Atlas.txt: Linking Geo-referenced Data to Text for NLG

Kavita E. Thomas and Somayajulu Sripada
Department of Computing Science
University of Aberdeen
Aberdeen, UK
{tkavita,ssripada}@csd.abdn.ac.uk

Abstract

Geo-referenced data which are often communicated via maps are inaccessible to the visually impaired population. We summarise existing approaches to improving accessibility of geo-referenced data and present the Atlas.txt project which aims to produce textual summaries of such data which can be read out via a screenreader. We outline issues involved in generating descriptions of geo-referenced data and present initial work on content determination based on knowledge acquisition from both parallel corpus analysis and input from visually impaired people. In our corpus analysis we build an ontology containing abstract representations of expert-written sentences which we associate with macros containing sequences of data analysis methods. This helps us to identify which data analysis methods need to be applied to generate text from data.

1 Introduction

Currently there is a plethora of geo-referenced statistical data such as census data publicly accessible on the World Wide Web. Much of this data is provided by governmental organisations like National Statistics which provides the UK census 2001 data online[1]. Such organisations are legally required to make this data accessible to user groups with visual impairments. However the majority of this data is currently displayed in the form of chloropleth maps which shade regions according to the mean value of a given variable in that region, as is typically exemplified in Figure 1[2] and is described by the expert authored text:

"Of Scotland's Census Day population of 5,062,011, close to 2 million people live in large urban areas". Such maps of geo-referenced data aid sighted users to quickly comprehend spatial patterns and trends, e.g., that the Scottish population is largely concentrated among the urban areas in Figure 1. However, visually impaired users have to listen to each of the individual frequency values corresponding to the census area units before comprehending the spatial distribution of population. The Atlas.txt project aims to improve access to such geo-referenced data using NLG technology by determining the salient features which best explain the data and communicating this relevant information as text summaries like the one above, which can then be read out via a screen-reader. The project initially addresses predominantly census data, but aims to develop techniques that can be applied on other geo-referenced data such as land-use data.

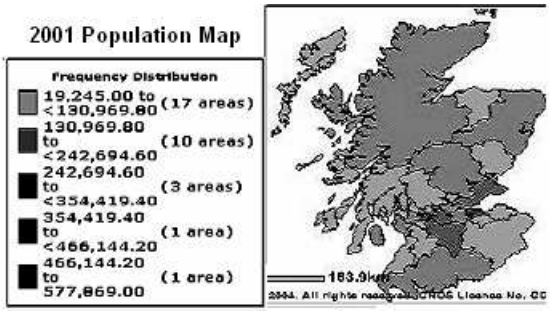

Figure 1: A map showing geo-referenced data

2 Related Work

Systems which generate descriptions of numerical data are not uncommon, e.g., the FOG system (Goldberg et al, 1994), TREND (Boyd, 1997) and MULTI-METEO (Coch, 1998) all generate textual summaries of numerical weather data. (Ferres et al, 2006) generates textual descriptions of informa-

[1] At http://www.statistics.gov.uk/census2001/census2001.asp .

[2] Scotland's Census Results Online, www.scrol.gov.uk.

tion in graphs and enables user querying. SumTime (Sripada et al, 2003) summarises time-series data. While there is no prior work on generating textual descriptions of geo-referenced data, there have been studies on describing spatial data in the context of route directions ((Geldof, 2003); (Marciniak and Strube, 2004)), scene descriptions (Novak, 1986), geometric descriptions (Mitkov, 1990) and spatial descriptions (Ebert et al, 1996). RoadSafe[3], which generates weather forecast texts for road maintenance based on spatio-temporal weather prediction data, also explores similar issues. However, in the Atlas.txt project our main focus is on describing geo-spatial data to visually impaired users.

Other projects address the issue of accessibility with other user interaction paradigms in mind, for example haptic interfaces and non-speech sonic interfaces. Of these, the most closely related are sonification (non-speech audio) approaches to data communication. (Stockman, 2004) discusses issues involved with sonification of spreadsheet data. Of particular interest is the iSonic (Zhao, 2005) data exploration tool which uses sonification techniques to aid interactive exploration of geo-referenced data displayed as chloropleth maps. Such approaches can be seen as complementary to our approach.

