces a single configuration. Categorization and realization can produce multiple outputs. We understand individual location prepositions to denote multiple ***spatial relations*** $p^s_u(Object_{\text{trajector}}, Object_{\text{landmark}})$ corresponding to senses s and use types u. We adopt the trajectory and landmark terms [14] in preference to the cognitive linguistics literature's Figure and Ground as the more general terms.[2]

It seems clear that it is possible to systematically avoid generating unsuitable preposition senses and use types only if that distinction is marked at the spatial relation level. Having prepositions denote single spatial relations (as Herskovits appears to do when she refers (p. 160) to "*the* spatial relation denoted by the preposition", our italics) makes it impossible to systematically avoid generating descriptions that can only be understood as a different sense or use type inconsistent with the observed scene.

Domain-specific Spatial Relation Sets: In some contexts there may not be sufficiently fine-grained information available in the scene configuration to distinguish different preposition senses and use types. We deem this to be the case if two spatial relations (SR) consistently co-occur in SR sets selected by annotators (Section IV). In that case we consider them, in this context only, and not in any strict lexical sense, ***synonymous***, which in turn we interpret as license to merge them into one SR. In the present context, for reasons of practicality we assume that a preposition's senses and use types are all synonymous in this sense, as distinguishing even just senses would make data annotation (Section IV) a specialist task and training data much harder to obtain. It would also increase data sparsity. In fact, we go one step further (Section IV-B) and in some special cases consider spatial relations that are denoted by different prepositions synonymous in the above sense. However, in those cases we do justify the merge with co-occurrence statistics from data.

Automatic Image Description: In our work we start from a still image in which the plane of view is fixed. Some of our component techniques could be interpreted as combining to correspond to schematization: (1) object identification yields an abstraction of the objects as rectangles (the bounding boxes); (2) geometric feature computation (a) yields further abstractions of objects e.g. as points (centroids of bounding boxes), and (b) makes available other spatial information such as distance between objects; and (3) scenes are 'configured' into pairs of objects where one is the trajector and the other the landmark. Our methods categorize pair configurations into spatial relations, before generating an NL description for them.

III. Related Research in Image Description

The aim of image labelling (or tagging, or indexing) is to identify regions in an image that are meaningful to a human observer, and to attach labels to them that capture that meaning in some way. Image labelling goes back at least to the 1960s (for an overview of early image labelling work see Rosenfeld and Azriel [15]). A simple form of description can be generated from such region labels, but it would not be much more than a list. A step further is recent work on visual relationship detection [16]–[18] where relations between objects are identified in addition to the objects themselves.

Image description proper starts where a summarizing description of the whole image is aimed for, which involves prioritizing more important elements and relationships in a fully realized NL description of the input image. A fairly basic division is between (i) methods that create a new description for a given image from scratch, and (ii) methods that measure the similarity of a new image with other images for which descriptions exist, and then use one or more of those descriptions to create a description for the new image. See [19]–[23] for examples of the latter; our focus here is on methods of the

[1] For full list see Herskovits, 1997, p. 193.

[2] In expressions about spatial relationships, "typically one entity is taken as a reference point (or area) with respect to which the other is located" [14, p. 648]; the former is the *landmark* and the latter is the *trajector*. E.g. in *She is on the balcony*, she is the trajector and the balcony is the landmark.

former type. A recent survey by Bernardi et al. [24] reviews both types in detail.

Methods that create a new description for a given image from scratch can be said to have the main three component steps mentioned in the introduction: (1) identification of type and, optionally, location of objects and background/scene in the image; (2) detection of attributes, relations and activities involving objects from Step 1; and (3) generation of a word string from a representation of the output from Steps 1 and 2. For Step 1, some systems identify labelled regions [25], [26], others directly map images to words [27]. For Step 2, systems determine object attributes [26], [28], spatial relationships [29]–[31], activities [26], [30], etc. In Step 3, systems differ in the amount of linguistic knowledge they bring to bear on the generation process. Some view the task as similar to a linearization problem where the aim is to work out a likely string of words containing the labels, relations and attributes from Steps 1 and 2 [27], [32]; others employ templates to slot the latter into [29], [30], while still others use grammar-based techniques to construct descriptions [33], [34].

Identifying the spatial relationships between pairs of objects in images is an important part of image description, but is rarely addressed as a separate subtask in its own right. If a method produces spatial prepositions, it tends to be as a side-effect of the overall method [33], [35], or else relationships are not between objects, but e.g. between objects and the scene [29]. An example of preposition selection as a separate subtask is Elliott & Keller's work [30] who base the mapping on manually composed rules. Spatial relations also play a role in referring expression generation [36], [37] where the problem is, however, often simplified as a content selection task from known symbolic representations of objects and scene.

Mostly closely related to our work for Step 2 is work by Ramisa et al., 2015 [38] and Hürlimann & Bos, 2016 [39]. In both, various visual and verbal features computed for a given image are used to predict prepositions to describe the spatial relations between a pair of objects in the image. We have ourselves previously reported work on predicting English [40], [41] and French [31], [42] prepositions.

IV. Image Data and Annotations

Our main data source is the VOC'08 corpus of images [43] in which objects have been annotated with rectangular bounding boxes and object class labels. We collected additional annotations for images (Section IV-C) which list, for each object pair, a set of prepositions selected by human annotators as correctly describing the spatial relationship between the objects.

A. Source Data Sets

VOC'08 9K: The data from the PASCAL VOC 2008 Shared Task Competition (VOC'08) consists of 8,776 images and 20,739 objects in 20 object classes. In each image, every object belonging to one of the 20 VOC'08 object classes is annotated for class, bounding box, viewpoint, truncation, occlusion, and identification difficulty [43], examples of all of which can be seen in Figure 1.[3] Of these annotations we use the following:

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

VOC'08 1K: Using Mechanical Turk, Rashtchian et al. [4] collected five descriptions each for 1,000 VOC'08 images selected from the larger 9K set (see above) randomly but ensuring there were 50 images from each VOC'08 class. Contributors had to have high hit rates and pass a language competence test before creating descriptions, leading to relatively high quality with few grammatical or spelling mistakes. See Figure 1 for an example.

B. Spatial Relations for Annotation

In order to determine the set of spatial relations (SRs) to be used by our annotators, we proceeded as follows. From the VOC'08 1K data set we obtained a set of candidate prepositions by parsing the 5,000 descriptions with the Stanford Parser version 3.5.2[4] with the PCFG model, extracting the nmod:*prep* prepositional modifier relations, and manually removing the non-spatial ones. This gave us a set of 38 English prepositions.

In order to obtain an analogous set of prepositions for French, as a first step we asked two French native speakers to compile the list of possible translations of the English prepositions, and to check these against 200 random example images from our corpus. The full list for French had 21 prepositions and these were reduced to a smaller set, on the basis of an earlier batch of annotations [42], by eliminating (i) prepositions that were used fewer than three times by annotators (*en haut de, parmi*), and (ii) those which co-occur with another preposition more than 60%[5] of the times they occur in total (*á l'interieur de, en dessous de*), in accordance with the general sense of synonymity defined in Section II. We found this kind of co-occurrence to be highly imbalanced, e.g. the likelihood of seeing *á l'interieur de* given *dans* is 0.43, whereas the likelihood of seeing *dans* given *á l'interieur de* is 0.91. We take this as justification for merging *á l'interieur de* into *dans*, rather than the other way around. The whole process leaves a set of 17 French prepositions:

$\mathbf{V_F}$ = {*à côté de, á 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*}

As discussed in Section II, we make the domain-specific assumption that there is a one-to-one correspondence between prepositions and the SRs they denote. While our machine learning task is SR detection, we ask annotators to annotate our data with the corresponding prepositions (a more human-friendly task).

