Abstract

The current interest in data acquisition and analysis has resulted in a large number of solutions available to the public. However, anyone other than professionals in the field can find it difficult to make sense of this sea of data. This demo showcases a tool that produces general static reports (as opposed to query or intention based systems of past NLG interest) of combined text and graphics given any spreadsheet sent by email.

1 Introduction

The current interest in data acquisition and analysis has resulted in a large number of solutions available to the public (Microsoft Power BI,1 Pentaho,2 etc.). However, anyone other than professionals in the field can find it difficult to make sense of this sea of data. Report generation from tabular data has a long tradition in NLG (Fasciano and Lapalme, 1996; Kerpedjie et al., 1997; Yu et al., 2007; Hunter et al., 2012). However, these systems assume that a knowledgeable user can guide the system with explicit communicative intentions in the form of queries or emphasis in particular columns or relations (Fasciano, 1996; Labbé et al., 2015). How to fulfill those expectations when confronted with a novice user can span whole research projects in smart User Interfaces. Instead, in this demo we present a tool that produces general static reports of combined text and graphics given any spreadsheet. Our tool incorporates concepts of surprise, popularized from the KDD community (Guillet and Hamilton, 2007) and employed lateraly in other NLG systems (Molina et al., 2011).

Our system is based on the ANA architecture (Kukich, 1983): fact generation, message generation, content planning and tactical generation. It takes any spreadsheet in Excel, CSV and OpenDocument format sent by email3 and produces a OpenDocument text document with a textual description of the data and embedded graphs, a form of multimedia generation (André, 2000).

It addresses two traditional conditions in report generation (Kittredge and Polguere, 2000): a primary interest in objective or fixed type data and a conceptual summarization over said data. Two other conditions are approximated (a temporal dimension in the data, which is attempted using a number of heuristics) or left for potential follow up consulting engagements (recurrent situation of communication).

Similar to (Molina et al., 2011), we seek to summarize relevant facts with explanatory descriptions and graphical information. However, we have a different main goal which is to provide an overview of any tabular data without extra domain knowledge provided by the user. We also share the secondary goal of producing reports that are informative and persuasive, useful for non-expert users and have a uniform style.

2 Structure of the Demo

Our demo shows a number of spreadsheets (Figure 1 (a), adapted from Foreman (2013)) from which

---

1http://powerbi.microsoft.com/
2http://pentaho.com
3To the address get@thedatapreport.com
the audience can change the data with a provided OpenCalc instance running in the machine. Then the spreadsheet will be submitted to the system and the resulting multi-page report will be shown and discussed (Figure 1 (b)).

Acknowledgements

The author would like to thank the reviewers and Annie Ying for comments and references.

References


