Generating Descriptions of Visual Data Anchored in Spatial Relations

Adrian Muscat[1], Anja Belz[2]

[1]University of Malta, Malta [2]University of Brighton, United-Kingdom

March 5-7, 2018

Introduction

Humans prioritise:

- Salient entities
- Attributes (colour, size, etc.).
- Relationships linking the entities and their surroundings

Truth Content / Conjecture

A man *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**.

Image 2008_008320 from VOC 2008 with annotations and image descriptions obtained by Rashtchian et al. (2010). (BB=bounding box; image from http://lear.inrialpes.fr/RecogWorkshop08/documents/everingham.pdf)

Our work

We focus on entities and spatial relations as aspects of image description that require no (or very little) conjecture.

e.g. *a boy* **on** *a bicycle*; *a dog* **in** *a lake*

Questions:

1. To what level of accuracy can spatial relations be determined by machine learning methods?
2. To what extent do human authors agree when determining spatial relations from still images?
3. What level of quality can be achieved with image descriptions anchored in automatically detected spatial relations?

Our Approach

A three-step task:

1. entity identification (which we do not address ourselves);
2. identifying the 3D spatial relations (SRs) between pairs of entities on the basis of 2D geometric features and language features
3. generating natural language (NL) descriptions from sets of spatial relations.

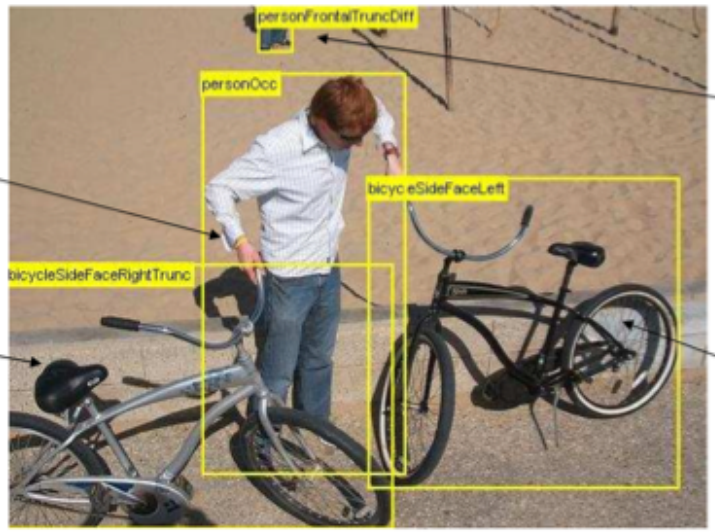
Machine Learning methods to generate prepositions.
Standard NLG pipeline to generate descriptions.
Human evaluations to test our methods

Language and Space

How language marks spatial distinctions and patterns, and ascribes structure to space. e.g (Herskovits 1997)

- Speakers do not take into account the full, complex details, but access radically simplified representations.
  - Schematisation reduces the complex entities and their arrangements to simple geometric descriptions such as points, lines, etc.
- Spatial relation terms are the principal means available to speakers for the description of location.
  - Typically one entity is taken as a reference point (*landmark*) with respect to which the other is located (*trajector*).
- Literature is not very explicit about how linguistic spatial relations formally map to prepositions. However prepositions have multiple senses each of which has multiple use types.
- Schematisation is parametrised by linguistic goal and is language specific among other things, and produces a single configuration.

Spatial Relations Dataset



- Data source: VOC'08 corpus of images (Everingham et al. 2010)
- Collected additional annotations from humans a set of prepositions
- Parsed Pascal Sentences Rashtchian et al. 2010 to obtain a set of candidate prepositions (38)
- From the latter mapped a corresponding set in French

1. Abstraction of the objects as rectangles.
2. Geometric feature computation yields further abstractions of objects e.g. as points, lines, trajectories, and distance between objects.
3. Scenes are 'configured' into pairs of objects where one is the trajector and the other the landmark.
4. Generate spatial relations (SR) and then a NL description

Language and Geometrical Feature Set

$F0$	Object label L_s (definition depends on learning method).
$F1$	Object label L_o (definition depends on learning method).
$F2$	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, normalised by image diagonal.
$F6$	Area of overlap of bounding boxes normalised by the area of the smaller bounding box.
$F7$	Distance between centroids divided by sum of square root of areas/2 (approximated average width of bounding boxes).
$F8$	Position of Obj_s relative to Obj_o expressed as one of 4 categories, depending on the angle with the vertical axis.
$F9$–$F12$	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 $F12$.
$F13$	Aspect ratio of box of Obj_s.
$F14$	Aspect ratio of box of Obj_o.
$F15$	GloVe word vector for L_s.
$F16$	GloVe word vector for L_o.