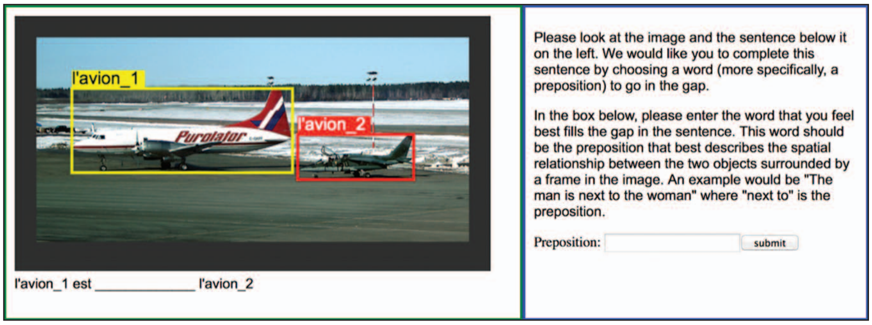
[3] Image adapted from: http://lear.inrialpes.fr/RecogWorkshop08/documents/everingham.pdf

[4] http://nlp.stanford.edu/software/lex-parser.shtml#Download

[5] This is a very high threshold and far above co-occurrence percentages for any other preposition pairs.

C. Annotation

For our annotation experiments, we selected all images with two and three objects in bounding boxes from the VOC'08 data, giving a set of 1,554 images (about 18% of the corpus). For each object pair O_i and O_j in each image, and for both orderings of the object labels, L_i, L_j and L_j, L_i, the task for annotators was to select (i) the single best preposition for the given pair, and (ii) the possible prepositions for the given pair (selected from a given list) that accurately described the relationship between the two objects in the pair. Figure 2 is a screen grab from our annotation tool showing the first annotation task (free-text entry of single best preposition).



l'avion_1 est ____________ l'avion_2

Please look at the image and the sentence below it on the left. We would like you to complete this sentence by choosing a word (more specifically, a preposition) to go in the gap.

In the box below, please enter the word that you feel best fills the gap in the sentence. This word should be the preposition that best describes the spatial relationship between the two objects surrounded by a frame in the image. An example would be "The man is next to the woman" where "next to" is the preposition.

Preposition:

FIGURE 2 Screen grab of annotation tool, showing free-text entry of single best preposition task (displaying image 2008_008025 from Pascal VOC 2008).

For the French annotation interface, we replaced the English VOC'08 object class labels (Section IV-A) with their French equivalents (used also for language features, Section V). Even though in the first annotation task, annotators were not limited to the prepositions shown in the second task, they did not use any others (a few typos we corrected manually). As it would have been virtually impossible to remember the exact list of prepositions and only use those, we interpret this as meaning that annotators did not feel other prepositions were needed.

We used pairwise kappa to assess inter-annotator and intra-annotator agreement. 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 size of set overlap. 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 *single best* prepositions (annotation task 1), average inter-annotator agreement was 0.67, and average intra-annotator agreement was 0.81. For *all possible* prepositions (annotation task 2), average inter-annotator agreement was 0.63, and average intra-annotator agreement was 0.77. These would have been higher if one of the annotators had not had much lower kappas than the others, an issue we return to in Section VIII-B.

D. Different Versions of Data Used in Experiments

Our complete data set has 1,020 images with two objects (2,040 ordered object pairs), and 534 images with three objects (3,204 ordered object pairs), from which we use a total of 5,240 ordered object pairs. For each ordered object pair Obj_s, Obj_o we have the set of prepositions $p_i^{s,o}$ selected for it by the annotators. Each such triple $(Obj_s, Obj_o, p_i^{s,o})$ becomes an individual data set instance. There are altogether 11,291 data set instances, corresponding to 2.2 prepositions per ordered pair. For ease of reference, Table 1 provides an overview of the different versions of the French data we use in the experiments reported in this paper, along with the shorthand names used for each in the results tables.

V. Machine Learning Methods

In this section, we describe the different machine learning (ML) models we test (Section V-C), the features we use to train the models (Section V-A; Table 2), and the set-up within which we tune model hyper-parameters, train the models, and investigate subsets of features (Section V-B).

A. Features

The ML methods described in the following section all use the feature set shown in Table 2. F0, F1, F15 and F16 are language features. F0 is the class label of the first object, F1 of the second (e.g. *person*). F15 and F16 are GloVe word vectors [44] for the labels each of length 50.[6] F2–F14 are visual features measuring various aspects of the geometries of the bounding boxes (BBs). Most features express a property of just one of the objects, but F4–F9 express a property of both objects jointly, e.g. F6 is the normalized BB overlap.

B. Machine Learning Set-up

Several of our ML methods (see Section V-C) have hyper-parameters (HPs) that need to be tuned in order to get good performance out of them, via some form of hyper-parameter optimization (HPO). At the same time, we wanted to investigate the role played by the different features in a feature

TABLE 1 Different versions of data set used in experiments (SR = spatial relation; *n* training instances).

NAME	DESCRIPTION	*n*
DS-F	COMPLETE DATA SET, BEST/ALL SR ANNOTATIONS	11,291
DS-F-BEST	DS-F WITH BEST-SR ANNOTATIONS ONLY	5,240
DS-F-ALL	DS-F WITH ALL-SR ANNOTATIONS ONLY	9,737
DS-F-ALL-SUB	DS-F-ALL REDUCED TO SIZE OF DS-F-BEST	5,240

[6] GloVe is a count-based method for creating distributed word representations; essentially, it starts with the word co-occurrence matrix computed from a given corpus, factorizes it into a lower-dimensional word-by-features matrix, and uses the rows in the matrix (indexed by the words) as word vectors.

TABLE 2 Language and visual features as used by the ML methods in Section V-C.
Note that the 17 numbered features correspond to feature vectors of length between 114 and 138, depending on method.

$F0$:	OBJECT LABEL L_s — DEFINITION DEPENDS ON LEARNING METHOD.	NB, DT, RF: {0, 1,…, 19}; OTHERS: 1-HOT ENCODING (20 BITS)
$F1$:	OBJECT LABEL L_o — DEFINITION DEPENDS ON LEARNING METHOD.	
$F2$:	AREA OF BOUNDING BOX OF Obj_s NORMALIZED BY IMAGE SIZE.	[0, 1]
$F3$:	AREA OF BOUNDING BOX OF Obj_o NORMALIZED BY IMAGE SIZE.	[0, 1]
$F4$:	RATIO OF Obj_s BOUNDING BOX AREA TO THAT OF Obj_o.	[0, SIZE OF Obj_s]
$F5$:	DISTANCE BETWEEN BOUNDING BOX CENTROIDS, NORMALIZED BY IMAGE DIAGONAL.	[0, 1]
$F6$:	AREA OF OVERLAP OF BOUNDING BOXES NORMALIZED BY THE AREA OF THE SMALLER BOUNDING BOX.	[0, 1]
$F7$:	DISTANCE BETWEEN CENTROIDS DIVIDED BY SUM OF SQUARE ROOT OF AREAS/2 (APPROXIMATED AVERAGE WIDTH OF BOUNDING BOXES).	[0, ~20]
$F8$:	POSITION OF Obj_s RELATIVE TO Obj_o EXPRESSED AS ONE OF 4 CATEGORIES, DEPENDING ON THE ANGLE WITH THE VERTICAL AXIS.	NB, DT, RF: {0, 1, 2, 3}; OTHERS: 1-HOT ENCODING (4 BITS)
$F9$–$F12$:	LET DISTANCE FROM IMAGE EDGE OF LEFT AND RIGHT EDGES BE $a1, b1$ FOR FIRST BOX AND $a2, b2$ FOR SECOND BOX: $F9 = (a2 - a1)/(b1 - a1), F10 = (b2 - a1)/(b1 - a1)$. SIMILARLY FOR THE TOP AND BOTTOM EDGES, GIVING $F11$ AND $F12$.	[~-40, ~+40]
$F13$:	ASPECT RATIO OF BOX OF Obj_s.	[0, ~10]
$F14$:	ASPECT RATIO OF BOX OF Obj_o.	
$F15$:	GLOVE WORD VECTOR FOR L_s.	HERE: $\sim[-2, +3]$
$F16$:	GLOVE WORD VECTOR FOR L_o.	