3 Knowledge Acquisition

In order to elicit end-user requirements and coordinate evaluation subjects, we interviewed a visually impaired volunteer at the Grampian Society for the Blind[4] (GSB) who is responsible for helping blind users with computer accessibility. He demonstrated how he and other blind users he knows currently explore data; if the data is in tabular form, they use spreadsheet applications which compute descriptive statistics such as the mean and range of a column (row) and use these statistics to gradually build a mental picture of the distribution of the data. Analysing geo-referenced data additionally needs to be associated with corresponding geographical areas. Textual descriptions, he felt, would certainly help visually impaired users to quickly gain an overview of the underlying data provided they present the same information a sighted person extracts from a visual representation of the data.

3.1 Parallel Corpus

In order to model the process of mapping geo-referenced data to its textual description we need to understand how humans perform this activity. There is a huge amount of geo-referenced data online, and we start by considering a few such expert (i.e., statistician) written texts and their corresponding data, which can be found online via organisations like National Statistics. Using such parallel data and text corpora for knowledge acquisition (KA) is a fairly common approach (Sripada et al, 2003). An example from Lambeth Council's website of the sort of texts we are looking at follows below, and the corresponding data appears in Table 1. Such pairings of sentences and tables containing the data communicated in the sentences form our growing corpus. The corpus collection is in an initial stage, and we currently have about 300 such pairs of sentences and data tables gleaned from a handful of online documents.

(1) *Example 1:* People from Non-white ethnic groups constituted 30.3 % of Lambeth's population in 1991, compared with 6.1% of the country as a whole.

(2) *Example 2:* Black African residents have increased in Lambeth by 5.1%. This is also reflected London wide and nationally.

We distinguish content depending on whether it classifies strictly spatial data (e.g., the first sentence in Ex. 2), compares areas (e.g., the second sentence in Ex. 2 and the last phrase in Ex. 1), or compares spatial data across different times. Another way we can distinguish message content has to do with whether messages describe factual information or communicate inferences. Following (Law et al, 2005), we distinguish between *descriptive* messages which are based on data analysis and factual features of the data, and *interpretative* messages which involve expert knowledge about the domain, which can include everything from inferences on the data to the communication of specific domain knowledge. The interpretative features which arise generally involve inferences drawn by the experts about the data, e.g., explanations, cause-effect, etc. Although performing these sorts of inferences is currently beyond the project's scope, our initial studies show a predominance of descriptive messages like those in Ex. 1 and 2, which seems to indicate that addressing these sorts of messages will enable us to address a decent proportion of the sorts of messages we want to communicate. Another point to bear in mind is that the documents we're webscraping are designed for sighted users, which means that one cannot assume that the modalities of text and data

[3] See www.csd.abdn.ac.uk/research/roadsafe.

[4] See www.grampianblind.co.uk

	White		Black Caribbean		Black African		Black Other		Indian		Pakistani		Bangladeshi		Chinese	
	1991	2001	1991	2001	1991	2001	1991	2001	1991	2001	1991	2001	1991	2001	1991	2001
Lambeth	69.7	62.4	12.6	12.1	6.5	11.6	2.7	4.9	2.1	2	0.8	1	0.7	0.8	1.3	1.3
Inner London	74.4	65.7	7.1	6.9	4.4	8.3	2	3.3	3	3.1	1.2	1.6	2.8	4.6	1.1	1.4
Greater London	79.8	71.2	4.4	4.8	2.4	5.3	1.2	2.3	5.2	6.1	1.3	2	1.3	2.1	0.8	1.1

Table 1: The Corresponding Data for Examples 1 and 2

contain the same information. However, although the two modes often contain complementary information, informal analysis indicates that official reporting of geo-referenced data often contains both maps and text to communicate important information, making webscraping fairly viable for corpus collection.

