Human evaluation of robot-generated spatial descriptions

Simon Dobnik and Stephen G Pulman

Computing Laboratory, University of Oxford
Wolfson Building, Parks Road, Oxford OX1 3QD, United Kingdom
{simon.dobnik,stephen.pulman}@comlab.ox.ac.uk
http://www.clg.ox.ac.uk

Abstract. We describe a system where the semantics of spatial referential expressions have been automatically learned by finding mappings between symbolic natural language descriptions of the environment and non-symbolic representations from the sensory data of a mobile robot used for localisation and map building (SLAM). Although the success of learning can be measured by examining classifier performance on held-out data, this does not in itself guarantee that the descriptions generated will be natural and informative for a human observer. In this paper we describe the results of an evaluation of our embodied robotic system by human observers.

Key words: spatial expressions, machine learning, mobile robots, embodied multi-modal conversational agents, evaluation

1 Introduction

A conversational robot must be able to refer to and resolve references to the environment in which it is located with its human conversational partner. Mapping between the linguistic and non-linguistic representations is commonly performed by first identifying some parameters of the physical world on the basis of psychological evidence and then integrating them into customised functions [1, 2]. However, in a real robotic system which has been primarily built for tasks such as map building, localisation and navigation the information required by such models may not be readily available. Our approach attempts to use a simple model of space and motion that is available to a mobile robot and show that a mapping between its representations and highly abstract natural language spatial descriptions can be learned: that the robot can display a human-like performance in "understanding" and generating spatial descriptions or motion in new environments. In this paper we focus on the evaluation of the robot's performance from the point of view of a human conversational partner.

2 Learning spatial descriptions

Spatial descriptions may be about the identity of objects in a scene [3], about the spatial relations between the objects in a scene [4] or about the route that a moving

object can take in a scene [5]. The scene may be a small artificial town on a table top, a building with rooms or a real town. Our scenario is a larger room, a lab, which is constrained by walls, and which contains life-sized objects such as a chest, a box, a table, a pillar, a stack of tyres, a chair, a desk and shelves. The natural language descriptions that can be made in this environment belong to two categories: they can be descriptions of the robot's motion such as "You're going forward slowly" or, when the robot is stationary, descriptions of relations between the objects in the scene "The table is to the left of the chair". We consider descriptions of motion spatial description because their meaning is also relative to the environment in which they are used.

We use an ATRV-JR mobile robot designed by iRobot which runs middle-ware called MOOS.[1] The system runs an odometry component which provides information about the robot's motion such as its ⟨R-Heading⟩[2] and ⟨Speed⟩ and the SLAM localisation component [6] which uses a previously built 2-dimensional SLAM map to localise the robot. The objects were grounded on the map manually by taking a centre point of the cloud of points representing them, for example: chair ⟨0.6234,0.2132⟩ (⟨X⟩ and ⟨Y⟩). Our representation of the state of the robot and the space around it is thus extremely simple but the values of such representations are very accurate.

A group of four non-expert volunteers was invited to provide linguistic spatial descriptions of the robot and its environment. Each was first familiarised with the scene, the names of the objects and the different types of motion that the robot can produce. Then they were instructed to describe the motion and the location of objects from the perspective of the robot. This ensured that all directionals were used unambiguously from a single reference frame [7]. Two datasets were created. The linguistic descriptions in the first dataset (*Simple*) were made by a single describer and were restricted to a pre-defined small vocabulary (16 words) that appeared as choices on a computer screen. The second dataset (*All*) was created by all four participants who could use unrestricted vocabulary and sentences. Such descriptions show considerable lexical variation (46 words) but their syntactic structure is limited and in most cases similar to the examples above.[3] The two settings were intended to show the effects of subjectivity on the datasets and the models produced. To preserve the naturalness of the situation we used speech recognition (with some consequent noise in the language).

