Predicting Visual Spatial Relations in the Maltese Language

Adrian Muscat, Albert Gatt
adrian.muscat@um.edu.mt

Abstract

In this paper, the automatic detection of spatial prepositions between objects depicted in an image is studied. The ultimate aim is to incorporate the findings in a system that automatically describes images in a natural language. Whereas the explicit prediction of spatial relations in images has been previously studied in English and French, the work reported in this paper addresses relations in Maltese, an understudied language. A dataset consisting of images, spatial prepositions in Maltese, and a number of geometrical and language features is assembled from previous works. A number of predictive models are developed and the results are evaluated in terms of agreement with human-selected prepositions. The relative importance of the features in predicting the various relations is discussed and the paper concludes with a discussion on future work.

Introduction

The task of predicting relationships between objects depicted in an image is a fundamental problem in both Image Understanding (IM) and Natural Language Generation (NLG) and has useful applications in, for example, the development of assistive technology for the visually impaired, text-based querying of image databases, as well as education, such as in automated assessment.

Relationships between objects in images are often referred to as visual relations (Lu, Krishna, Bernstein, and Fei-Fei, 2016). Two important types of visual relations are actions, typically expressed using verbs (e.g. a person *riding* a horse; or person *kicks* ball) and spatial relations, typically expressed using prepositions (e.g. a bottle *on* a table; or a car *behind* a gate). As an example of the latter, consider the image on the left in Figure 1. The spatial relationship between the chair ('siġġu') and the sofa ('sufan') could be captured by prepositions such as 'ħdejn' ('near') or 'viċin' ('next to'). Presumably, the choice of preposition depends on both features of the language (here, Maltese) and the spatial configuration of the objects.

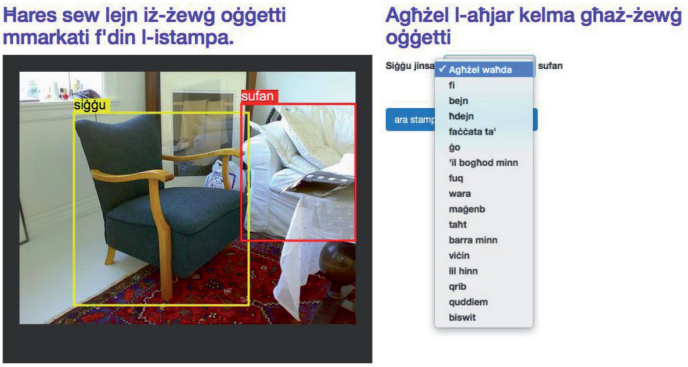


Figure 1. Screenshot of the crowdsourcing platform described in Section 2.0. The objects of interest are surrounded by bounding boxes and marked with labels siġġu ('chair') and sufan ('sofa). Users selected a preposition from the dropdown menu shown on the right.

The earliest attempt at detecting visual relationships considered the development of visual phrase detectors (Sadeghi & Farhadi, 2011). In this work a model is required for each unique phrase (where each phrase is a triplet <subject, object, relation>) and therefore enough examples per unique phrase are required during training. Whereas the visual phrase system works well with a small number of unique phrases, its complexity grows exponentially with the number of objects and relations. One solution is therefore to detect the objects and relations separately. In this regard, two methods have been studied; one based on manually engineered geometrical features (Belz & Muscat, 2015; Ramisa et al., 2015) and the second method based on features detected in deep convolutional neural networks (Lu et al.,, 2016; Yu et al., 2017). Furthermore, since these studies make use of machine learning models, various visual relations datasets have been collected or developed, some of which are publicly available. These include the ViSen dataset (Ramisa et al., 2015), based on prepositions obtained by parsing human-authored image descriptions in MScoco (Lin et al., 2014) and Flickr32k (Young et al., 2014; Plummer et al., 2015); and the Visual Relationship Dataset (VRD; Lu et al., 2016), for which explicit human annotations of objects and relations where crowd-sourced.

The explicit prediction of spatial relations in images has been studied in the English and French languages. The work reported in this paper addresses the prediction of spatial prepositions in Maltese, an understudied language. The ultimate aim is to incorporate the findings in a system that describes images in natural language (Maltese). Additionally, the results from this study can be useful in products such as Augmentative and Alternative Communication (AAC) apps for the Maltese language (Abela, 2018).