ML Models and Results

- Compare six ML models
- Hyperparameter optimisation
- Feature optimisation (wrapper method)
- Apply cross-validation

Method	Tukey Grouping	$Acc(1)$	$Acc(2)$	$Acc(3)$	$Acc(4)$
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**

Results per Preposition (priot to FO)

Spatial relation	Freq	ML Model					
		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	–	–

'–' means the relation was never predicted by the method; 0 means predictions were wrong every time.

Feature Optimisation - Greedy Backward Wrapper Method

ML	Removed features in order of elimination	Optimised feature set (features shown in order of continued elimination)	$Acc(1)$		Acc_P	
			AFO	BFO	AFO	BFO
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

- Distributed word vector representations offer no advantage over the words themselves in our context.
- F12 is the single most useful feature. Captures the extent to which the top of the landmark object extends into the trajector. This seems to help with in *front of/behind* type 3D relations.

Human Evaluations - Preposition Prediction

DS-F									
	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

Human evaluation results: cells show number of times a system received an assessment; *last column shows average rating for Description Quality to give perspective on rank.

Variation in Quality of Annotations

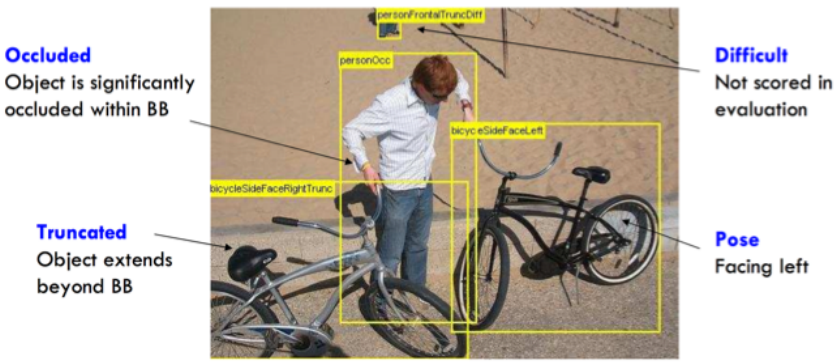
Method	Intra-annotator agreement	*Acc*(1)	*Acc*(2)
BL	$\kappa = 0.83$	61.3%	74.4%
	$\kappa = 0.7$	59.5%	71.3%
NB	$\kappa = 0.83$	76.4%	87.3%
	$\kappa = 0.7$	64.1%	80.2%
DT	$\kappa = 0.83$	78.4%	87.5%
	$\kappa = 0.7$	69.1%	83.4%
LR	$\kappa = 0.83$	61.3%	74.4%
	$\kappa = 0.7$	59.5%	87%
SVM	$\kappa = 0.83$	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%

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

Description Generation

Standard NLG pipeline (Reiter and Dale 2000):

- Step 1: The ML methods are used to generate the set of all pairwise spatial relations (SRs) for a given image.



$\{next\_to(bicycle_1, bicycle_2),$
$in\_front\_of(bicycle_1, person),$
$next\_to(bicycle_2, bicycle_1),$
$next\_to(bicycle_2, person),$
$behind(person, bicycle_1),$
$next\_to(person, bicycle_2)\}$

- Step 2: Content selection reduces this set to those that will be realised in the image description ordering them at the same time.

... Description Generation

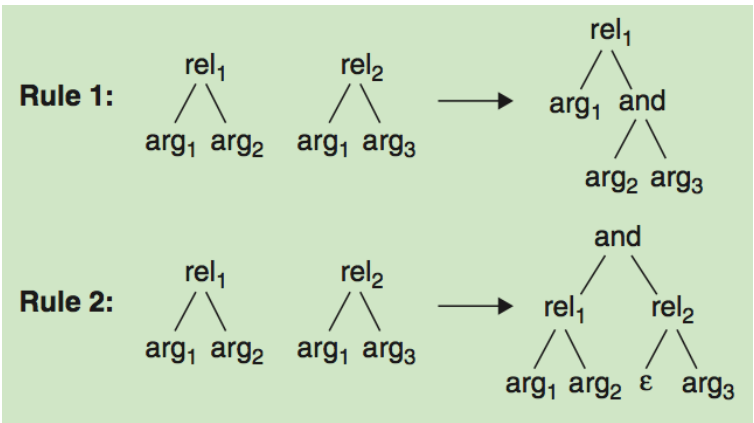
- Step 2: Content selection reduces this set to those that will be realised in the image description ordering them at the same time.