To turn MOOS log files (where both linguistic and non-linguistic information was recorded) to learning instances a few processing steps had to be performed: the locations of objects were expressed relative to the robot (rather than being global values relative to some random point where the robot has started) and their values were normalised (given the estimated size of the room or the maximum speed of the robot in the current session). This ensured that the models that were built could be later applied to new contexts. Words from natural language descriptions

[1] MOOS was designed by Paul M. Newman (Mobile Robotics Group, Department of Engineering, University of Oxford). We would like to thank him and members of his group for introducing us to mobile robotics.

[2] The attributes used in learning are marked with angled brackets.

[3] Complex descriptions such as "the chair is to the left of the table and behind the sofa" were simplified as two descriptions of relation.

were tagged to one of the categories ⟨Verb⟩, ⟨Direction⟩, ⟨Heading⟩, ⟨Manner⟩ and ⟨Relation⟩ which were also the target classes to be learned. The learning was accomplished with the Weka toolkit [8] which includes a range of offline supervised classifier implementations and a common framework to represent the data and evaluate the results. Each of the target linguistic classes was learned separately and not all attributes were used in each learning exercise. For example, to learn the category of ⟨Verb⟩ we only used the ⟨R-Heading⟩ and ⟨Speed⟩ attributes and to learn the category ⟨Relation⟩ we used the attributes ⟨LO_x⟩, ⟨LO_y⟩, ⟨REFO_x⟩ and ⟨REFO_y⟩ where LO stands for a located object and REFO stands for a reference object. Including all attributes resulted in a considerably lower classifier accuracy since may spurious relations were discovered.

The classifiers that were used in the human evaluation experiments described in the following sections were produced by the J48 learner which is the Weka's implementation of the ID3/C4.5 decision tree learner [9]. Their estimated accuracies obtained by a stratified 10-fold cross-validation are given in the last column of Table 1 for both *Simple* and *All* datasets. Note that these values are not the best values that we obtained. The accuracy of the motion categories was improved by a better method of combining a set of temporally sequential observations from the robotic log to instances. We also compared the performance of different machine learning methods on our datasets.

3 Evaluation by humans

The evaluation of machine learning classifiers by a stratified 10-fold cross-validation tests the degree to which the descriptions learned will generalise correctly to new cases. However, it does not tell us whether the models that are built will result in linguistic behaviour natural to humans. In order to know this we carried out a user study. We integrated the classifiers to a simple system that generates descriptions called *pDescriber*. This considers the current (normalised) values of the same attributes that were used in learning and predicts linguistic target classes. If the robot is moving, it generates descriptions of motion; if it is stationary it generates descriptions of object relations. The values of the predicted categories are applied to syntactic patterns such as "I'm ⟨Verb⟩ing" or "⟨LO⟩ is ⟨Relation⟩ the ⟨REFO⟩" which produce sentences that are subsequently pronounced by a speech synthesiser, for example "I'm reversing" and "You are behind the chair".

A new room was set up. Most of the objects were the same as in the data collection exercise but their placement was different. Five subjects were invited to the lab for approximately an hour each. None of them had participated in data collection. After being introduced the scene, they were explained that they should indicate whether they agree with the description that was generated by the robot given its current state and that of the environment. This gave us simple binary data. If they disagreed with the description, they had a chance to provide a better description. Note that the descriptions were not evaluated as utterances but per linguistic category. For example, for each utterance the system would query the evaluator whether "right" was a good word to describe the robot's heading in which it was moving or whether "to the left of" was a good description of the relation

between the chair and the table. The evaluators were also invited to make qualitative judgements about the appropriateness of the descriptions which we noted down. For approximately one half of the session the system used the classifiers built from the *Simple* dataset and the other half it used the classifiers built from the *All* dataset.

4 Evaluator-system agreement

The central part of Table 1 shows the measured accuracies from each evaluator per category. As explained in the previous section, accuracy is measured as evaluator agreement with the system on the choice of description. The penultimate column contains the accuracies when all evaluators are considered together. The last column contains the estimated accuracies of the classifier that the system was using to produce these descriptions. The table is split into two parts each containing the results from one configuration of the system (*J48-Simple* and *J48-All*).