In this work, machine learning models to predict prepositions in Maltese are developed and are used to study to what extent spatial prepositions can be

automatically detected. A dataset for the Maltese language is assembled (section 2) and a total of thirty-one geometrical features and two language features are computed (section 3). For prediction, a number of predictive models – more specifically a baseline k-Nearest Neighbour (kNN) model, a Decision Tree (DT), a Support Vector Machine (SVM), a logistic regression (LR) and Random Forest (RF) – are developed. The results are evaluated in terms of agreement with human-selected relations (section 4). An ablation study yields insight into the usefulness of the various features, how the models discriminate between near synonymous prepositions and how the 2D geometrical features model space in a 3D world (section 5). The paper concludes with a discussion on future work (section 6).

The Maltese Spatial Prepositions Dataset

The dataset for Maltese spatial prepositions collected in Farrugia (2017) is the starting point. This dataset is based on the VOC2010 (Everingham et al., 2010) image dataset, which provides ground truth annotations for the object label, bounding box in pixels, *pose, difficult* and *occlusion*, the latter three being binary variables. New human-generated annotations that specify the spatial relations between pairs of objects in Maltese were added using a purposely built crowdsourcing platform (Farrugia, 2017; Muscat and Belz, 2017). A screenshot is shown in Fig. 1. On the platform, human annotators were shown the image with the bounding boxes highlighting the two objects. The annotators were then asked to choose a suitable preposition from a pre-defined list of spatial prepositions, as shown in Table 1. The final dataset consisted of 4332 labelled object pairs selected from 2399 unique images, which correspond to the same number of unique object pairs. The entries in the dataset therefore consist of <image, image size, bounding box for subject, bounding box for object, subject label, object label, preposition> and the average number of prepositions per object pair is 1.81. Table 1 gives the distribution of the output labels (prepositions) over the dataset. The average number of occurrences per preposition is 270.8 with a standard deviation of 185.2. Table 2 gives the distribution of the object categories over the dataset. The average number of occurrences per object is 433.2 with a standard deviation of 520.8.

Table 1: Number of occurrences per preposition

Preposition	Freq
Barra minn	54
bejn	7
biswit	253
Faċċata ta’	166
Fi	18
Fuq	515

Preposition	Freq
Ġo	44
Ħdejn	509
’Il boghod minn	252
Lil hinn	130
Maġenb	539

Preposition	Freq
qrib	255
Quddiem	488
Taħt	486
Viċin	324
Wara	292

Table 2: Number of occurrences per object category

Object label	Freq	Object label	Freq	Object label	Freq
Ajruplan	244	Qattus	566	Persuna	2613
Rota	273	Siġġu	386	Pjanta	202
Għasfur	347	Baqra	229	Nagħġa	248
Dgħajsa	265	Mejda	78	Sufan	327
Flixkun	213	Kelb	749	Tren	259
Xarabank	172	Żiemel	409	Televixin	269

The Vision and Language Features

The model input features are derived from both the vision and the language domain, assuming that for a given object pair, the most suitable preposition depends on the language as well as on the spatial configuration. The language features used are the object labels converted to one-hot vectors and it is left up to the model to compute or discover the distribution of the prepositions over the object labels. The geometrical features are computed from the sizes of the image and the bounding boxes and the union of the bounding boxes.

Table 3 lists the geometrical features considered, most of which are adopted from Ramisa et al.,(2015), Belz & Muscat (2015), and Muscat & Belz (2017). The features are organised in ten distinct groups. The features within each group are of the same type, however computed differently or using different inputs. For example group B represents area of each object normalised by either the area of the Union or the area of the Image, thus four different features in total. The groups are referred to in the following discussion on correlation between features.

Table 3: Geometrical features computed from image size, bounding boxes and union box.

#	Description of geometrical feature	Group
1, 2	**Aspect** Ratio of each Object	A
3...6	**Area** of each Object normalized by Union, Image area.	B
7...10	Area of Object **overlap** normalized by Union, Image, minimum and total area.	C
11..12	**Distance** between bounding box **centroids** normalized by union and image diagonal	D
13..16	**Diagonal** of each object normalised by image and union diagonal	E
17	Union diagonal normalised by image diagonal	F
18..20	**Euclidean** distance and **distance-size** ratio in between objects normalised by union, image.	G

21:24	**Ratio** of **object areas** and diagonals (max:min, trajector: landmark)	H
25..27	Trajector **position relative** to landmark (categorical and vector)	I
28..31	**Ratio of object limits** (minimum, maximum in x and y directions)	J

Intuitively, some of the features are correlated. The Pearson Correlation Coefficient was computed and Table 4 groups the features for the cases where the magnitude of the coefficient is >0.7 and, separately >0.5. As expected, a number of features show strong correlation within groups (since the difference is often the normaliser). For example normalised areas (B) and normalised diagonals (E) are strongly correlated. Surprisingly features in group I are not strongly correlated, probably due to different representations. However more importantly features in group J are not strongly correlated and only some of these are weakly correlated. Not surprisingly groups J and I are weakly correlated and also G, C and D.