Random Chaining (**RC**)	$\{in\_front\_of(bicycle_1, person),\ next\_to(person, bicycle_2)\}$
Random Fanning (**RF**)	$\{next\_to(bicycle_1, bicycle_2),\ in\_front\_of(bicycle_1, person)\}$
Biggest-first Chaining (**BFC**)	$\{next\_to(bicycle_2, bicycle_1),\ in\_front\_of(bicycle_1, person)\}$
Biggest-first Fanning (**BFF**)	$\{next\_to(bicycle_2, bicycle_1),\ next\_to(bicycle_2, person)\}$
Human-centric Biggest-first Chaining (**HCBFC**)	$\{next\_to(person, bicycle_2)\},\ next\_to(bicycle_2, bicycle_1)\}$
Human-centric Biggest-first Fanning (**HCBFF**)	$\{next\_to(person, bicycle_2)\},\ behind(person, bicycle_1)\}$

- Step 3: Aggregation: repeated REs or SRs can be elided.

... Description Generation

- Step 3: Aggregation: repeated REs or SRs can be elided.



BFF :
$\{next\_to(bicycle_2, and(bicycle_1, person))\}$

HCBFF:
$\{and(next\_to(person, bicycle_2), behind(\epsilon, bicycle_1))\}$

- Step 4: Referring Expression Generation (REG) determines the type of referring expression (RE) for each object.
  REG rules based on previous work, (Belz and Varges 2007):

... Description Generation

- Step 5: Surface realisation outputs the final, fully realised image description.

Random Chaining:	a bicycle in front of a person who is next to a second bicycle
Random Fanning:	a bicycle next to a second bicycle and in front of a person
Biggest-first Chaining:	a bicycle next to a second bicycle which is in front of a person
Biggest-first Fanning:	a bicycle next to a second bicycle and a person
Human-centric Biggest-first Chaining:	a person behind a bicycle which is next to a second bicycle
Human-centric Biggest-first Fanning:	a person next to a bicycle and behind a second bicycle

Similar to pattern identified by both (Herskovits 1997) and (Huddleston and Pullum 2002) as the most frequent syntactic realisation of location expressions.

Human Evaluations - NLG

	Description correct?			
SR detection method	RF	LR	SVM	BL
Avg rank per NLG meth	1.4	2.2	3.4	3.2
Average rating	3.67	3.33	2.83	2.86

	Description complete?			
SR detection method	RF	LR	SVM	BL
Avg rank per NLG meth	1.7	2.5	2.9	2.9
Average rating	3.61	3.28	3.19	3.25

	Description natural?			
SR detection method	RF	LR	SVM	BL
Avg rank per NLG meth	2.7	3	2	1.8
Average rating	3.72	3.58	3.78	3.83

- Generate 144 descriptions from six new images, six native speakers.
- Significance difference : (RF, BL), (RF, SVM) for correctness.
- Chaining strategies are better in terms of naturalness;
- Biggest-first better in terms of correctness;
- Person-first is most useful for completeness.

Related Work

Step 2: Most closely related to our work is work by (Ramisa et al. 2015) and (Huasrrlimann and Bos 2016).
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.

Step 3: (Bernardi et.al.,2016) distinguish between methods that generate a description from scratch and those that stitch retrieved phrases together. Our work falls under the former.

Conclusions

- To what level of accuracy can spatial relations be determined by ML methods?
  Best method (random forests) received high human ratings and highest accuracy rate of 89.1%

- To what extent do human authors agree when determining spatial relations from still images?
  Average inter-annotator pairwise kappa scores of 0.67.

- What level of quality can be achieved with image descriptions anchored in spatial relations?
  The descriptions received good ratings.
  What about an application context?
  There is currently no standard for doing this.

Acknowledgements

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

A. Muscat and A. Belz, "Learning to generate descriptions of visual data anchored in spatial relations," *IEEE Computational Intelligence Magazine*, vol. 12, no. 3, pp. 29-42, Aug 2017.