Algorithm 1 HPO nested within cross-validation nested within FO.

```
for a given feature set $F = \{f_1, f_2, \ldots, f_m\}$ do
  for each of $k$ runs of the cross-validation do
    Assign development, test and core training sets;
    for each grid point $G = \{v_1, v_2, \ldots, v_n\}$ in HP space to be tested do
      Train a model on core training set and test on development set, using HPs $G$ and features $F$;
      Compute $Acc(1)_{dev}$ (definition see following section) on the development set;
      Select $G_{best}$ with the highest $Acc(1)_{dev}$;
    end for
    Using $G_{best}$ and $F$, retrain the model on combined core training and development sets;
    Test the trained model on test set, yielding $Acc(1)_{test}$;
  end for
  Compute $Acc(1)^k_{test}$, the mean of the $k$ $Acc(1)_{test}$ scores, as final score for $F$;
end for
```

optimization (FO) framework. In FO, the task is to find the subset of features that maximizes a given performance metric, and this space can be searched exhaustively, at least in principle. For HPO, it is not possible to test all combinations of HP values because many HPs are continuous-valued, and some are unbounded; this is usually addressed by manually discretizing and bounding HP values which yields a grid of points in HP space which can be searched exhaustively, e.g. by grid, random or sequential search [45], [46].

For a full exploration of the feature and (discretized, bounded) HP spaces, the basic model training process would have to be repeated for every feature subset, for every combination of HP values. Bearing in mind, moreover, that there are usually multiple ML methods, this tends to be computationally infeasible. Our data set is small, and optimal HP values depend on the feature set being used; for these reasons, we cannot use the complexity-reducing option of fine-tuning HP values in a separate process and discarding the data used for it afterwards. Instead, we opt for a similarly[7] parsimonious set-up that incorporates a **grid search of the HP space** embedded within a **k-fold cross-validation** regime over $1/k$ development set, $1/k$ test set, and $(k-2)/k$ core training set, and can be summarized in informal pseudo-code as shown in Algorithm 1. This protocol is in turn embedded in a **greedy backward feature elimination** procedure which in each iteration removes the feature whose removal most improves, or minimizes worsening of, the $Acc(1)^k_{\text{test}}$ score; the order in which features are eliminated gives an idea of their relative usefulness and informs our discussion of features in Section VII-B.

For clarity, the results shown in Table 3 are the $Acc(1)^k_{\text{test}}$ scores for the complete feature set (Section V-A); also shown are the corresponding $Acc(2)$, $Acc(3)$, and $Acc(4)$ scores.

[7] Assuming one would always want to involve cross-validation because of the small size of the data set.

C. Models

Using the features from Section V-A, we separately trained models of the six types below. We use the scikit-learn[8] library to implement the DT, LR, SVM and RF models. All models output the probability vector for the prepositions, from which the evaluation measures (see following section) are calculated.

Baseline (BL): Select most frequent spatial relation (SR) for an object pair; back off to most frequent SR overall.

Naive Bayes (NB): A very simple classifier which assumes that each feature is conditionally independent of every other feature given the SR. We use the object label (F0, F1) probabilities as the prior and all other features (F2-F16) as the likelihood. The latter are modelled with a Gaussian distribution, except for F4 and F8 (for details see [31]).

Decision Tree (DT): Classifies data in a series of decisions based on conjunctions of features. Values for the maximum tree depth [2, 20], and minimum samples at split and leaf [1, 40], are determined by hyper-parameter optimization (HPO).

Logistic Regression (LR): A linear classifier which models the SR probabilities with a logistic function. The value for the inverse of regularization constant [0.1, 100.0] is determined by HPO. The regularization is set to L1-norm and the model uses one-versus-rest multi-class classification.

Support Vector Machine (SVM): A non-probabilistic binary linear classifier solving the multiclass case via (here) one-versus-one classification. The RBF kernel parameters, C [0.1, 100.0] and gamma [0.001, 1.0] are determined by HPO.

Random Forests (RF): A meta-estimator comprising multiple decision-tree classifiers fitted to sub-samples of the data, using averaging to improve predictive accuracy and to control overfitting. The number of estimators [10, 150], maximum features [1, 156], maximum tree depth [2, 20], and minimum samples at split and leaf [1, 40], are determined by HPO.

VI. Evaluation Methods for SR Prediction

System-level Accuracy: We use four different variants of system-level Accuracy, denoted $Acc(n)$, $n \in \{1, 2, 3, 4\}$, which return Accuracy rates for the top n outputs produced by systems, such that a system output is considered correct if at least one of the target (human-selected) outputs is in the top n system outputs (for $n = 1$ this yields standard Accuracy).

Weighted Average Per-relation Precision: This measure, denoted Acc_P, computes the weighted mean of individual per-relation precision scores. Individual precision for a relation p is the proportion of times that p is in the human-selected set of reference outputs (target outputs) in those cases where p is returned as the top preposition by a method.

Human-assessed Measures: For the human evaluations, we first randomly selected four of the test instances (ordered object pairs with labels L_s, L_o for a specific image) from each of the five test sets from our five cross-validation runs. We then collected the outputs (top-ranked prepositions) $p_M^{s,o}$ produced for the 20 selected test instances by the baseline system and the three overall best methods M, all *as trained in the corresponding run of the cross-validation*, thereby creating 80 individual evaluation items $(L_s, L_o, p_M^{s,o})$.

TABLE 3 *Acc*(*n*) **results for all ML methods on DS-F data; statistical significance of** *Acc*(1) **differences indicated in second column (from one-way ANOVA with post-hoc Tukey HSD test).**

METHOD	TUKEY GROUP	*Acc*(1)	*Acc*(2)	*Acc*(3)	*Acc*(4)
BL	D	66.4	79.7	88.3	92.9
NB	D	68.5	83.8	90.8	94.7
DT	C	75.2	87.9	93.0	95.2
LR	C	77.1	90.9	95.4	97.8
SVM	B	79.7	91.9	96.0	97.9
RF	A	**82.4**	**92.2**	**96.4**	**98.0**

We recruited four native speakers as evaluators and assigned 20 evaluation items to each of them. The assignment ensured that each evaluator saw each method and each image/object-pair combination the same number of times. We achieved this via a Latin Square experimental design, using five sequenced 4 × 4 Latin Squares, so that evaluator E_i evaluates the ith row of every square. For details, see Kow & Belz, 2012 [47].

We presented the evaluation items to each evaluator in randomized order using an interface almost identical to the left half of Figure 2, except that we displayed the preposition selected by the given method instead of the empty slot (so evaluators were assessing simple 'L_s est $p_M^{s,o}$ L_o' phrases, e.g. *le chien est devant la personne*). This was shown alongside two questions: 'Is this statement true?' (YES, NO, UNSURE), and 'Is this a good way to describe the spatial relationship between the two objects?' (VERY GOOD, GOOD, OK, NOT VERY GOOD, BAD). Evaluators were first shown a page of instructions with examples and three practice items.

Results for this evaluation method are reported in Section VII-C below. A separate set of human evaluation results using a similar experimental design, but for the fully realized, complete image descriptions, is reported in Section IX-F.

VII. Primary Results

The primary results were obtained with the ML set-up described in Section V-B, with $k = 5$ and dividing the data into k non-overlapping subsets by means of **stratified sampling**, rather than random sampling, to ensure approximately equal representation of spatial relation classes [48].