Table 4: Correlated features within groups (single) and across groups (tuple)

Pearson Coefficient	Groups
>0.7	B; (B, E); C; G; H,
>0.5	(G, C, D); (J, I); (B, C, E, H); H; I; J;

The Machine Learning Models and Evaluation Metrics

The models considered in this study are (a) k-Nearest Neighbour (kNN), which is the baseline model, (c) Decision Tree (DT) model, (c) Support Vector Machine (SVM) model, (b) a logistic regression (LR) model and Random Forest (RF) model. The dataset is split into a train set, a development set, and a test set. The dataset was split on the basis of the unique image list (of size 2399) such that an image is unique to either the train, development or the test split and is not shared in between the splits. Additionally stratified sampling was used to ensure similar preposition distributions over all three splits. The development set was used to tune the hyperparameters using a grid search and was then concatenated to the training set for the final training of the models. The results quoted are those obtained from the test set.

The automatic detection of spatial prepositions is a multi-label problem, in the sense that there may be more than one preposition that is suitable for the object pair depicted in the image. In the Maltese spatial preposition dataset, half of the entries are annotated with two distinct prepositions, while a few are annotated with three distinct prepositions. This poses a problem in evaluation and most researchers use *recall@k* when evaluating the model, avoiding the use of *precision,* since the latter underestimates the model's accuracy. However *recall@k* results in the problem of

selecting the correct preposition out of *k* possibilities, which is an issue that has to be addressed for real-world applications. In Yu et al., (2017) the value of *k* is treated as a hyperparameter and in Belz and Muscat (2015), the annotators are asked to choose all suitable prepositions which allowed the computation of *precision* over the set of suitable prepositions in addition to *recall@k*. In this paper *recall@k*, (*k* = {1,2,3,4}), is the evaluation metric used in all experiments. Furthermore a multi-label set (note: this is not a complete set) was assembled by grouping all unique prepositions selected for the same object pair.

5.0 Results and Discussion

The primary results (recall@1), given in Table 5, were obtained after hyper-parameter optimisation. The 'All features' column are the results obtained when using all the language and vision features. The second column considers a selected set of features, following the correlation study in section 3, and results are given for both the single-label set and multi-label set. The kNN, LR and SVM models benefit from the selected feature set, while there is no change in the DT and RF model results. The latter observation is probably due to the feature selection methods built into these models. Furthermore, the SVM model benefited most from the feature selection. As expected the multi-label results are higher than the single label ones. The fifth and sixth column consider language and vision features separately. The vision features bring in more information compared to the language features. This result is expected since spatial prepositions are partly a function of spatial configuration. However other studies (Ramisa et al., 2015; Lu et al., 2016) report contrasting observations. These differences can however be attributed to skewed and long-tailed datasets. In contrast, the Maltese prepositions dataset is relatively balanced. Overall, the RF model obtains the highest score, closely followed by the LR and SVM models. The LR model fares best with the 'Vision Feature' set, which is probably due to the fact that the LR model fares better with real-valued features than the RF model, which is based on decision trees.

Table 5. Primary results

recall@1	All features	Selected Features		Language Features	Vision Features
		Single label	Multi-label		
kNN	31.3	32.0	40.6	29.3	31.1
LR	36.3	37.3	46.1	31.6	**35.2**
DT	31.3	31.3	41.3	27.1	31.3
SVM	32.3	36.4	44.3	28.9	34.3
RF	**37.6**	**37.6**	**46.5**	**32.0**	34.9

Table 6 expands the results for the 'selected feature' set to a set of *k* values, {1,2,3,4}. At *k*=4, scores reach the 84.1% mark. Probably, *k*=3 is a suitable value for this dataset, at which the score is 63.4% for the single-label. This warrants a deeper investigation into what the models are learning. To motivate such a discussion, Table 7 tabulates the scores per preposition for the best three models (RF, LR, SVM) and Fig. 2 depicts the confusion matrix as a Hinton diagram.

