Enabling Access to Geo-referenced Information: *Atlas.txt*

Kavita E. Thomas
Department of Computing Science
University of Aberdeen
Scotland
tkavita@abdn.ac.uk

Livia Sumegi & Leo Ferres
Human-Oriented Technology Laboratory
Carleton University
Canada
lferres@ccs.carleton.ca

Somayajulu Sripada
Department of Computing Science
University of Aberdeen
Scotland
yaji.sripada@abdn.ac.uk

ABSTRACT

We present Atlas.txt, a novel data-to-text natural language generation system which enables access to geo-referenced information like online census data. We first discuss initial findings from an accessibility study on geo-referenced data and outline needs requirements for visually-impaired users of such data. We then present work towards realising our data-to-text system and indicate how it aims to address this issue.

Keywords

visually-impaired access, geo-referenced data

1. INTRODUCTION

Geo-referenced data is data which has a geographic component, and is distributed over a region on Earth. Such data is often visualised as thematic choropleth maps, as shown in Figure 1 which comes from Scotland's Census Results Online (www.scrol.gov.uk), where colour shading indicates density of distribution, so that dark or intense regions on such maps indicate where the variable, for example unemployment, is most frequently located. Most governments make geo-referenced data like census results available to the public online, and they are obliged to make this data accessible. However data displayed as thematic maps is currently inaccessible to the visually-impaired, so the focus of this work is to make this information accessible. Geo-referenced data is frequently stored in long data tables in spreadsheets which hinders visually-impaired users, who must rely on screenreaders to read out the values in the data table, from quickly and easily forming an overview of how the data is distributed. For example, although the table corresponding to Figure 1 only has 31 rows (corresponding to the 31 council areas of Scotland), visually-impaired users still need to listen to each of these values and remember them in order to form a mental picture of the overall trends in population distribution which sighted users can infer from the map in under a minute.

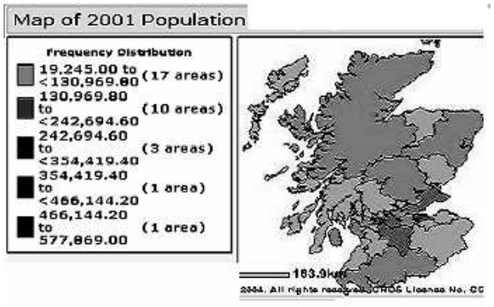
Map of 2001 Population
Frequency Distribution
19.245.00 to <130,969.80 (17 areas)
130,969.80 to <242,694.60 (10 areas)
242,694.60 to <354,419.40 (3 areas)
354,419.40 to <466,144.20 (1 area)
466,144.20 to 577,869.00 (1 area)
183.9km
2004, All rights reserved. SCROS Licence No. CC

Figure 1: A typical map showing geo-referenced data

We conducted a usability study on a group of blind and sighted users to determine the blind users' relative difficulties in interpreting such data. We describe this study and present our initial findings and the implications they pose. We indicate issues which applications aiming to make such data accessible to the visually-impaired need to account for. One such issue is familiarity. If visually-impaired users are unfamiliar with the geography of the region described, they are unable to visualise how the data is distributed even if they can remember all the read-out table values. Sighted users on the other hand, can interpret how geo-referenced data is distributed from a map even if they are completely unfamiliar with the region described. In effect, this prevents visually-impaired users from exploring new regions' characteristics and presents a serious obstacle for them.

We present Atlas.txt, a novel data-to-text natural language generation (NLG) system under development which enables access to geo-referenced information. Atlas.txt aims to analyse geo-referenced data in order to determine where salient clusters lie and then communicate this information in texts which can be read out via a screenreader. We indicate some additional considerations based on our usability study which we will incorporate into Atlas.txt to make the system more accessible.

2. RELATED WORK

W4A2008, April 21–22, 2008, Beijing, China.

Systems which generate descriptions of numerical data are not uncommon, e.g., the FOG system [6], TREND [1] and MULTI-METEO [2] all generate textual summaries of numerical weather data. The iGraph system [4] generates textual descriptions of information in graphs and enables user querying. SumTime[1] summarises time-series data and RoadSafe[2] generates travel advisories and looks at spatial 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 ([5], [8]), scene descriptions [11], geometric descriptions [10] and spatial descriptions [3].

Data-to-text generation projects like Atlas.txt differ from the traditional three-stage pipeline generation architecture of most NLG systems [12] because they need to analyse the data in order to determine data abstractions before then going on to the traditional three stages of document planning, micro-planning and realisation, as was put forward by [13].