A. Results Using All Features

Table 3 shows $Acc(n)$ results for models trained on the set of all features. The RF model has the highest scores for $n = 1$ (the main results), and for higher values of n although with decreasing margins. The LR model is far behind for $n = 1$, but catches up at $n = 4$. Baseline (BL) results remain behind by a substantial

[8] http://scikit-learn.org

margin, and NB and DT scores are lower than those of the best methods for all *n*.

Table 4 shows the corresponding set of relation-level Acc_P scores in order of relation frequency (in brackets); the actual (system-level) Acc_P results are shown in the bottom row. Note that the system-level Acc_P score for SVM is strongly affected by the relation-level Acc_P score for a_cote_de, a very high frequency item. Generally, the higher frequency relations are predicted with higher precision, but higher frequency is not necessary for good generalization, as demonstrated by *dans* (74 occurrences) and *autour de* (42) for which the overall best and fourth best Acc_P scores, respectively, are obtained. It seems likely that some relations are harder to identify than others from 2D images in our domain.

In each row in the table, the best score for the relation indicated in the first column is highlighted in bold. The RF method has the best Acc_P score for 10 of the relations; NB has three of the best (although one is joint best and the other two, for a_l'exterieur_de and par_dela are very low); LR and SVM have two each.

B. Feature Elimination

Table 5 shows results from greedy backward feature elimination which excludes the least useful feature in each iteration (see Section V-B). We let the procedure continue until all but one feature had been eliminated; for all learning methods there was an initial improvement in the *Acc*(1) score (second column in Table 5) after which *Acc*(1) got worse again (third column in Table 5).

TABLE 4 Relation-level Acc_P results (system-level results in bottom row) on DS-F data, shown in order of SR frequency (*n*) in annotations. '–' means the relation was never predicted by the method; 0 means predictions were wrong every time.

	RELATION-LEVEL Acc_P					
SPATIAL RELATION (*n*)	**BL**	**NB**	**DT**	**LR**	**SVM**	**RF**
pres_de (2,808)	73.5	**83.9**	78.1	80.4	81.1	**83.9**
a_cote_de (1,740)	17.9	79.5	73.3	50.4	0.0	**94.2**
devant (1,353)	48.0	72.1	72.5	76.8	78.6	**79.5**
derriere (1,300)	58.9	55.6	71.2	80.4	**80.8**	75.9
au_niveau_de (1,132)	0.0	61.5	62.4	44.4	55.0	**72.3**
contre (718)	58.9	53.9	39.9	59.9	65.5	**71.6**
sous (525)	56.2	65.5	71.7	74.3	77.1	**81.9**
loin_de (470)	58.3	40.2	60.0	70.3	75.5	**83.8**
sur (443)	56.2	75.1	77.5	76.5	78.1	**82.3**
en_face_de (333)	57.1	48.6	41.7	61.8	35.6	**91.5**
au_dessus_de (143)	50.2	–	50.0	**76.9**	50.0	50.0
le_long_de (83)	–	0.0	–	**39.3**	–	–
dans (74)	–	41.7	33.3	56.7	**95.0**	76.0
a_l'exterieur_de (51)	14.0	**36.8**	–	36.1	0.0	–
par_dela (47)	–	**10.2**	–	–	–	–
autour_de (42)	0.0	38.7	63.2	65.5	83.6	**87.5**
aucun (28)	–	–	0.0	0.0	–	–
SYSTEM Acc_P (11,290)	50.2	68.5	68.7	68.0	61.8	81.6

TABLE 5 Results on DS-F data for feature optimization (FO) on *Acc*(1), using greedy backward feature elimination.

			Acc(1)		Acc_P	
METHOD	**REMOVED FEATURES IN ORDER OF ELIMINATION**	**OPTIMIZED FEATURE SET (FEATURES SHOWN IN ORDER OF CONTINUED ELIMINATION)**	**AFTER FO**	**BEFORE FO**	**AFTER FO**	**BEFORE FO**
BL	N/A	(0–16)	(66.4)	(66.4)	(50.2)	(50.2)
NB	16, 15, 13, 7, 3, 11, 9, 2, 14, 10	4, 6, 5, 12, 8, 0, 1	74.2	68.5	71.2	68.5
DT	11, 3, 10, 9, 13, 5, 1	14, 0, 2, 7, 4, 8, 6, 15, 16, 12	76.6	75.2	69.5	68.7
LR	3, 2, 16, 0	14, 9, 10, 5, 13, 11, 6, 4, 8, 7, 1, 15, 12	77.7	77.1	68.2	68.0
SVM	13, 15, 16, 4	10, 9, 14, 2, 8, 6, 5, 3, 11, 1, 0, 7, 12	80.3	79.7	78.6	61.8
RF	6, 8, 9, 2, 5, 13, 14	1, 0, 3, 10, 11, 4, 7, 15, 16, 12	82.6	82.4	81.2	81.6

Note we are optimizing on *Acc*(1); *Acc*$_P$ is also shown for cross-reference. FO leads to improvement even for methods such as RF that can switch features off via weights, although except for NB, improvement is small. For all methods except RF, improvement in *Acc*(1) is paralleled by improvement in *Acc*$_P$.

The elimination process also enables insights into the relative usefulness of features. For the LR and SVM models, four features can be removed before *Acc*(1) gets worse again, for DT and RF it is seven, and for NB ten. There are patterns in the order in which features are removed that can be observed across ML methods. The single most useful feature is F12 (capturing the extent to which the top of the landmark object extends into the trajector object, expressed as a proportion of the trajector's height) which is retained until the end by DT, LT, SVM, and RF, and is in the last four for NB. F4 (size ratio), F6 (normalized overlap) and F8 (angle between centroids) are in the optimized feature set for four out of five methods.

There is evidence that F0/F1 (the trajector and landmark class labels, interpretable as words, e.g. person, bicycle), and F15/F16 (word vectors for these same class labels), can substitute for each other: models tend to retain F0 if F15 has been eliminated (and vice versa), and to retain F1 if F16 has been eliminated (and vice versa). E.g. NB and SVM eliminate F15 and F16 early on and retain F0 and F1 in the final four features; LR eliminates F16 and F0 early on and retains F1 and F15 in the final three features. There is evidence of this for all methods, without a clear preference for getting rid of simple labels sooner than lengthy word vectors; this indicates that for this particular task word vectors afford no advantage over words.

Among the *least* useful features are F13 (aspect ratio of trajector object) which is eliminated by four out of five methods (while results are still improving), F2/F3 (BB sizes relative to image size), F9 (left edge displacement), and F16 (word vector for landmark label), the latter four features being eliminated by three out of five methods. Feature elimination and retention patterns would probably be more uniform if it was feasible to carry out tests for all subsets of features; as it is, the greedy algorithm cannot fully take account of the interactions between features.

C. Human Evaluation Results

Table 6 shows human evaluation results (method see Section VI). Evaluators broadly agree with the automatic metrics in terms of ranking the three systems, for both truth and quality. The RF model stands out for producing statements that are true in all cases except one where the evaluator was unsure, are good or very good in 80% of all cases, with no bad ones at all. The last column shows averages over quality ratings, but this needs to be interpreted with caution, because the distances between the scores cannot be assumed to be equal in all cases. We ran a non-parametric test (Wilcoxon) which found the differences between RF, SVM, LR and the baseline significant for truth, and between RF and BL for quality, but a Dunn-Bonferroni correction left only the difference between RF and BL as significant in both cases.

TABLE 6 Human evaluation of SR detection (on DS-F data): cells show number of times a system received a rating; *average rating for perspective on rank (but see in text for caveat).