Table 6. Recall @k for k={1,2,3,4} for the single-label and multi-label sets

recall@k	1		2		3		4	
	Single-label	Multi-label	Single-label	Multi-label	Single-label	Multi-label	Single-label	Multi-label
kNN	32.0	40.6	47.2	60.2	57.5	70.7	66.6	79.5
LR	37.3	46.1	52.3	63.6	62.9	74.2	71.3	83.4
DT	31.3	41.3	44.7	55.9	55.8	67.6	63.9	77.0
SVM	36.4	44.3	51.8	63.0	61.3	74.7	69.8	**84.1**
RF	**37.6**	**46.5**	**53.6**	**66.5**	**63.4**	**77.0**	**71.9**	83.6

Considering recall@1 scores, the prepositions 'bejn' and 'ġo' are never recalled, while 'fi' is only recalled by the LR model, 'biswit' is only recalled by the RF model and RF fails to recall 'qrib'. While this failure can be attributed to the small number of examples for bejn (7), ġo (44) and fi (18), it is not so for 'qrib' and 'biswit', which are represented by 255 and 253 examples, respectively. However from the Hinton diagram, Fig.2, 'qrib' is being predicted mostly as 'maġenb', 'quddiem', 'viċin' and 'wara', which may all be plausible substitutes. Similarly, 'biswit' is exchanged for 'maġenb' and 'ħdejn' and to a lessor extent for 'quddiem' and 'faċċata ta'. The models score low for the preposition 'wara'. It is clear that the geometrical features do not act as a proxy for depth, which is missing in the feature set. On the other hand, 'quddiem' scores 38%, which is just higher than the average. In this case some features may be substituting for depth. The scores for the remaining prepositions, in general, can be improved by considering near-synonyms as indicated in the Hinton diagram. However there remains the question of what features are needed for fine-grained distinctions. Language priors may offer a solution. Table 7 highlights the highest score obtained per preposition. There is no one model that is clearly the best and the top scores are approximately equally shared by all three models.

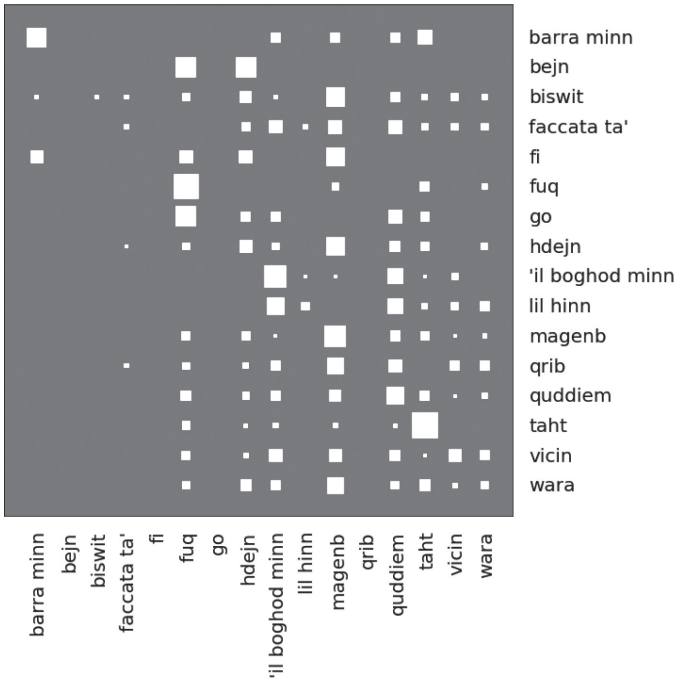
Table 7. Recall@k per preposition

Model	RF				SVM				LR			
recall@k	1	2	3	4	1	2	3	4	1	2	3	4
barra minn	44	67	67	67	**67**	67	67	67	56	56	56	67
bejn	0	0	0	0	0	0	0	0	0	0	0	0
biswit	**2**	2	22	31	0	0	8	20	0	0	24	39
faċċata ta'	3	3	23	32	**16**	16	16	26	10	26	42	48
fi	0	0	0	0	0	0	0	0	**20**	20	20	20
fuq	**73**	83	84	88	**73**	84	85	87	72	76	78	80
ġo	0	0	10	20	0	0	0	10	0	20	20	20
ħdejn	19	66	75	86	12	64	77	89	**22**	54	71	78
'il bogħod minn	**59**	76	86	88	**59**	75	82	86	53	80	82	86
lil hinn	8	17	33	38	0	12	25	29	**12**	21	25	38
maġenb	53	75	84	94	**57**	74	83	92	50	74	83	91
qrib	0	2	4	13	2	4	9	18	**7**	13	20	31
quddiem	36	57	82	86	**38**	60	83	90	24	46	65	88
taħt	81	88	90	92	77	88	88	92	**93**	93	95	95
viċin	**19**	26	36	68	6	9	36	60	8	23	40	49
wara	7	25	32	46	4	9	21	32	**9**	25	39	52

