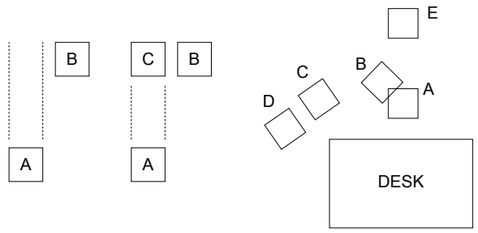
Learning spatial referential words with mobile robots

Simon Dobnik

Computational Linguistics Group, Oxford University

Introduction

The meaning of spatial words can only be evaluated by establishing a reference to the properties of the environment in which the word is used.



- Is B behind A? - Which chair is in front of the desk?

- How fast is fast?

We need to evaluate the size of the scene, the perspective at which it is viewed, typical behaviour and properties of objects, and the configuration of other objects in the scene.

Aim

Encoding semantic rules that define these expressions by hand is hard: we describe physical world with precise interval measures as opposed to NL expressions which are ambiguous and refer to discrete regions.

We can **automatically learn the rules** (WEKA: Witten and Frank, 2005) from descriptions of scenes made by human commentators in a mobile robotics setting and information that a mobile robot has about itself and its environment (SLAM: Dissanayake et al., 2001).

The rules that are learned offline should be embedded in the system that drives a mobile robot so that this can produce new descriptions of new scenes, answer questions about the scenes or accept motion commands.

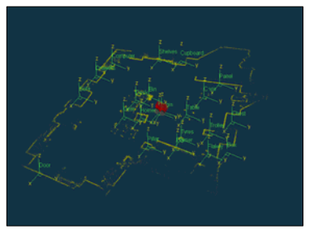
If the robot can use such expressions in a manner which is natural to a human observer, then we can be sure that we captured something important about their semantics.

Method

We collect a dataset containing:

- NL descriptions made by human observers: *You're going forward slowly. Now you're turning right. The chair is to the left of you. The table is further away than the chair.*
- numeric descriptions about the location and the state of the robot (SLAM) obtained from its sensors.

A robot in the environment containing objects such as a table, a chest and a pillar. The robot changed its position around the room displaying various kinds of motion and allowing for different configurations of objects. Human describers described the scenes from the robot's perspective.



A representation of the environment as internalised by the robot using SLAM (Simultaneous Localisation and Mapping) software (Newman, 2007). The environment is represented as clouds of points. Objects are not discrete entities and they have been grounded (i.e. given names) manually.

- Information was structured to 'instances' – lists of attribute values (e.g. *speed*, *verb*, *relation*) that are fed to machine learning algorithms.

```
0.001, 0.234, turning, left, none, none
0.432, 0.002, going, right, forward, fast
0.456, 0.234, 0.001, 0.221, right
0.134, 0.342, 0.154, 0.581, in-front-of
```

- Machine learning (Naïve Bayes and J48 Decision Trees) tries to find a theory about this data to account for all the instances in the training set.
- ML is performed offline. The knowledge is integrated (with some extra bits!) into two NLP applications that run on the robot (*pDescriber* and *pDialogue)*: one generates new descriptions and the other answers questions about the scene or performs motion commands.

Results

The accuracy of classifiers measured by 10-fold cross-validation):

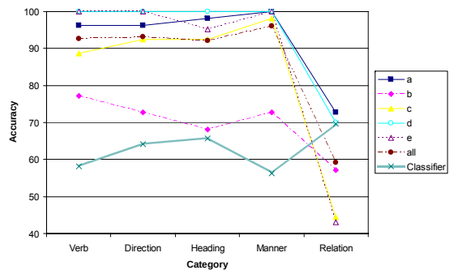
Classifier	*Verb*	*Direct.*	*Head.*	*Manner*	*Relat.*
Decision Trees	65.4%	73.4%	73.1%	65.4%	74.9%
Naive Bayes	62.1%	70.1%	67.8%	57.4%	77.3%

The words that were learned are:

Attribute	
Verb	moving, turning, stopped, going, reversing, creeping, continuing, edging
Direction	none, backward, forward, on the spot
Heading	none, left, right, clockwise, straight, anticlockwise, 180 degrees, around, in a straight line, hard, straight ahead
Manner	none, slowly, gently, moderately, fast, rapidly, tightly, imperceptibly, at a walking pace, quickly
Relation	to the right of, in front of, to the left of, behind, facing, far from, close to, opposite of, parallel to, near, after, next

There are 4 to 13 words per category. Thus, if a classifier were choosing words at random (baseline) the accuracy would be between 25% (max.) and 7.7% (min).

The performance of the system was also evaluated live on the robot in a similar yet different environment from which the learning data was collected. Humans were asked whether they agree or not with the generated description. If they did not agree they were given a chance to enter a better one.



For motion categories, the valuators a, b, c, d *and* e *judged the performance of the live system to be better than the underlying classifier. For the* Relation *category, their judgements were slightly lower than the performance of the classifier.*

Possible explanations

- Motion words are semantically less restrictive (i.e. ambiguous) than words relating objects and thus choosing an incorrect word (from the classifier point of view) does not mean that the description is bad.
- The grounding model of object relations may require inclusion of other features of objects than just topological ones (e.g. properties of objects and preceding discourse).
- The non-automatic knowledge that was used to create instances and integrate the classifiers into a language generation system may not be ideal.
- The training data could involve systematic errors or describer variation which were incorporated to the models.

The consistency of evaluators' judgements was tested by finding a statistical correlation between the judgements of each evaluator and the rest of the group. All correlation coefficients (except for evaluator *b*) are statistically significant at the level $\alpha = 0.05$.

Conclusion

We show a full cycle of how descriptions are learned from human comments and the properties of the environment internalised by a mobile robot and later used by the system to describe the environment back to humans. Future work includes enriching the learning feature sets and exploring whether the integration of the localisation and linguistic subsystems can bring mutual benefits.

Contact:

Centre for Linguistics and Philology
Walton Street, Oxford, OX1 2HG, UK

Ph: +44 1865 280 401
Email: simon.dobnik@clg.ox.ac.uk
Web: http://users.ox.ac.uk/~lady0641

Acknowledgements

Many thanks to Stephen Pulman at Computing Laboratory at Oxford University for discussion and ideas. Special thanks also to Paul Newman and the Mobile Robotics Group at Oxford University for help on SLAM.