	STATEMENT TRUE?			GOOD DESCRIPTION?					
	YES	NO	UNSURE	5 = VERY GOOD	4 = GOOD	3 = OK	2 = NOT VERY GOOD	1 = BAD	*AVERAGE
BL	10	10	0	3	4	3	3	7	2.65
LR	16	3	1	5	3	5	5	2	3.2
SVM	15	1	4	4	3	7	5	1	3.2
RF	19	0	1	5	11	0	4	0	3.85

TABLE 7 *Acc*(1), *Acc*(2) and *Acc*(3) results using best prepositions ('best'); all prepositions ('all'), and a randomly selected subset ('all-sub') of 'all' equal in size to 'best'.

		Acc(1)	*Acc*(2)	*Acc*(3)
DT	DS-F-BEST	51.6	71.8	83.1
DT	DS-F-ALL	67.7	81.4	91.0
DT	DS-F-ALL-SUB	64.7	80.9	88.8
NB	DS-F-BEST	57.6	74.8	84.0
NB	DS-F-ALL	64.7	80.9	90.4
NB	DS-F-ALL-SUB	61.2	78.8	88.3
LR	DS-F-BEST	**59.3**	78.8	88.8
LR	DS-F-ALL	**74.9**	89.2	94.2
LR	DS-F-ALL-SUB	**73.6**	88.4	93.9

VIII. Further Analysis

A. Training on Best SRs vs. All SRs

In this section, we compare results for training on best prepositions only vs. training on all possible prepositions. For this specific set of experiments (previously reported [42]), we used manually set hyper-parameters, a simple leave-one-out cross-validation set-up, no feature optimization, and slightly different versions of NB and DT, all as described in [42].[9] DS-F contains more than twice the number of instances for all possible prepositions (9,278) than for best prepositions only (4,140); we therefore also report (DS-F-all-sub in Table 7) results for a randomly reduced subset of the all-prepositions data of the same size as the best-prepositions-only data (averaged over four different random reductions).

[9] We did not rerun these experiments, because the validity of the claim being made does not depend on the learning method or data set being used.

[Humans] do not take into account a scene's full, complex details, but instead access radically simplified representations.

The results in Table 7 clearly show the benefit of training on all possible prepositions compared to best only, although it is less marked for the NB method. While results for DS-F-all are higher than for DS-F-all-sub, and one aspect of this is likely to be larger training set size, the DS-F-all-sub results nevertheless show clearly that the biggest factor is training on all possible prepositions rather than just the best (that being the only difference between DS-F-best and DS-F-all-sub).

B. Different Quality Annotations

There is a marked difference in intra-annotator agreement for the three annotators (see also Section IV-C): Annotator 0: $\kappa = 0.7$; Annotator 1: $\kappa = 0.78$; Annotator 2: $\kappa = 0.83$. We tested what happens if we train on annotations of different quality, as captured by intra-annotator agreement. We separately trained only on the worst annotator vs. only on the best. To ensure equal data set sizes, we reduced the larger set randomly to match the size of the smaller set. The $Acc(n)$ evaluation results in Table 8 paint a very clear picture: training only on the most self-consistent annotations leads to much better results.

IX. Language Generation Component

The methods described above are used within our image description approach to generate the set of all best pairwise spatial relations (SRs) for a given image. This set contains more SRs than can be used in a description: two SRs for each pair of objects (one for each ordering), both possibly of the same type.

Section IX-B describes how the first step in the language generation component (content selection/ordering) reduces this set to those SRs that will be realized in the image description. Sections IX-C to IX-E describe the remaining steps: aggregation, referring expression generation (REG), and surface realization which outputs the final, fully realized image description. We start with a look at related research.

TABLE 8 ***Acc*(1) and *Acc*(2) results, methods trained separately on annotations of high quality ('$\kappa = 0.83$'), and on annotations of lower quality ('$\kappa = 0.7$').**

	DS-F		
METHOD	**INTRA-ANNOTATOR AGREEMENT**	***Acc*(1)**	***Acc*(2)**
BL	$\kappa = 0.83$	61.3%	74.4%
	$\kappa = 0.7$	59.5%	71.3%
NB	$\kappa = 0.83$	76.4%	87.3%
	$\kappa = 0.7$	64.1%	80.2%
DT	$\kappa = 0.83$	78.4%	87.5%
	$\kappa = 0.7$	69.1%	83.4%
LR	$\kappa = 0.83$	61.3%	74.4%
	$\kappa = 0.7$	59.5%	87%
SVM	$\kappa = 0.83$	86.5%	94.9%
	$\kappa = 0.7$	70.6%	80.2%
RF	$\kappa = 0.83$	89.1%	96.8%
	$\kappa = 0.7$	77.1%	90.5%

A. Related Research in NLG

We previously mentioned (Section III) work in image description with a distinctly NLG flavor, some using template-based techniques [29], [30], some a grammar-based approach [33], [34] to assemble descriptions. Outside of image description, spatial relations play a role in REG [36], [37] where the task is often framed as a content selection problem from known symbolic representations of objects and scene, and the aim is to uniquely identify the referent. In fact, the kinds of descriptions we generate (see below) can be seen as containing relational referring expressions. Viethen and Dale [36] provide evidence that relations are frequently used in referring even where not necessary for identification. Krahmer and van Deemter observe that if relations are important and frequently used, then this may call into question the validity of the incremental generation algorithm, perhaps the most influential in REG, because it does not readily accommodate relations [49, p. 184].

The steps we perform in our NLG component are readily recognisable as a traditional NLG pipeline [50] with standard symbolic NLG techniques. Our strategies for realization of spatial relations and their two objects produce combinations of the pattern NP_{subj} *is preposition* NP_{obj}, although we leave out the copula in some cases, and aggregation sometimes results in elided subject NPs and coordinated object NPs. This is the pattern identified by both [10] and [14] as the most frequent syntactic realization of location expressions, with the trajector assuming subject, and the landmark object, position.

B. Content Selection and Ordering

We distinguish two cases: images with two identified objects and images with three. In the former case, the SR set contains just two SRs of which we select one in a simpler process as described at the end of this section. For images with three annotated objects, there are initially six SRs; for example, for the image in Figure 1, the SR set might be the following:

{*next_to*(*bicycle*$_1$, *bicycle*$_2$), *in_front_of*(*bicycle*$_1$, *person*), *next_to*(*bicycle*$_2$, *bicycle*$_1$), *next_to*(*bicycle*$_2$, *person*), *behind*(*person*, *bicycle*$_1$), *next_to*(*person*, *bicycle*$_2$)}

The task now is to select and order the subset of SRs that will be expressed in the description. We make two assumptions: (i) that we will want to include just one SR with the same two arguments, e.g. have either *in_front_of*(*bicycle*$_1$, *person*), or *behind*(*person*, *bicycle*$_1$), but not both; and (ii) while each object must appear in at least one SR, it does not have to appear in more than one. On this basis, we apply one of the following six content selection/ordering strategies.