3.2 Corpus Analysis

We index data tables like the one shown in Table 1 to the corresponding texts which describe them in a database, thereby linking information in different modalities and forming a parallel corpus. Essentially the core of KA from corpus analysis lies in the mapping from a (manually interpreted) abstract representation (AR) of an expert textual description or $message$ (Reiter and Dale, 2000) to an AR of the data, as is shown in Figure 3. Note that the textual analysis takes place on the sentential level currently; we leave higher-level discourse structure for future work. Abstract representations of data (ADR) contain the results of applying data analysis methods (which can range from simple descriptive statistics to spatial data mining methods) on the raw data found in tables in the documents.

AR of textual descriptions proceeds in several steps. We started building up an ontology of spatial relations by labelling texts at the sentential level according to the messages communicated; these message labels abstract over texts and indicate the primary spatial relation communicated, e.g., $category : location - density$, which for example labels the text: "The Black Caribbean community is concentrated in the wards around Brixton". Here $category$ is a higher level class in the ontology corresponding to messages communicating information to do with a category of data, in this case the Black Caribbean population, and $location - density$ indicates the particular feature ($location$) and property of this feature ($density$) under discussion. Message labels indicate information content. We indicate sentential rhetorical/discourse content via a set of around ten message predicates adopted from McKeown's message predicates (McKeown, 1985) and RST (Mann and Thompson, 1988). These message predicates communicate contrast, causality, etc. Message predicates and message labels constitute the highest level of abstraction at the sentential level in our ATR.

At a lower level we create ARs for the messages which can be seen as an intermediary level of abstraction between message labels and predicates and the texts themselves and correlate roughly to cue-phrases. We find all cue-phrases which indicate spatial relations in the corpora, e.g., "constituted", "concentrated", to get a range of the sorts of contexts these cues can appear in and then abstract over the cues, forming mostly synonymous classes of cues which can be associated with the same message labels and predicates. The text above has as its message content, the frame $concentrated[Group, Location]$.

3.2.1 Initial Methodology and Findings

We map from ATRs to ADRs by (manually) associating simple sequences of data analysis methods (which we term *macros*) with ATRs, so the previous text's ATR is associated with the method spatial segmentation. The text "A higher number of Black African people live in the north of the borough than in the south" would be associated with first spatial segmentation (Haining, 2003), and then, given that the distribution is not uniform, summation of the population values for the category in geographic regions with different frequencies.

The macros serve as an initial indication of which data analysis methods we should apply to the data and in which order so that we can generate the kinds of texts produced by the experts. We expect that the findings from data analysis will mostly lie at the sentential level, e.g., descriptive statistical information, or information from spatial segmentation. The idea is that these mappings between data and text can be used as a backbone for describing new data sets which share similar findings from spatial data analysis.

From an initial analysis of 70 texts, we found that spatial segmentation was the most common method invoked, occurring 17% of the time, followed by comparisons of values, which occurred 14% of the time. We found that 13% of the texts involved a sequence of data analysis methods, e.g., segmenta-

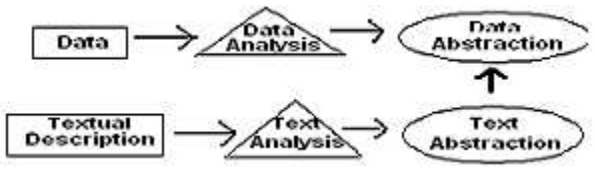

Figure 2: The Text to Data Mapping Process

tion followed by breakdown analysis, and another 13% invoked multiple (non-sequential) data analysis methods like reporting the minimum and maximum values. This leaves the majority (74%) as simply invoking a single data analysis method. Ignoring whether methods are invoked sequentially or in multiples, we found that the largest number of methods invoked involve simply giving significant values, ranking values, summing them or comparing them. However the extent to which these findings (as opposed to others arising from the same data) are significant needs to be resolved so that given a data set, we can determine which findings are "interesting".