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 sonfication (non-speech audio) approaches to data communication. [14] discusses issues involved with sonification of spreadsheet data. Of particular interest is the sonification-based data exploration tool iSonic [16] which uses sonification techniques to aid interactive exploration of geo-referenced data displayed as choropleth (shaded thematic) maps. Such approaches can be seen as complementary to our approach, which is to generate textual descriptions of geo-referenced data, and future work will need to investigate the advantages and disadvantages of each interaction paradigm.

3. USER NEEDS STUDY

A preliminary informal user needs analysis (see [15]) into the needs and requirements of visually-impaired individuals on geo-referenced map-reading tasks like interpreting census maps indicated that blind people have similar spatial and comprehension abilities as sighted people, which echoed work by [7]. What is unclear are the specific informational requirements of blind people and the sort of language which best helps them to understand spatial data. We conducted a study into visually-impaired users' needs requirements for accessing geo-referenced information. The purpose of this study was to (1) investigate blind users' requirements for understanding geo-referenced data like census maps, (2) determine whether expert-written texts enable better access to geo-referenced data than via screen-read data tables, and (3) assess the impact of familiarity with a region/area on understanding expert summaries.

Our main hypotheses were:

1. Blind participants will perform tasks more accurately and understand the data better given texts as opposed to tables

2. Sighted participants will have no overall difference between task performance or comprehension of maps as opposed to texts

3. Blind participants will make more corrections of expert-written texts (tested in Part 2, section 3.1) than sighted participants, and in particular, their corrections will often involve simplifying complex spatial expressions which refer to abstractions over regions (e.g., "Eastern seaboard regions") which they are unfamiliar with

4. No overall difference between groups for familiarity when participants are given texts

5. Sighted participants will perform tasks more accurately and understand the data better given maps than blind participants given tables especially for unfamiliar regions

3.1 Experiment Design and Methodology

The stimuli consist of expert-produced maps, tables and texts from census data published online at Statistics Canada (StatCan) and Scottish Census Results Online (SCROL). The experiment has three parts:

- Part 1: Blind people were presented with 2 read-out data tables and 2 texts, and sighted participants were presented with 2 maps and the same 2 texts that the blind participants got. Here the goal was to compare within-participants for blind and sighted groups how their task performance and data comprehension differs depending on whether they are presented with tables/maps vs. texts.

- Part 2: Here the goal was to comparatively evaluate blind and sighted participants' corrections of 2 expert-written texts. Both blind and sighted participants were presented with the same texts.

- Part 3: Blind people were presented with 2 read-out data tables and 2 texts, and sighted participants were presented with 2 maps and the same 2 texts as the blind participants got, a pair each (map and text for sighted participants, and table and text for blind participants) from familiar (Canada) and unfamiliar (Scotland) regions of the world. Here the goal is to compare blind and sighted participants' task performance and data comprehension on texts vs. maps/tables from familiar and unfamiliar areas.

Texts tended to have 1-2 paragraphs with under 5 sentences per paragraph, and were drawn from online documents found on the websites of the census authorities mentioned above. Tables and maps were roughly matched on complexity of the data described, as we couldn't find the corresponding data tables for many of the maps we found online at these census websites. This matching process was a very intuitive process, and ideally we should have used the tables corresponding to the maps in order to avoid comparability issues between maps and tables.

3.2 Initial Findings

This study had 5 sighted and 5 blind participants, roughly matched on their familiarity with data analysis and statistics, all of whom were Ottawa locals. This is too small a group for us to deduce statistically significant findings. This is a frequent problem in studies involving visually-impaired subjects, as it is very hard to find participants, particularly when we additionally want to find visually-impaired participants who vary in level of expertise with data analysis. However, qualitatively we can draw some initial conclusions from both the testing process and an initial look at the results. There were two main areas in which blind participants

[1] More information can be found at http://www.csd.abdn.ac.uk/research/sumtime.

[2] See http://www.csd.abdn.ac.uk/research/roadsafe.

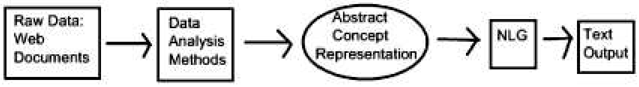

Figure 2: The Data to Text Mapping

had more difficulties than sighted ones. These are: inferring trends, and mentally visualising geographically unfamiliar data. Spreadsheet applications can indicate what the maximum and minimum values in a column are, but in order to gain a sense of trends in the data, blind users would need to hear and roughly remember or rank all the values in the table.

In fact, gaining an overview was uniformly very difficult for blind participants, even for those with much experience in data analysis, because it requires them to remember large numbers of column values and then coordinate those values with their corresponding geographic region. A participant whose job involves statistical data analysis indicated that while screenreaders are invaluable tools for making data accessible, they do not enable gaining a quick or easy overview of general trends in numerical data.