Random Chaining (RC): Randomly choose the first SR $p_{1,2}(L_1, L_2)$, then keep choosing the next SR randomly from among those that have L_2 as their first argument and do not have a previous first argument as their second, until no SRs match the constraints. For example, this strategy might select the following SR subset given the candidate SR set above:

{in_front_of(bicycle$_1$, person), next_to(person, bicycle$_2$)}

Random Fanning (RF): Randomly choose an object Obj_1, then select all SRs $p_{1,i}(L_1, L_i)$, in random order. E.g. for the above SR set:

{next_to(bicycle$_1$, bicycle$_2$), in_front_of(bicycle$_1$, person)}

Biggest-first Chaining (BFC): Select the two objects Obj_1 and Obj_2 that have the biggest and second biggest bounding box (BB) areas, respectively. Select SR $p_{1,2}(L_1, L_2)$, then keep choosing the next SR $p_{2,i}(L_2, L_i)$ such that Obj_i is the next biggest object and L_i has not been a first argument yet. E.g. for the above SR set:

{next_to(bicycle$_2$, bicycle$_1$), in_front_of(bicycle$_1$, person)}

Biggest-first Fanning (BFF): Select the object Obj_1 with the biggest BB area, then select all SRs $p_{1,i}(L_1, L_i)$, in order of L_i BB area sizes. E.g. for the above SR set:

{next_to(bicycle$_2$, bicycle$_1$), next_to(bicycle$_2$, person)}

Human-centric Biggest-first Chaining (HCBFC): Select the two objects Obj_1 and Obj_2 that have the biggest and second biggest BB areas, respectively, among objects with a person label while available. Choose SR $p_{1,2}(L_1, L_2)$, then keep choosing the next SR $p_{2,i}(L_2, L_i)$ such that Obj_i is the next biggest object, is of type person if available, and L_i has not been a first argument yet. E.g. for the above SR set:

{next_to(person, bicycle$_2$), next_to(bicycle$_2$, bicycle$_1$)}

Human-centric Biggest-first Fanning (HCBFF): Select Obj_1 with the biggest BB area from among all objects with a person label while available, then select all SRs $p_{1,i}(L_1, L_i)$, in order of the Obj_i BB area sizes, adding those where L_i is a person label first. E.g. for the above SR set:

{next_to(person, bicycle$_2$), behind(person, bicycle$_1$)}

If the image has only two objects, then for the random content selection/ordering strategies we select one of the two SRs at random; for the biggest-first strategies we select the SR with the bigger Obj_1 BB area, and if they are equal then the SR with the bigger Obj_2 BB area; for the human-centric strategies we select the SR that has a person as Obj_1, and if both do then the SR with the bigger Obj_1 BB area, and if those are equal in size, then the SR with the bigger Obj_2 BB area. The ultimate back-off is always random selection.

C. Aggregation

The aggregation strategies apply when there is a shared first argument or if there is a shared first argument and spatial relation. This can only happen for the fanning strategies. The two aggregation rules we use transform two relations into one nested relation, and can most easily be explained graphically as shown in Figure 3.

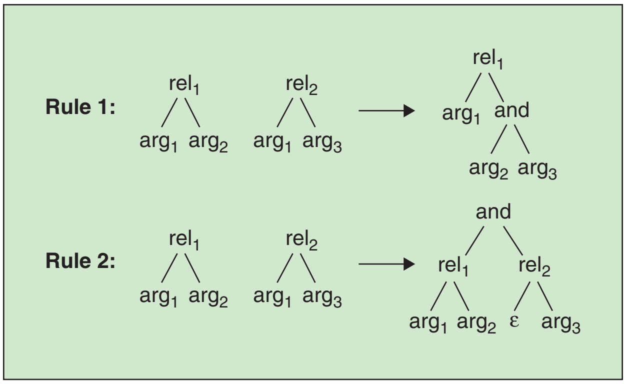


FIGURE 3 Aggregation rules (ϵ = the null string).

After aggregation, e.g. the SR sets created by Biggest-first Fanning (BFF) and Human-centric Biggest-first Fanning (HCBFF) look as follows:

BFF: {*next_to* (*bicycle*$_2$, *and* (*bicycle*$_1$, *person*))}

HCBFF: {*and* (*next_to* (*person*, *bicycle*$_2$), *behind* (ϵ, *bicycle*$_1$))}

D. Referring Expressions

For referring expression generation (REG), we process the SRs in order, overwriting first and/or second arguments as required by the rules below; 'mention' here means 'is an argument in an SR already processed'. We use the following REG rules (based on previous work [51]):

1) First argument in the current SR is the second argument in the last SR: relative pronoun, e.g. who.
2) First mention of object O and first mention of any object of O's type T_O: indefinite determiner and noun, e.g. a bicycle.
3) First mention of object O and there have been mentions of $n - 1$ other objects of O's type T_O: indefinite determiner, ordinate numeral nth, and noun, e.g. a second bicycle.
4) Non-first mention of object O and there have been no mentions of other objects of O's type T_O: definite determiner and noun, e.g. the bicycle.
5) Non-first mention of object O, there have been mentions of other objects of O's type T_O, and O is the nth object to be mentioned: definite determiner, ordinate numeral nth, noun, e.g. the first bicycle.

Note that the aggregation rules (previous section), as a side effect, address one aspect of REG, namely ellipsis, by setting a repeated first argument to the null string ϵ. After REG, the two examples from above look as follows:

BFF: {*next_to* (a bicycle, *and* (a second bicycle, a person))}

HCBFF: {*and* (*next_to* (a person, a bicycle), *behind* (ϵ, a second bicycle))}

E. Surface Realization

The surface realization stage lexicalizes the spatial relations and inserts copulas where required (e.g. after relative pronouns) after which it linearizes the SR trees left to right. The final steps are morphological processing, agreement checking,

TABLE 9 Evaluators' scores for three best SR detection methods plus baseline, within descriptions generated by same NLG strategy.

	DESCRIPTION CORRECT?				DESCRIPTION COMPLETE?				DESCRIPTION NATURAL?			
SR DETECTION METHOD	RF	LR	SVM	BL	RF	LR	SVM	BL	RF	LR	SVM	BL
AVERAGE RANK PER NLG METHOD	1.4	2.2	3.4	3.2	1.7	2.5	2.9	2.9	2.7	3	2	1.8
AVERAGE RATING	3.67	3.33	2.83	2.86	3.61	3.28	3.19	3.25	3.72	3.58	3.78	3.83

upper/lower-casing, and formatting. After surface realization, the final realizations of all SR sets from above look as follows:

RC:	a bicycle in front of a person who is next to a second bicycle
RF:	a bicycle next to a second bicycle and in front of a person
BFC:	a bicycle next to a second bicycle which is in front of a person
BFF:	a bicycle next to a second bicycle and a person
HCBFC:	a person behind a bicycle which is next to a second bicycle
HCBFF:	a person next to a bicycle and behind a second bicycle

F. Evaluation

We use the same basic experimental design as for the other human evaluations (Section VI). However, this time we could not use items from previous test sets for evaluation, because those correspond to object pairs, not images. Instead we randomly selected six *new* images from those with four bounding boxes in the VOC'08 data set, but where one of the bounding boxes is very small and labeled 'difficult.' As this usually means highly truncated or barely visible, we can ignore the difficult object, and treat the image as having three annotated objects (the image in Figure 1 is an example). For each image (6) and each of our three best optimized spatial-relation identification methods (LR, SVM, RF) plus the baseline (4), we generated a description with each description generation strategy (6), resulting in a total of 144 (6 × 4 × 6) descriptions to be evaluated. Using six native speakers as evaluators, and four sequenced 6 × 6 Latin squares, we presented each image description alongside evaluation questions in a way very similar to Figure 2: the image with three bounding boxes to the left, the image description underneath, and the evaluation questions in the right half of the screen:

1) Is the description correct (true)? 5=HIGHLY CORRECT, 4=CORRECT, 3=OK, 2=NOT VERY CORRECT, 1=HIGHLY INCORRECT
2) How complete a description of the objects in bounding boxes is this? 5=VERY COMPLETE, 4=COMPLETE, 3=OK, 2=NOT VERY COMPLETE, 1=VERY INCOMPLETE
3) Is it a natural-sounding description? 5=VERY NATURAL, 4=NATURAL, 3=OK, 2=NOT VERY NATURAL, 1=VERY UNNATURAL

We looked at how evaluators' scores ranked the four SR detection methods within descriptions generated by the same NLG strategy, and then averaged rank for each SR detection method over all NLG strategies, giving the results in the second and third row in Table 9, respectively.

