Describing Time-Varying Data

Sarah Boyd
Microsoft Research Institute
Macquarie University
Sydney, Australia 2109
sarahb@mri.mq.edu.au

1 Introduction

Automatically generating textual descriptions of numerical data is a particularly useful task with the explosion of accessible online information. A textual description allows people to absorb key features more easily, especially when there are large amounts of data involved. Some work has been done in automatically generating text from data; e.g. Mittal et al. [1995] and Robin and McKeown [1996]; in these cases, apart from simple runs in the data, more complicated temporal features of the data have not been described. However, a great deal of data is time-varying in nature: e.g. stockmarket prices, government figures, patient medical records, computer network statistics and weather data, and phenomena such as trends and dramatic changes in the data are of interest.

2 Content Determination

Working out what to say in Natural Language Generation is an important problem. When generating descriptions of time-varying data, as well as commenting on the standard properties of max, min and mean, what is of interest is the behaviour of the variable over time such as dramatic changes, trends and degree of variability in the data. For example, in an expert description of monthly temperature, the "general decrease" over the whole month was mentioned along with "a steady recovery" in the second week of the month and a "general downward trend" in the last two weeks of the month. Identifying trends like these at different scales of analysis is a difficult problem. This research will address the content selection problem when generating descriptions of time-varying data using the mathematical techniques of wavelets and scale space theory. Wavelets and scale space theory can analyse a signal at multiple granularities and extract the perceptually salient features in the data. These perceptually salient features are worth including in a description.

3 Preliminary Experiments

A prototype system has been implemented in Matlab which extracts the four most significant dramatic changes in a time-varying signal using the multiscale edge detector wavelet described in Mallat and Zhong [1992] and scale space theory [Witkin, 1983]. This system has been tested on ten sets of monthly temperature data and the forty features automatically extracted were compared to handwritten descriptions of the same datasets. 18/19 features chosen by the expert were also extracted automatically.

4 Future Work

Future work is planned in three broad areas:

l)language analysis — analysing real texts to confirm the types of phenomena described and how they are realised.

2) *inferencing theory* — determining appropriate techniques to infer descriptions at appropriate scales of detail for different datasets and comparing these chosen techniques with existing methods such as statistical modelling.

3)*system development* — implementing the techniques and integrating with real language output.

References

[Robin and McKeown, 1996] Robin J. and K. McKeown Empirically Designing and Evaluating a New Revision-Based Model for Summary Generation. *Artificial Intelligence* Vol 85 1996.

[Mittal *et* al., 1995] Mittal, V. O., S. Roth, J. D. Moore, J. Mattis and G. Carenini Generating Explanatory Captions for Information Graphics. *Proceedings IJCAI 1995.*

[Witkin, 1983] Witkin, A.P. Scale Space Filtering. *Proceedings IJCAI 1988.*