Conclusion and Future Work

This paper evaluated a supervised machine learning setup for the automatic detection of prepositions that describe the spatial relationships between object pairs in images. The experiments were carried out for the Maltese language and recall scores were used in the evaluation. A Maltese spatial prepositions dataset obtained from previous work was augmented with multi-labels and a set of geometrical features. Scores of 46.1% (recall@1) and 84.1% (recall@4) were recorded and an analysis of per preposition recall scores was carried out. The latter analysis motivated a discussion for future work. In the experiments, language features were represented by one-hot integer vectors. An alternative is to make use of distributed representations, such as word2vec (Mikolov et al., 2013), though these would need to be trained specifically for Maltese. Previous results indicate that distributed representations contribute mostly in the case of unseen object pairs, since intuitively objects in unseen object pairs are substituted by simliar ones. For example 'kelb' instead of 'qattus'. Finally, if prepositions can be grouped into near synonym sets on the basis of geometrical features (for example euclidean distance) and on co-occurence statistics (which is partly addressed with a multi-label set), then it should be possible to study whether the machine is learning the similarities among prepositions.

Figure 2. The confusion matrix for the Random Forest (RF) model. The size of the white squares is proportional to the percentage score. For example 'taħt' has a score of 81% on the diagonal.



References

Belz, A., Muscat, A., Aberton, M. and Benjelloun, S. (2015). Describing Spatial Relationships between Objects in Images in English and French. *Workshop on Vision and Language 2015 (VL'15): Vision and Language Integration Meets Cognitive Systems, EMNLP-2015*, Lisbon, Portugal.

Abela, S. (2018). Development of an Augmentative and Alternative Communication App for the Maltese Language. *Dissertation, BSc in Computer Engineering*.

Everingham, M., Van Gool, L., Williams, C., Winn, J. and Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International journal of computer vision*, 88(2), 303–338.

Farrugia, A. (2017). An Investigation into Spatial Relations in Maltese for Image Captioning. *Dissertation, BSc HLT, University of Malta.*

Lin, T. Y., Maire, M. Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., and Zitnick, C. L., (2014). Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Lu, C., Krishna, R., Bernstein, M. and Fei-Fei, L. (2016). Visual relationship detection with language priors. European Conference on Computer Vision.

Mikolov, T., Sutskever, I, Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems,*. 26(2), 3111-3119

Muscat, A. & Belz, A. (2017). Learning to Generate Descriptions of Visual Data Anchored in Spatial Relations. *IEEE Computational Intelligence Magazine*, 12(3), 29-42

Plummer, B., Wang, L., Cervantes, C., Caicedo, J., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image- to-sentence models. *CoRR*, abs/1505.04870.

Ramisa, A., Wang, J., Lu, Y., Dellandrea, E., Moreno-Noguer, F., and Gaizauskas, R. (2015). Combining geometric, textual and visual features for predicting prepositions in image descriptions. *EMNLP, Lisbon, Portugal, September 2015.*

Sadeghi, M.A. & Farhadi, A. (2011). Recognition using visual phrases. *Computer Vision and Pattern Recognition (CVPR) 2011, 1745-1752.*

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2, 67–78.

Yu, R., Li, A., Morariu, V. I. and Davis, L. S. (2017). Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation. *IEEE International Conference on Computer Vision (ICCV)*, 1068-1076.

Bio-notes

Prof. Adrian Muscat is an Associate Professor at the Department of Communications and Computer Engineering, University of Malta. His research interests are in pattern recognition models applied to Image Understanding, with special emphasis on semantic relations in images, and in Discrete Event Simulation and optimisation mainly applied to transport.

Dr Albert Gatt is Director of the Institute of Linguistics and Language Technology. His primary research interests are in Natural Language Processing, particularly the automatic generation of text from non-linguistic information, and the relationship between language and perceptual input, especially vision.