Models are ranked identically for correctness and completeness (and identically to the last two evaluations). The corresponding average ratings for each SR detection method show very similar trends. A non-parametric test (Wilcoxon with Dunn-Bonferroni correction) showed significance for the difference between RF and BL, and between RF and SVM, for correctness. Results for naturalness are hard to interpret, considering how very close average ratings are.

Results were inconclusive for evaluating different NLG strategies; some trends were that the chaining strategies are all ranked more highly on average than the fanning strategies in terms of naturalness; biggest-first may be an advantage over other methods in terms of correctness; and person-first and biggest-first may be best and worst, respectively, for completeness.

X. Discussion

A. Contributions and Limitations

In producing our results, we gained some insights into the relative utility of different ML methods, annotation strategies, and features for the task of spatial relation (SR) detection. The optimized RF model achieves the highest SR detection accuracy rates (up to 89.1%) as well as the highest human scores. Moreover, the optimized RF model does similarly well across all but one SR that it makes predictions for, including less frequent ones, i.e. high performance is not achieved at the price of specializing on high-frequency items.

There are some clear conclusions for data annotation: (1) it is worth ensuring high levels of self-consistency in annotators; and (2) it is better to ask annotators to provide all possible solutions, than just the best one. Although there is likely to be variation in what different annotators consider 'best', our results show that having several annotators does not make up for not having variation explicitly included in the annotations.

In terms of features, we found that distributed word vector representations offer no advantage over the words themselves in our context. The single most useful feature was F12 which captures the extent to which the top of the landmark object extends into the trajector object, expressed as a proportion of the trajector's height.[10] Why F4 (size ratio), F6 (normalized overlap) and F8 (angle between centroids) are useful features seems clear; as for F12, its value is positive if

[10] We are grateful to David Hogg for suggesting F8, F9, F10 and F11.

the top of the landmark is higher than the bottom of the trajector, and negative otherwise, and its magnitude increases with the distance between the two edges (greater negative values indicating greater vertical distance between the bounding boxes, and greater positive values indicating greater vertical overlap). It seems likely that this helps with above/below and relations involving a notion of physical proximity/distance, but perhaps most importantly with in front of/behind type 3D relations (especially because F12 is normalized by trajector height). The latter are particularly hard to detect in a 2D setup.

There are things we perceive only because it suits the requirements of our language.

We cannot claim that we capture all aspects, or even all important aspects, of every picture: we focus on SRs and do not address object attributes, scene background or activities. Nevertheless, good average completeness ratings (between OK and GOOD) for the best methods indicate that our descriptions capture a lot of the important content a lot of the time.

We perform content selection from the complete set of SRs, selecting one SR each for the object pairs we know about, and assign trajector (Figure) and landmark (Ground) roles as a side effect of determining syntactic structure. The evidence from cognitive linguistics suggests that in human scene description, linguistic goals drive schematization and configuration of just the main scene components, including assignment of Figure role (an object that is moving or is conceptually movable) and Ground role (an object conceived of as stationary) [10, p. 167]. While what we do seems sufficient in our context, in descriptions of images that have region annotations covering the entire image, these strategies would not suffice and something like Talmy's primary breakup of a spatial scene may have to be performed early on, perhaps learnt from data, to account for how language "mark[s] out one portion within a scene for primary focus and […] characterize[s] its spatial disposition in terms of a second portion […] and sometimes also a third […]" [9, p. 182].

B. Language Structures Space

Herskovits [10], [12] and others have argued persuasively that image description by humans is driven by linguistic goals and subject to language-specific constraints to the point of *language-induced percepts*, meaning that there are things we perceive only because it suits the requirements of our language. It is language that structures space, not vice versa. This implies that work on automatic detection of relationships between objects in images should be linguistically informed, if only to define relationships in such a way that they correspond clearly to linguistic entities. This surely is important whether the aim is to generate human-like image descriptions, or simply to create descriptions that are useful for, or are rated highly, by humans. In this context it seems a missed opportunity for recent research on relationship detection in images [16], [17] to use sets of relationships that have very little linguistic or cognitive grounding—at the least such relationships will not be ideal as inputs to language generation.

C. Using Human-Authored Image Descriptions

Our methods do not learn or evaluate from human-authored language. Automatic image description work, strongly reliant as it currently is on supervised machine learning methods, tends to start by gathering human-authored image descriptions, applying ML methods to learning from such human examples, and evaluating the machine-generated descriptions against the human-authored ones using some text similarity metric such as BLEU [52]. What is entirely lacking in this setup is any kind of extrinsic context [53]. Descriptions are deemed of high quality in proportion to their similarity to the human-authored descriptions, but the question what they are good *for* is not addressed. Similarly, in the human evaluations of our fully realized image descriptions reported in Section IX-F, evaluators were simply asked whether a description was correct/natural-sounding/complete. They were not told what the intended purpose or application context was for the descriptions. This may suffice for simple descriptions anchored in SRs. More generally, the application contexts in which one might wish to use (high quality) automatically generated image descriptions differ so greatly in terms of required focus, level of detail, style, etc. that it is surely impossible to say whether an image description is a good one without a specified extrinsic context, such as an application which it is supposed to be good for.

XI. Concluding Comments

The questions from the beginning of this paper were: (1) to what level of accuracy can spatial relations (SRs) be determined by ML methods; (2) to what extent do human authors agree when determining SRs from still images; and (3) what level of quality can be achieved with image descriptions anchored in SRs. Regarding (1), our best method (random forests) achieved a highest accuracy rate of 89.1% (when trained on high-quality annotations) which is far higher than our previous best results of 75% [42], and higher than best accuracy results of 80% reported in comparable work [38].

Regarding (2), our annotators achieved average pairwise kappa scores of 0.67, which is good, especially considering that one of the annotators had substantially lower agreement scores than the others, and that kappa is a relatively conservative estimate of agreement.

Regarding (3), our complete image descriptions were judged highly by evaluators which is perhaps surprising considering they consist solely of referring expressions for objects and prepositions describing the SRs between them. However, as mentioned in the discussion of limitations in the preceding section, such results need to come with the caveat that for a true measure of quality, evaluation of automatic image description methods needs to incorporate an application context, a specification of what the generated descriptions are *for*. There is currently no common method for this in the image description field, something we plan to address in future research.

References

[1] A. Workman, M. Herbert, H. Rowlatt, and D. Williams, "Getting connected: Engaging people with sight loss in accessible technology," Tech. Rep: Royal National Institute of Blind People, 2013.

[2] F. Barton, G. Bradbrook, and G. Broome, "Digital accessibility: A brief landscaping," Citizens Online, UK, Tech. Rep., Mar. 2015.

[3] "Visual impairment and blindness," World Health Organization, Fact Sheet No 282, Aug. 2014.

[4] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proc. NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, California, June 6, 2010, pp. 139–147.

[5] E. Reiter and A. Belz, "An investigation into the validity of some metrics for automatically evaluating natural language generation systems," *Comput. Linguist.*, vol. 35, no. 4, pp. 529–558, Dec. 2009.

[6] G. Leech, *Towards a Semantic Description of English*. New York: Longman Press, 1969.

[7] D. Bennett, *Spatial and Temporal Uses of English Prepositions: an Essay in Stratificational Semantics*. New York: Longman, 1975.

[8] R. Jackendoff, "Semantics of spatial expressions," in *Semantics and Cognition*. Cambridge, Massachusetts: MIT press, 1983, ch. 9, pp. 161–187.

[9] L. Talmy, "How language structures space," in *Spatial Orientation: Theory, Research and Application*, H. L. Pick Jr. and L. P. Acredolo, Eds. New York: Plenum Press, 1983, ch. 3, pp. 225–282.

[10] A. Herskovits, "Language, spatial cognition, and vision," in *Spatial and temporal reasoning*, O. Stock, Ed. Norwell, MA, USA: Kluwer, 1997, ch. 6, pp. 155–202.

