Applying spatial knowledge from a scene description task to question answering

Simon Dobnik

Department of Philosophy, Linguistics and Theory of Science
Box 200, 405 30 Göteborg, Sweden
simon@dobnik.net
http://flov.gu.se

This presentation extends our previous work in which we build and test a mobile robot which learns grounded semantic representations of spatial concepts from human descriptions and its own perception through sensors of a room containing real objects. The learning is performed offline as a machine learning classification task. In [1] we show that the learning of spatial concepts is successful when the classifiers are evaluated. In [2] we argue that classifier evaluation is not enough to show that the robot acquired human-like spatial knowledge which generalises to new spatial configurations. We therefore integrated the classifiers in our own NL generation system (*pDescriber*) which produces grounded descriptions of spatial scenes such as "The table is to the left of the chair" and allows humans to observe the acquired spatial knowledge. The discourse setting in which these descriptions are made is identical to the one in which they were sampled before they were learned. In this contribution we examine whether we can use data-sets and classifiers from the scene description task to answer questions that (A) locate objects: "Where is the table?" - "The table is to the left of the chair"; (B) confirming object description: "Is the table to the left of the chair?" - "No, the table is near the chair."; (C) find objects: "What is to the left of the chair?" - "The pillars, the tyres and the wall are to the left of the chair"; and (D) reference objects: "What is the chair to the left of?" - "The chair is to the left of the table, the desk and the wall". We see the task as an experimentally constrained form of dialogue which contains only two dialogue acts (information request and answer) which are always performed by the same illocutionary partner: a human directs questions to the robot. Since the dialogue is situated both spatially and in discourse we do expect to find effects of semantic coordination of human observers when interpreting the robot's responses.

Generating question answers (*pDialogue*) requires more steps than generating descriptions and hence more factors may influence the evaluation of spatial knowledge. User utterances must be interpreted as questions and their content must be matched against dialogue rules which specify how to answer them. Most dialogue rules require an application of ML classifiers that take linguistic descriptions and predict perceptual properties rather than reverse (*pDescriber*). The classification tells us what state of perception corresponds to a description. The dialogue rules must then issue commands that bring the robot to this state or find a configuration of objects that holds in the state. The resulting knowledge is used to generate natural language sentences.

The system was individually evaluated by 13 non-expert volunteers in a room environment different from that used in data collection for ML. Each evaluator considered the robot's answers to 55 questions which were scripted and were automatically asked by the evaluation software at four distinct room locations (L1 to L4). This ensured that various spatial and linguistic conditions were covered. The evaluators' task was to indicate whether each robot's answer is an intuitive or natural description on a scale from 1 (bad) to 5 (best). Each run took between 45 to 60 minutes to complete. We estimated evaluator agreement by calculating Pearson's correlation coefficient between the scores of each evaluator per particular question-answer pair and the mean of such scores over all other evaluators. The overall agreement of 0.583 (the mean of correlation coefficients from all 13 folds) shows that there is a considerable consensus between the evaluators on the performance of the system.

To estimate the accuracy of the system the evaluator scores were normalised to values between 0 and 1 (1=0, 2=0.25, 3=0.5, 4=0.75, 5=1) and summed. The accuracy per question type is as follows: A - 43.5%, B - 54.2%, C - 54.7%, D 56.9% and mean - 52.3%. The steps involved in answering questions A are identical to generating a description in *pDescriber* (59.3%) [2] – one or two objects are selected at random and the relation between them is classified – but the estimated performance of *pDialogue* on questions A is lower by 15.8%. The result suggests that a new discourse setting affects the interpretation of spatial descriptions. When the system generates a description on its own, a human hearer understands it as a general statement about the scene that both are observing. However, when an agent in conversation asks a question, they expect an informative and relevant answer which helps them to interpret the scene. Choosing a salient reference object is particularly important. Objects can be salient in their visual properties (visual-salience) or through being previously discussed and located in dialogue (discourse salience). The modelling of both kinds of salience is an object of our future investigations.

We also tested two other properties affecting the semantics of spatial descriptions in a situated discourse. The difference in evaluation scores for questions-answer pairs that involved (a) objects that were in the robot's visual field (L1 and L2) and those that were not (L3) is statistically significant (t-test: $a > b$; $\alpha = 2P = 0.000$). The interlocutors expect the robot to change its orientation towards objects referred to in questions and answers. Secondly, the difference in evaluation scores (a) where the spatial description in questions was unambiguous in terms of the reference frame (C at L1: "What is to the left of you?" – intrinsic) and (b) where a question could be answered using an alternative reference frame ("What is to the left of the chair?) is not statistically significant (t-test: $a = b$; $\alpha = 2P = 0.61 > 0.05$.). This shows that human observers align to the reference frame chosen by the robot (relative to itself) and do not insist on changing it.

References

1. Dobnik, S.: Learning spatial referential words with mobile robots. In: Proceedings of the 9th Annual CLUK Research colloquim. The Open University, Milton Keynes, United Kingdom (2006)

2. Dobnik, S., Pulman, S.: Human evaluation of robot-generated spatial descriptions. In: Proceedings of the Workshop on Computational Models of Spatial Language Interpretation (CoSLI). Portland, Oregon, USA (2010)