Automatic Reports from Spreadsheets: Data Analysis for the Rest of Us

Pablo Ariel Duboue
Textualization.com
White Plains, New York, USA

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; Kerpedjiev 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 email[3] 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

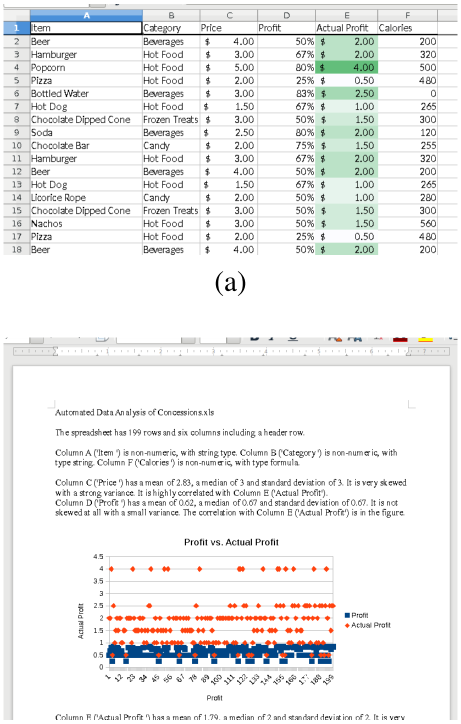
Our demo shows a number of spreadsheets (Figure 1 (a), adapted from Foreman (2013)) from which

[1] `http://powerbi.microsoft.com/`

[2] `http://pentaho.com`

[3] To the address `get@thedatareport.com`

Proceedings of The 9th International Natural Language Generation conference, pages 244–245,
Edinburgh, UK, September 5-8 2016. ©2016 Association for Computational Linguistics



	A	B	C	D	E	F
1	Item	Category	Price	Profit	Actual Profit	Calories
2	Beer	Beverages	$ 4.00	50%	$ 2.00	200
3	Hamburger	Hot Food	$ 3.00	67%	$ 2.00	320
4	Popcorn	Hot Food	$ 5.00	80%	$ 4.00	500
5	Pizza	Hot Food	$ 2.00	25%	$ 0.50	480
6	Bottled Water	Beverages	$ 3.00	83%	$ 2.50	0
7	Hot Dog	Hot Food	$ 1.50	67%	$ 1.00	265
8	Chocolate Dipped Cone	Frozen Treats	$ 3.00	50%	$ 1.50	300
9	Soda	Beverages	$ 2.50	80%	$ 2.00	120
10	Chocolate Bar	Candy	$ 2.00	75%	$ 1.50	255
11	Hamburger	Hot Food	$ 3.00	67%	$ 2.00	320
12	Beer	Beverages	$ 4.00	50%	$ 2.00	200
13	Hot Dog	Hot Food	$ 1.50	67%	$ 1.00	265
14	Licorice Rope	Candy	$ 2.00	50%	$ 1.00	280
15	Chocolate Dipped Cone	Frozen Treats	$ 3.00	50%	$ 1.50	300
16	Nachos	Hot Food	$ 3.00	50%	$ 1.50	560
17	Pizza	Hot Food	$ 2.00	25%	$ 0.50	480
18	Beer	Beverages	$ 4.00	50%	$ 2.00	200

(a)

Automated Data Analysis of Concessions.xls

The spreadsheet has 199 rows and six columns including a header row.

Column A ('Item ') is non-numeric, with string type. Column B ('Category ') is non-numeric, with type string. Column F ('Calories ') is non-numeric, with type formula.

Column C ('Price ') has a mean of 2.83, a median of 3 and standard deviation of 3. It is very skewed with a strong variance. It is highly correlated with Column E ('Actual Profit').
Column D ('Profit ') has a mean of 0.62, a median of 0.67 and standard deviation of 0.67. It is not skewed at all with a small variance. The correlation with Column E ('Actual Profit') is in the figure.

Column E ('Actual Profit ') has a mean of 1.79, a median of 2 and standard deviation of 2. It is very

(b)

Figure 1: (a) Input data, adapted from Foreman (2013); (b) Example output.

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

Elisabeth André. 2000. The generation of multimedia presentations. *Handbook of natural language processing*, pages 305–327.

Massimo Fasciano and Guy Lapalme. 1996. Postgraphe: a system for the generation of statistical graphics and text. In *Proceedings of the Eighth International Workshop on Natural Language Generation*, pages 51–60.

Massimo Fasciano. 1996. *Génération intégrée de textes et de graphiques statistiques*. Université de Montréal.

John W Foreman. 2013. *Data smart: using data science to transform information into insight*. John Wiley & Sons.

Fabrice Guillet and Howard J Hamilton. 2007. *Quality measures in data mining*, volume 43. Springer.

James Hunter, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, and Cindy Sykes. 2012. Automatic generation of natural language nursing shift summaries in neonatal intensive care: Bt-nurse. *Artificial intelligence in medicine*, 56(3):157–172.

Stephan Kerpedjiev, Giuseppe Carenini, Steven F Roth, and Johanna D Moore. 1997. Autobrief: a multimedia presentation system for assisting data analysis. *Computer Standards & Interfaces*, 18(6):583–593.

Robert I Kittredge and Alain Polguere. 2000. The generation of reports from databases. *Handbook of natural language processing*, pages 261–304.

Karen Kukich. 1983. Design of a knowledge-based report generator. In *Proc. of ACL*.

Cyril Labbé, Claudia Roncancio, and Damien Bras. 2015. A personal storytelling about your favorite data. In *Proc. of ENLG 2015*, September.

Martin Molina, Amanda Stent, and Enrique Parodi. 2011. Generating automated news to explain the meaning of sensor data. In *International Symposium on Intelligent Data Analysis*, pages 282–293. Springer.

Jin Yu, Ehud Reiter, Jim Hunter, and Chris Mellish. 2007. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13(01):25–49.