Table 1. System performance *vs.* classifier performance

Category			Evaluators						Classifier
			a	b	c	d	e	All	J48
									Simple
Motion		$n =$	36	17	14	2	21	90	–
	Verb		100	**88.24**	100	100	95.24	96.67	89.02
	Direction		100	**76.47**	100	100	100	95.56	87.80
	Heading		100	**82.35**	100	100	**85.71**	**93.33**	97.56
	Manner		100	82.35	100	100	100	96.67	70.73
Relation		$n =$	65	23	19	53	22	182	–
	Relation		**67.69**	**65.22**	**68.42**	**66.04**	**59.09**	**65.93**	75.90
									All
Motion		$n =$	53	22	53	7	41	176	–
	Verb		96.23	77.27	88.68	100	100	92.61	48.22
	Direction		96.23	72.73	92.45	100	100	93.18	55.68
	Heading		98.11	68.18	92.45	100	95.12	92.05	60.77
	Manner		100	72.73	98.11	100	100	96.02	54.70
Relation		$n =$	66	28	72	110	58	334	–
	Relation		72.73	**57.14**	**44.44**	70.00	**43.10**	**59.28**	69.12

How do the results from the evaluation of the system by humans and the evaluation of the underlying classifiers compare? The classifier accuracies are the average accuracies obtained through a 10-fold cross-validation. In the human evaluation of the system the accuracy is determined on an independent test set. In both cases the reported accuracy is the ratio between the number of agreements with the system or correct classifications over the total number of considered testing instances. There is a slight difference between the two situations in how a positive match is made. In cross-validation the correct value of the class is pre-defined and hidden from the classifier and this is matched with the predicted class. In human evaluation an evaluator hears the generated description before they give their evaluation.

This description is the one that is predicted by a classifier given the attributes representing the robot's current internal state. In this respect it is possible that the system unavoidably biases the evaluator, since other possible descriptions are never produced. Furthermore, when evaluating the system in this way, the observers are not always just evaluating the classifiers. For example, when generating descriptions of object relations the located and the reference objects are chosen at random and the classifier is used to predict the best relation between the two. The description may be evaluated as unsuitable because of an unfortunate choice of objects even though the spatial relation between them is correct.

A quick look at the table reveals that the evaluators considered the system performance to be better than the accuracy of the underlying classifiers on most classes of motion descriptions (J48-Simple classifier: $\bar{x} = 86.28\%$, J48-Simple evaluators: $\bar{x} = 95.56\%$; J48-All classifier $\bar{x} = 54.84\%$, J48-All evaluators: 93.47%). To make the comparison easier we mark the values where the *opposite* is true, when the system is evaluated to perform worse than its classifiers, in bold. The evaluator accuracies are quite similar across categories, even for the ⟨Manner⟩ category on which the the classifiers perform less well than others. This is more the case with the *Simple* configuration than *All*. On the contrary, the system was considered to perform less well than its classifiers on the ⟨Relation⟩ category by approximately 10% in both cases (J48-Simple classifier: 75.90%, J48-Simple evaluators: 65.93%; J48-All classifier: 69.12%, J48-All evaluators: 59.28%).

The scores from evaluator b are lower than those from other evaluators, particularly on the motion classes and for the *J48-Simple* configuration. The numbers in lines starting with $n =$ indicate the size of the evaluation sample. Although the number of descriptions that the robot generated was not strictly controlled, a reasonable sample was obtained for each evaluator. The only exception is evaluator d who evaluated only a small number of descriptions of motion but on the other hand considered more descriptions of object relations.

An explanation why the evaluators consider the system to perform better than its underlying classifiers on the motion categories but not on the relation category could be that motion categories contain words that are less semantically restrictive. For example, the category ⟨Verb⟩ contains words such as "going", "moving" and "continuing" which all have a very similar reference for a human but not for a machine learner where the attribute values are assumed to be discrete. Consequently, an evaluator may accept such alternative. The categories ⟨Direction⟩, ⟨Heading⟩ and ⟨Manner⟩ contain words with clearer semantic divisions but they all also contain a word "none" which was assigned as a value of each category in machine learning dataset if a word for that category was not present. The meaning of this word is ambiguous between a default meaning and an anaphoric meaning. For the ⟨Direction⟩ category "none" has the same meaning as "straight". However, it can also refer anaphorically to the previously generated description of direction if this has not changed.