[11] J. D. Kelleher and F. J. Costello, "Applying computational models of spatial prepositions to visually situated dialog," *Comput. Linguist.*, vol. 35, no. 2, pp. 271–306, June 2009.

[12] A. Herskovits, *Language and Spatial Cognition*. New York, NY, USA: Cambridge Univ. Press, 1987.

[13] B. Landau and R. Jackendoff, "Whence and whither in spatial language and spatial cognition?," *Behav. Brain Sci.*, vol. 16, no. 2, pp. 255–265, June 1993.

[14] R. Huddleston and G. Pullum, *The Cambridge Grammar of the English Language*. Cambridge: Cambridge Univ. Press, 2002.

[15] A. Rosenfeld, "Iterative methods in image analysis," *Pattern Recognit.*, vol. 10, no. 3, pp. 181–187, June 1978.

[16] M. Yatskar, V. Ordonez, and A. Farhadi, "Stating the obvious: Extracting visual common sense knowledge," in *Proc. Conf. North American Chapter Association for Computational Linguistics: Human Language Technologies*, San Diego California, USA, June 12-17, 2016, pp. 193–198.

[17] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. 14th European Conf. Computer Vision*, Amsterdam, The Netherlands, Oct. 11–14, 2016, pp. 852–869.

[18] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, pp. 1–42, Feb. 2017.

[19] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Trans Assoc Comput Linguist.*, vol. 2, pp. 27–218, June 2014.

[20] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, June 7-12, 2015, pp. 3128–3137.

[21] R. Mason and E. Charniak, "Nonparametric method for data-driven image captioning," in *Proc. 52nd Annu. Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, June 22-27, 2014, vol. 2, pp. 592–598.

[22] A. Gupta, Y. Verma, and C. V. Jawahar, "Choosing linguistics over vision to describe images," in *Proc. 26th AAAI Conf. on Artificial Intelligence*, Toronto, Ontario, Canada, July 22-26, 2012, pp. 606–612.

[23] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proc. Advances in Neural Information Processing Systems*, Granada, Spain, Dec. 12-17, 2011, pp. 1143–1151.

[24] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, data sets, and evaluation measures," *J. Artif. Intell. Res.*, vol. 55, pp. 409–442, Jan. 2016.

[25] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Proc. 11th European Conf. on Computer Vision*, Crete, Greece, Sept. 5-11, 2010, pp. 15–29.

[26] M. Yatskar, L. Vanderwende, and L. Zettlemoyer, "See no evil, say no evil: Description generation from densely labeled images," in *Proc. 3rd Joint Conf. Lexical and Computational Semantics*, Dublin, Ireland, Aug. 23-24, 2014, pp. 110–120.

[27] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollazr, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig., "From captions to visual concepts and back," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, June 7-12, 2015, pp. 1473–1482.

[28] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Baby talk: Understanding and generating simple image descriptions," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, June 20–25, 2011, pp. 1601–1608.

[29] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *Proc. 16th Conf. on Empirical Methods in Natural Language Processing*, Edinburg, Scotland, July 27-31, 2011, pp. 444–454.

[30] D. Elliott and F. Keller, "Image description using visual dependency representations," in *Proc. 18th Conf. on Empirical Methods in Natural Language Processing*, Seattle, Oct. 18-21, 2013, pp. 1292–1302.

[31] A. Belz, A. Muscat, M. Aberton, and S. Benjelloun, "Describing spatial relationships between objects in images in english and french," in *Proc. 4th Workshop on Vision and Language*, Lisbon, Portugal, Sept. 18, 2015, pp. 104–113.

[32] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proc. 15th Conf. on Computational Natural Language Learning*, Portland, Oregon, June 23/24, 2011, pp. 220–228.

[33] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé, III, "Midge: Generating image descriptions from computer vision detections," in *Proc. 13th Conf. of the European Chapter of the Association for Computational Linguistics*, Avignon, France, Apr. 23-27, 2012, pp. 747–756.

[34] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "Treetalk: Composition and compression of trees for image descriptions," *Trans. Assoc. Comput. Linguist.*, vol. 2, no. 10, pp. 351–362, June 2014.

[35] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. Berg, "Babytalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.

[36] J. Viethen and R. Dale, "The use of spatial relations in referring expression generation," in *Proc. 5th Int. Natural Language Generation Conf.*, Salt Fork, Ohio, June 12–14, 2008, pp. 59–67.

[37] D. Golland, P. Liang, and D. Klein, "A game-theoretic approach to generating spatial descriptions," in *Proc. 15th Conf. on Empirical Methods in Natural Language Processing*, Massachusetts, Oct. 9-11, 2010, pp. 410–419.

[38] A. Ramisa, J. Wang, Y. Lu, E. Dellandrea, F. Moreno-Noguer, and R. Gaizauskas, "Combining geometric, textual and visual features for predicting prepositions in image descriptions," in *Proc. 20th Conf. on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, Sept. 17-21, 2015, pp. 214–220.

[39] M. Hürlimann and J. Bos, "Combining lexical and spatial knowledge to predict spatial relations between objects in images," in *Proc. 5th Workshop on Vision and Language*, Berlin, Germany, Aug. 12, 2016, pp. 10–18.

[40] A. Muscat and A. Belz, "Generating descriptions of spatial relations between objects in images," in *Proc. 15th European Workshop on Natural Language Generation*, Brighton, UK, Sept. 10/11, 2015, pp. 100–104.

[41] A. Muscat, A. Belz, and B. Birmingham, "Exploring different preposition sets, models and feature sets in automatic generation of spatial image descriptions," in *Proc. 5th Workshop on Vision and Language*, Berlin, Germany, Aug. 12, 2016, pp. 65–69.

[42] A. Belz, A. Muscat, B. Birmingham, J. Levacher, J. Pain, and A. Quinquenel, "Effect of data annotation, feature selection and model choice on spatial description generation in french," in *Proc. 9th Int. Natural Language Generation Conf.*, Edinburgh, UK, Sept. 5-8, 2016, pp. 237–241.

[43] M. Everingham, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 303–338, June 2010.

[44] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. 19th Conf. Empirical Methods in Natural Language Processing*, Doha, Qatar, Oct. 25-29, 2014, pp. 1532–1543.

[45] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 281–305, Feb. 2012.

[46] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Proc. 25th Annu. Conf. on Neural Information Processing Systems*, Granada, Spain, Dec. 12-17, 2011.

[47] E. Kow and A. Belz, "Lg-eval: A toolkit for creating online language evaluation experiments," in *Proc. 8th Int. Conf. on Language Resources and Evaluation*, Istanbul, Turkey, May 21-27, 2012, pp. 4033–4037.

[48] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. on Artificial Intelligence*, San Francisco, Aug. 20-25, 1995, pp. 1137–1143.

[49] E. Krahmer and K. V. Deemter, "Computational generation of referring expressions: A survey," *Comput. Linguist.*, vol. 38, no. 1, pp. 173–218, Mar. 2012.

[50] E. Reiter and R. Dale, *Building Natural Language Generation Systems*. New York, NY, USA: Cambridge Univ. Press, 2000.

[51] A. Belz and S. Varges, "Generation of repeated references to discourse entities," in *Proc. 11th European Workshop on Natural Language Generation*, Schloss Dagstuhl, Germany, June 17-20, 2007, pp. 9–16.

[52] K. Papineni, S. Roukos, T. Ward, and W. j. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting on Association for Computational Linguistics*, Philadelphia, Pennsylvania, July 7–12, 2002, pp. 311–318.

[53] A. Belz, "That's nice… what can you do with it?," *Comput. Linguist.*, vol. 35, no. 1, pp. 118–119, Mar. 2009.