4 Future Work

This paper describes work in progress on the Atlas.txt project, which just started in January 2007. As such, our goal here is to introduce the project and initial KA work in the hopes that we will receive useful feedback about our initial methodology which will then guide future work. In this paper we have presented initial ideas about mapping geospatial data to text via the data analysis macros necessary for communicating the corresponding texts with the goal of driving generation of similar sentences for unseen data. We still need to account for discourse level structuring of these messages. We have implemented our ATR in the SimpleNLG framework and toolkit[5], enabling us to generate text from ATRs. However much work still needs to be done toward implementing the data to ADR and ADR to ATR stages.

An immediate area of future work involves analysing more corpora in order to expand our set of data analysis macros. We also need to specifically investigate the extent to which the data analysis macros we associate with ATRs actually do produce the information communicated in the texts. Additionally we need a better understanding of the informational requirements of visually-impaired end-users, and the extent to which results from a survey we are currently running including "think aloud" descriptions of census data from sighted users needs to be adapted for the visually-impaired.

[5]See www.csd.abdn.ac.uk/ẽreiter/simplenlg/.

References

S. Boyd. 1997. Detecting and Describing Patterns in Time-varying Data Using Wavelets. In *Advances in Intelligent Data Analysis: Reasoning About Data, Lecture Notes in Computer Science*, 1280.

J. Coch. 1998. Multimeteo: Multilingual Production of Weather Forecasts. *ELRA Newsletter*, 3:2.

C. Ebert, D. Glatz, M. Jansche, R. Meyer-Klabunde, R. Porzel. 1996. From Conceptualization to Formulation in Generating Spatial Descriptions. In *Proceedings fo the 5th European Conference on Cognitive Modelling*, 96-39.

L. Ferres, A. Parush, S. Roberts, G. Lindgaard. (2006). 2006. Helping People with Visual Impairments Gain Access to Graphical Information Through Natural Language: The iGraph System. In *Proceedings of ICCHP 2006, Lecture Notes in Computer Science*, 4061.

S. Geldof. 2003. Corpus Analysis for NLG. In *Proceedings of the 9th EWNLG*.

E. Goldberg, N. Driedger, R. L. Kittredge. 1994. Using Natural-Language Processing to Produce Weather Forecasts. *IEEE Expert*, 9:2.

R. Haining. 2003. Spatial Data Analysis: Theory and Practice. *Cambridge University Press*.

A. S. Law, Y. Freer, J. R. Hunter, R. H. Logie, N. McIntosh, J. Quinn. 2005. A Comparison of Graphical and Textual Presentations of Time Series Data to Support Medical Decision Making in the Neonatal Intensive Care Unit. *Journal of Clinical Monitoring and Computing*, 19:183–194.

T. Marciniak, M. Strube. 2004. Classification-based Generation Using TAG. In *Proceedings of Natural Language Generation: 3rd International Conference, Lecture Notes in Artificial Intelligence*, 3123.

W. Mann, S. Thompson. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. In *Text:3*.

R. Mitkov. 1990. A Text-Generation System for Explaining Concepts in Geometry. In *Proceedings of the 13th Conference on Computational Linguistics*.

H. J. Novak. 1986. Generating a Coherent Text Describing a Traffic Scene. In *Proceedings of the 11th Conference on Computational Linguistics*.

E. Reiter and R. Dale. 2000. Building Natural Language Generation Systems. Cambridge University Press.

S. Sripada, E. Reiter, J. Hunter, J. Yu. 2003. Exploiting a Parallel TEXT-DATA Corpus. In *Proceedings of Corpus Linguistics*.

S. Somayajulu, E. Reiter, I. Davy. 2003. SumTime-Mousam: Configurable Marine Weather Forecast Generator. In *Expert Update* 6(3):4-10.

T. Stockman. 2004. The Design and Evaluation of Auditory Access to Spreadsheets. In *Proceedigns of the 10th International Conference on Auditory Display*.

H. Zhao, C. Plaisant, B. Shneiderman. 2005. I Hear the Pattern - Interactive Sonification of Geographical Data Patterns. In *Proceedings of ACM SIGCHI Extended Abstracts on Human Factors in Computing Systems*.