Another explanation why the results are different for descriptions of motion and object relations is that learning and generating of the latter is more complex. It could be that our learning and generation models for descriptions of object relations capture human knowledge less well than the models for description of motion. We discuss some qualitative evidence for this in Section 6.

5 Inter-evaluator agreement

Agreement between individual evaluators demonstrates that the system has not been tuned to the vocabulary of the describers who provided descriptions for machine learning. Disagreement may be informative too: if evaluators collectively disagree it means that the generation task is not subjective, that there exists a consensus on what is a good description in a particular context and what is not.

Unfortunately, inter-evaluator agreement cannot be established directly, for example by calculating a κ coefficient, because not all evaluators evaluated the same set of items. The evaluators considered a closed set of words produced by the system. We can expect that the agreement of a single evaluator with the system will not be identical on every word that it produces. Some words are more difficult to learn than others. If so, the difference in the ratings for words should be consistent across evaluators. According to our model of agreement, an evaluator agrees with other evaluators if their accuracy scores per word correlate with the mean of accuracy scores per word of everyone else.

Table 2. Agreement of each evaluator with the rest of the group

Configuration	a:rest	b:rest	c:rest	d:rest	e:rest	Mean
J48-Simple	0.824**	0.382 ns	0.787**	0.907**	0.636*	0.707
J48-All	0.504*	0.048 ns	0.635**	0.756**	0.662**	0.521

Table 2 shows the Pearson's correlation coefficients r_{xy} obtained at each fold of correlation for both sets of classifiers. The last column contains the average correlation coefficient. The asterisks indicate the statistical significance levels of the coefficients obtained by a two-tailed t-test.[4]

We can see that except for the evaluator b there exists a moderate to high correlation between the scores of an individual evaluator and the mean scores of the rest of the group. The average correlation coefficient for the *J48-Simple* configuration is greater (0.707) than the average correlation coefficient for the *J48-All* configuration (0.521). All correlation coefficients, except in the case of evaluator b are statistically significant at the level $\alpha = 0.05$. In sum, apart from evaluator b there is a considerable consensus between the remaining four evaluators on the performance of the system. Thus, it has captured some universal knowledge.

6 Qualitative evaluation

Descriptive observations made by the evaluators are useful because they point out facts about spatial cognition and the shortcomings of the system that can be further improved [10, 11].

[4] * indicates that the correlation is significant at the 0.05 level, and ** indicate that it is significant at the 0.01 level. "ns" indicates that the correlation is not significant.

Ambiguity of heading and direction. The descriptions such as "left" and "right" are ambiguous when used to refer to motion. "Moving right" can mean moving forward with a heading in the clockwise direction. It can also mean making a sudden turn to the region that is to the right of the current location and then moving straight in that direction. Similarly, "moving backward" can mean that the robot is moving in the direction that is behind its back (reversing) or that it has reversed but is now moving forward in the direction that was previously behind its back. The second of each description pair is more complex and to learn such descriptions the learner would have to abstract over a set of actions rather than over physical descriptions of environment. Since while performing the second action "right" and "backward" may refer to the same state of the robot as "straight" and "forward" in our model, the robot is likely to over-generate such descriptions in cases where the first action was not performed.

Object shape. The SLAM map used in our model does not contain abstract representations of objects but only clouds of points. Each object is represented by a centre point. While this works reasonably well for objects that square-shaped, difficulties arise with objects that are markedly different in one dimensions such as "the wall" and "the barrier".