For our stimuli focusing on familiarity, we found that if blind participants were unfamiliar with the region described by the data, they were unable to mentally visualise it by having table values read out, as they did not know where the regions described were relatively located. Sighted participants on the other hand were easily able to perceive general trends given maps for unfamiliar areas.

We also found to our surprise that blind users also had more difficulties than sighted ones in interpreting geo-referenced data presented in expert-written texts. This is likely because the 200-300 word texts were dense and contained much specific information and blind users had greater difficulties in locating relevant information in order to answer specific questions related to the data, while sighted participants could simply skim through the texts again to find specific information. We read out texts to blind participants and enabled them to search texts for specific keywords, and we would then read out sentences containing these keywords in order to mimic the sort of interactivity they would have with a screenreader.

4. DEVELOPING THE ATLAS.TXT SYSTEM

Atlas.txt is a data-to-text NLG system under development which follows the pipeline architecture shown in Figure 2. It takes as input a table of geo-referenced data and, making use of geographical locations found in an electronic gazetteer resource, performs data analysis on the data. Clustering identifies the maximal and minimal clusters and data analysis also reveals geographic trends of increasing or decreasing values ([9]). This information is then passed to the NLG module which plans the text and realises it. A qualitative corpus study into how expert statisticians communicate geo-referenced data ([15]) indicated that messages communicating maximum values, minimum values and trends are the most commonly communicated messages, and Atlas.txt therefore aims to communicate these messages.

However given the findings of our accessibility study, it is clear that a system aiming to communicate geo-referenced data to the visually-impaired should also take into account both the issues of (1) familiarity (in case users are unfamiliar with the described region's geography) and (2) should provide an overview of the data.

Familiarity can be addressed by summarising salient facts about the geography of a place before presenting more detailed information about how the geo-referenced variable is distributed in the region. If we are describing unemployment in Aberdeen, we can start by presenting some salient facts about Aberdeen; that is, we aim to begin our texts with overviews like the following:

(1) Aberdeen is Scotland's third-largest city with a population of around 200,000. It is located in the Northeast of Scotland on the North Sea. The city centre is bounded on the North by the River Don and on the South by the River Dee, although the actual city limits extend considerably beyond the rivers.

The choice of information to be included in an introduction is an issue we will pursue in greater detail in future work. However it seems clear that the information contained here should be salient to both the information being described (i.e., the geo-referenced variable, in this case unemployment), and also, the introduction should enable visually-impaired users to mentally visualise the location of the region described in a larger (more familiar) region (in this case the UK), and also contain some basic information about the salient geographic orientation of the region itself. In this case, it is important to know that Aberdeen has the sea bounding it to the East and that the centre of the city is bounded by the two rivers. This information should introduce the region before then describing how the data itself is distributed, and we aim to produce texts like the following:

(2) Unemployment figures in Aberdeen are generally low at less than 3%, except in the inner wards where it is in the range of 3 - 6%. Unemployment is highest in the St. Machar and Tullos Hill areas, where it is around 6%.

However, visually-impaired people would be unable to locate the St. Machar and Tullos Hill areas as sighted users could, given the map corresponding to this data which is shown in Figure 3. So ideally we should also provide brief descriptions of any named areas to help blind users to create a mental visualisation of the data. Texts like the ones below would therefore be better:

(3) Unemployment figures in Aberdeen are generally low at less than 3%, except in the inner wards where it is in the range of 3 - 6%. Unemployment is highest in the St. Machar and Tullos Hill areas, where it is around 6%. St. Machar lies in the North of the city centre, just South of the River Don and near the coast, well within the city limits, while Tullos Hill lies to the Southwest of the city, South of the River Dee and is on the coast.

(4) Unemployment figures in Aberdeen are generally low at less than 3%, except in the inner wards where it is in the range of 3 - 6%. Unemployment is highest, at around 6%, in St. Machar, which is in the northern part of the city centre, and Tullos Hill, a coastal area in the Southwest of the city.

The choice of distinguishing description to apply is arguably less and more difficult than in traditional approaches to generating referring expressions (GRE). Traditional GRE

