Reference Object Choice in Spatial Language: Machine and Human Models

Michael Barclay

16th August 2010

Submitted by Michael John Barclay, to the University of Exeter as a thesis for the degree of Doctor of Philosophy by Research in Computer Science, August 2010.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

For my parents, who have always been my heroes

I had no idea when I started this research how important supervisors are and how different they can be in terms of the support and advice they give. All those of us who are supervised by Dr Galton know before long how fortunate we are. There is no one more generous with his time and ideas, or more constructive and supportive with his guidance. To Antony my sincerest thanks for all this, these last years have been the most enjoyable of my working life to date. Thanks also to Jonathan, who, though he only got the post of second supervisor recently, has provided some invaluable pointers and comments. I also need to thank the rest of the research group, Larry, Zena and Max for their help and support and again for making it such an enjoyable time. Last but not least all my love and thanks to Sadie, Tommo and Kitty who have been very forbearing and have put up with a lot of distractedness and not a lot of money for longer than they deserved.

Abstract

The thesis underpinning this study is as follows; it is possible to build machine models that are indistinguishable from the mental models used by humans to generate language to describe their environment. This is to say that the machine model should perform in such a way that a human listener could not discern whether a description of a scene was generated by a human or by the machine model.

Many linguistic processes are used to generate even simple scene descriptions and developing machine models of all of them is beyond the scope of this study. The goal of this study is, therefore, to model a sufficient part of the scene description process, operating in a sufficiently realistic environment, so that the likelihood of being able to build machine models of the remaining processes, operating in the real world, can be established.

The relatively under-researched process of reference object selection is chosen as the focus of this study. A reference object is, for instance, the 'table' in the phrase "The flowers are on the table". This study demonstrates that the reference selection process is of similar complexity to others involved in generating scene descriptions which include: assigning prepositions, selecting reference frames and disambiguating objects (usually termed 'generating referring expressions'). The secondary thesis of this study is therefore; it is possible to build a machine model that is indistinguishable from the mental models used by humans in selecting reference objects. Most of the practical work in the study is aimed at establishing this.

An environment sufficiently near to the real-world for the machine models to operate on is developed as part of this study. It consists of a series of 3-dimensional scenes containing multiple objects that are recognisable to humans and 'readable' by the machine models. The rationale for this approach is discussed. The performance of human subjects in describing this environment is evaluated, and measures by which the human performance can be compared to the performance of the machine models are discussed.

The machine models used in the study are variants on Bayesian networks. A new approach to learning the structure of a subset of Bayesian networks is presented. Simple existing Bayesian classifiers such as naive or tree augmented naive networks did not perform sufficiently well. A significant result of this study is that useful machine models for reference object choice are of such complexity that a machine learning approach is required. Earlier proposals based on sum-of weighted-factors or similar constructions will not produce satisfactory models.

Two differently derived sets of variables are used and compared in this study. Firstly variables derived from the basic geometry of the scene and the properties of objects are used. Models built from these variables match the choice of reference of a group of humans some 73% of the time, as compared with 90% for the median human subject. Secondly variables derived from 'ray casting' the scene are used. Ray cast variables performed much worse than anticipated, suggesting that humans use object knowledge as well as immediate perception in the reference choice task. Models combining geometric and ray-cast variables match the choice of reference of the group of humans some 76% of the time. Although niether of these

machine models are likely to be indistinguishable from a human, the reference choices are rarely, if ever, entirely ridiculous.

A secondary goal of the study is to contribute to the understanding of the process by which humans select reference objects. Several statistically significant results concerning the necessary complexity of the human models and the nature of the variables within them are established.

Problems that remain with both the representation of the near-real-world environment and the Bayesian models and variables used within them are detailed. While these problems cast some doubt on the results it is argued that solving these problems is possible and would, on balance, lead to improved performance of the machine models. This further supports the assertion that machine models producing reference choices indistinguishable from those of humans are possible.