Switching the reference frame. Although evaluators were told that the descriptions generated with the reference frame fixed on the robot or from "its perspective", it was very easy for them to switch from this relative reference frame to the intrinsic reference frame fixed on the reference object. Firstly, it became apparent that some switches to the intrinsic reference frame have been learned from the training data and such descriptions appeared appropriate in the current context. In this case, the majority of evaluators would accept such descriptions although they should not do so according to our instructions. Secondly, properties of some objects invite human describers or observers to use intrinsic rather than the relative reference frame. This is true for objects that are larger than describer (walls, barriers and cupboards), have an identifiable front and are animate (another robot). Only the intrinsic reference frame is possible when the robot describes its own location and consequently cannot serve as a reference object. "I'm in front of the chair" unambiguously means that the robot is located in the region around and orientated by the seating area of the chair. Note that the reference frame also applies to the projective descriptions of motion.

Reference to objects outside the robot's field of vision. There was a disagreement between the evaluators whether descriptions that cannot be "seen" by the robot are appropriate or not. Technically, "the vision field" of the robot is much greater than that of a human observer - it is the entire SLAM map which represents its mental map. Humans also use mental maps to imagine configurations of objects for tasks such as navigation and therefore descriptions of objects not in the visual focus of the describer may not be completely unnatural. In fact, particularly disapproved were those descriptions where only one of the objects was "visible".

Non-optimal choice of objects. The classifiers always attempt to predict the best description of relation between two objects and may do so but the description may

be judged inappropriate because of an unfortunate selection of objects. The latter can be accomplished by a contextual model which our system does not implement. Given that we are primarily interested in spatial relations itself the choice of objects at random seems to be reasonable. Some evaluators were more sympathetic to such descriptions than others. However, they all agreed that descriptions where the lack of object salience was coupled with the lack of the vision field salience were quite unacceptable.

7 Conclusion

Although our classifiers use a relatively simple (topological) representation of space primarily intended for localisation of a mobile robot we can conclude that they work surprisingly well in practice in replicating human linguistic competence. They fall short sometimes because they do not have access to non-topological information such as object shape, reference frame, discourse structure for modelling salience and world knowledge about the objects. Such data must be provided from other sources.

References

1. Regier, T., Carlson, L.A.: Grounding spatial language in perception: an empirical and computational investigation. Journal of Experimental Psychology: General 130(2), 273–298 (2001)
2. Coventry, K.R., Cangelosi, A., et al.: Spatial prepositions and vague quantifiers: implementing the functional geometric framework. In: Freksa, C., Knauff, M., et al. (eds.) Spatial Cognition, vol. IV, pp. 98–110. (2005)
3. Zender, H., Martínez-Mozos, O., et al.: Conceptual spatial representations for indoor mobile robots. Robotics and Autonomous Systems 56(6), 493–502 (2008)
4. Roy, D.K.: Learning visually-grounded words and syntax for a scene description task. Computer speech and language 16(3), 353–385 (2002)
5. Lauria, S., Kyriacou, T., et al.: Converting natural language route instructions into robot-executable procedures. In: Proceedings of Roman'02. pp. 223–228. (2002)
6. Bosse, M., Zlot, R.: Map matching and data association for large-scale two-dimensional laser scan-based SLAM. IJRR 27(6), 667–691 (2008)
7. Steels, L., Loetzsch, M.: Perspective alignment and spatial language. In: Coventry, K.R., Tenbrink, T., Bateman, J. (eds.) Spatial language and dialogue, Explorations in Language and Space, vol. 3, pp. 70–88. OUP (2009)
8. Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques. Morgan Kaufmann, 2nd edn. (2005)
9. Quinlan, J.: C4.5: programs for machine learning. Morgan Kaufmann (1993)
10. MacMahon, M., Stankiewicz, B., Kuipers, B.: Walk the talk: Connecting language, knowledge, and action in route instructions. In: Proceedings of AAAI-2006. pp. 1475–1482. (2006)
11. Moratz, R., Tenbrink, T.: Spatial reference in linguistic human-robot interaction: iterative, empirically supported development of a model of projective relations. Spatial Cognition & Computation 6(1), 63–107 (2009)