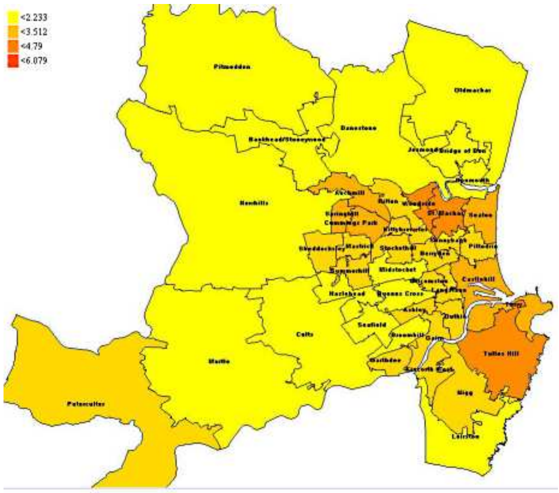

Figure 3: Aberdeen Unemployment Distribution

focuses on uniquely distinguishing an object from a set of distractors. E.g., a pet-owner going into a kennel to pick up her dog and seeing what the distractors (i.e., the other animals in the kennel) look like, would refer to her dog with the minimal number of descriptors necessary to distinguish her pet. So if there were only cats in the kennel aside from her dog, she'd refer to her pet as "the dog"; if there were other dogs, but none which was white like her dog, she'd refer to her pet as "the white dog", and so on. In our domain however, the object is a sub-region or group of sub-regions, and the distractors are other sub-regions in the overall region being considered. As can be seen in Figure 3, it is quite difficult to come up with a uniquely distinguishing description for St. Machar. However, in this domain, a uniquely distinguishing description is not as necessary as it is in domains like the kennel domain, since the goal of the description is to enable visualisation. This is why a description like the one in Example 4 above is sufficient; it enables the user to visualise where St. Machar and Tullos Hill are roughly, and that is sufficient information for an initial impression of the data.

5. FUTURE WORK

Ideally a system communicating geo-referenced data via text should be interactive, as our summaries may not provide enough detailed information for the user who wants to browse the data more interactively. Interactive data exploration for graphs is the focus of work by [4]. While interactivity is our goal, we realise that we first need to be able to generate summaries before focusing on interactivity, and that is the scope of this project at its current stage.

Future work within this stage of the project will involve investigating both the range of distinguishing descriptions we can employ in maximum, minimum and trend messages, and also explore some of the requirements on what information should be included in introduction messages. Then we will explore how these requirements can be phrased as procedures which can then be implemented.

6. ACKNOWLEDGEMENTS

The authors would like to thank Martin Lachance, Stacy Bleeks, Krista Kowalchuk, Bill Joyce and Brad Hawkes at Statistics Canada for their invaluable help and assistance.

7. REFERENCES

[1] S. Boyd. Detecting and describing patterns in time-varying data using wavelets. *Advances in Intelligent Data Analysis: Reasoning About Data, Lecture Notes in Computer Science*, 1280, 1997.

[2] J. Coch. Multimeteo: Multilingual production of weather forecasts. *ELRA Newsletter*, 3(2), 1998.

[3] C. Ebert, D. Glatz, M. Jansche, R. Meyer-Klabunde, and R. Porzel. From conceptualization to formulation in generating spatial descriptions. *Proceedings fo the 5th European Conference on Cognitive Modelling*, pages 96–39, 1996.

[4] L. Ferres, A. Parush, S. Roberts, and G. Lindgaard. Helping people with visual impairments gain access to graphical information through natural language: The igraph system. *Proceedings of ICCHP 2006, Lecture Notes in Computer Science*, 4061, 2006.

[5] S. Geldof. Corpus analysis for nlg. *Proceedings of the 9th EWNLG*, 2003.

[6] E. Goldberg, N. Driedger, and R. Kittredge. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2), 1994.

[7] R. Kitchin, M. Blades, and R. Golledge. Understanding spatial concepts at the geographic scale without the use of vision. *Human Geography*, 21(2), 1997.

[8] T. Marciniak and M. Strube. Classification-based generation using tag. *Proceedings of Natural Language Generation: 3rd International Conference, Lecture Notes in Artificial Intelligence*, 3123, 2004.

[9] A. Mehnert and P. Jackway. An improved seeded region growing algorithm. *Pattern Recognition Letters*, 18:10, 1997.

[10] R. Mitkov. A text-generation system for explaining concepts in geometry. *Proceedings of the 13th Conference on Computational Linguistics*, 1990.

[11] H. Novak. Generating a coherent text describing a traffic scene. *Proceedings of the 11th Conference on Computational Linguistics*, 1986.

[12] E. Reiter and R. Dale. Building natural language generation systems. *Cambridge University Press*, 2000.

[13] S. Sripada, E. Reiter, J. Hunter, and J. Yu. A two-stage model for content determination. *Proceedings of the 8th ACL-EWNLG*, 2001.

[14] T. Stockman. The design and evaluation of auditory access to spreadsheets. *Proceedigns of the 10th International Conference on Auditory Display*, 2004.

[15] K. Thomas and S. Sripada. Atlas.txt: First steps towards bridging the gap between geo-referenced data and text. *Proceedings of the 11th European Workshop on Natural Language Generation*, 2007.

[16] H. Zhao, C. Plaisant, and B. Shneiderman. I hear the pattern - interactive sonification of geographical data patterns. *Proceedings of ACM SIGCHI Extended Abstracts on Human Factors in Computing Systems*, 2005.