Textually Summarising Incomplete Data

Stephanie Inglis, Ehud Reiter and Somayajulu Sripada
Department of Computing Science, University of Aberdeen
Aberdeen, UK, AB24 3UE
r01si14@abdn.ac.uk, e.reiter@abdn.ac.uk, yaji.sripada@abdn.ac.uk

Abstract

Many data-to-text NLG systems work with data sets which are incomplete, ie some of the data is missing. We have worked with data journalists to understand how they describe incomplete data, and are building NLG algorithms based on these insights. A pilot evaluation showed mixed results, and highlighted several areas where we need to improve our system.

1 Introduction

Natural language generation systems which produce texts based on incomplete data can produce low quality or inaccurate reports (Reiter & Dale. 2000). Improving the quality of these reports means more accurate information can be concluded from datasets, which increases the impact of data being collected. All tasks described in this paper use the EmDat database (Guha-Sapir, Below, and Hoyois).

2 Related Work

Daniel et al. (2008) identified three major data quality issues which occur in a large number of datasets – incompleteness, inconsistency and incorrectness. The paper describes scenarios where data quality has a profound knock on effect, such as when ordering incorrect quantities of medical supplies. Based on findings from preliminary experiments involving non-experts, missing data was the most important of these issues. Therefore, missing data is the quality focus of this paper.

With automated journalism being used to report news in an unbiased fashion, care needs to be taken to improve generated texts. One way this has been done is through research on human written news (Van der Kaa & Krahmer, 2014).

3 Experts

3.1 Expert Identification

To design a system which models human behavior accurately, knowledge was elicited from experts – people who use datasets to create texts for humans. One such domain is journalism. The Guardian newspaper has a section dedicated specifically for this type of journalism called Datablog[1]. They use datasets to produce texts allowing non-experts to access the stories being told by the data. Some journalists from Datablog agreed to take part in a protocol analysis, which encourages the participant to speak aloud as they complete a task (Ericsson, 2006). This allows insight into how a data journalist produces text from a dataset, which can later be used to create an algorithm.

3.2 Normal Process

We first asked the journalists to describe in general terms how they would write an article based on incomplete data. Both journalists agreed that they

[1] https://www.theguardian.com/data

Proceedings of The 10th International Natural Language Generation conference, pages 228–232,
Santiago de Compostela, Spain, September 4-7 2017.

would not extrapolate gaps in the data and would only report the raw data to ensure accurate reports. Instead, they would contact the source of the data and enquire about the reasoning behind the gaps. If gaps could be remedied by contacting experts in the dataset's domain, this avenue would be explored. Otherwise, academic papers on the dataset would be consulted to see if another solution had been identified. If not, an attempt to locate an alternative dataset would be made. Both journalists also agreed they would not use a dataset if more than half the data was missing.

If no other datasets exist, they would question whether this story is appropriate. In one instance, journalists wanted to write a story about the number of sexual assaults on university campuses, and found that this data was not recorded. The story became that universities were not accurately recording sexual assaults on campuses (McVeigh, 2015).

If an incomplete dataset was used for a story, gaps would be communicated to the reader, for example by giving this information as footnotes.

3.3 Knowledge Elicitation

We asked journalists to participate in a formal knowledge elicitation task. They were asked to write texts to summarise data from the EM-Dat database.

The first dataset had no missing data. While unrealistic, this acted as a baseline to observe their methodology when missing data was not an issue. The second dataset had some data missing, but with enough data present to allow a report to be created. The final dataset had a large amount of missing data, which would likely be unsuitable for an article.

The journalists were asked to imagine they had to write an article using these datasets. For each dataset, they had to describe their steps to produce an article. They had access to the data in a spreadsheet on a laptop to allow them to manipulate data as they would normally to simulate normal working conditions as closely as possible.

> "I wouldn't say well the average looks to be between these periods this so therefore let's just extrapolate. I would not do that at all."

Figure 1 - Quote from transcript of Journalist 1, Dataset 3

A dictaphone recorded the journalists throughout.

3.4 Outcome Methodology

Both journalists followed similar methodologies. The journalists wanted to investigate the metadata first, such as the validity of the dataset itself, any bias that may be present from the database creator, and column headers definitions. They both also disregarded the current year as this was deemed incomplete since the year itself is currently incomplete.

Next, columns with no missing data were identified. The maximum value and minimum values of these columns were noted along with which years the maximum and minimum occurred. This was compared with the years of the maximum and minimum of other columns.

If no columns were complete, a threshold of present data would be decided, and applied consistently across all columns in the dataset. If data is missing, the phrase "of the data reported" was added.

Next the journalists looked at the rows for complete time periods, such as decades, otherwise for time periods with years divisible by 5. This was not always possible as events did not happen every year. However, just because a year is not present in the dataset does not mean an event did not occur; the entry may be missing from the database.

One journalist said they would not report an average figure as it did not make sense in this context. Events occur at different magnitudes and it would be highly inaccurate to assume all variables are evenly distributed across all events.

4 Algorithm

The algorithm that mimics the journalist methodology first computes the best block data and generates text to describe it as described below.

4.1 Preprocessing

A CSV file with the dataset is read, and columns with meaningful information are identified. Columns with redundant information were removed. These are columns acting as metadata which never have empty cells – such as year and occurrences, and columns that are the sum of other columns, which can later be calculated if required later.

If the current year has an entry, this is put to one side. The remaining data can now be used to search for the "best" block of data to talk about. A block is

defined as a subset of the dataset of any size, where data is either missing, or present. Rows within a block must remain contiguous, however all possible sets of columns, regardless of whether the columns in the set are next to each other are considered. Any gaps in the data are replaced with -1, to indicate that this cell has no data. Using 0 may be ambiguous as this is a legitimate number that could be reported.

4.2 Best Blocks

Instead of using the entire dataset with large areas of missing data to produce texts, blocks with more present data than missing data were identified.

We select the block with the highest score using the scoring function below:

$$Score(block) = \#DataElements(block) - \#MissingData(block)$$

If the total score is a negative number, more than 50% of the data is missing, and the block should be rejected for being too sparse.

If more than one block has the highest score, the block with the smallest percentage of missing data is chosen as the single "optimal" block.

4.3 Algorithm Functions

Once the optimal block has been selected, the algorithm looks for interesting elements to talk about.

If the block does not cover all rows of the dataset, text is added to give the time period discussed, in the form "Between *firstYear* and *lastYear*", with firstYear being the year of the first event, and lastYear being the year of the final event.

As each text output gives text for each column, one column was selected as the focus for the text. Both years and occurrences were ruled out as possible foci since they were not "meaningful" variables. For each column in the optimal block, the maximum values are reported in the form "the worst year for *column* as a result of *disaster* in *country* was *year* when there was *value*". This was the first thing both journalists considered with regards to the data itself:

"*Let's just sort it to start with because usually in headlines we think of like what was the worst year.*" – Journalist 1

"*When I look at the data I look at the year, the time series for the total deaths and the biggest number*" – Journalist 2

This is supported by research that people are more interested in negative headlines than positive headlines (Trussler & Soroka, 2014). Therefore, only the worst years are reported.

Next, the years in which there is a recorded event are investigated in descending order. The algorithm detects the number of consecutive years. This number is rounded down to the nearest multiple of 5. If this number is a multiple of 10, a sentence can report information about "the last *x* decades". Alternatively, if it is a multiple of 5 but not a multiple of 10, a sentence can report information about "the last *x* years". The rationale behind this is demonstrated from an excerpt from the transcript of participant 1 for dataset 2:

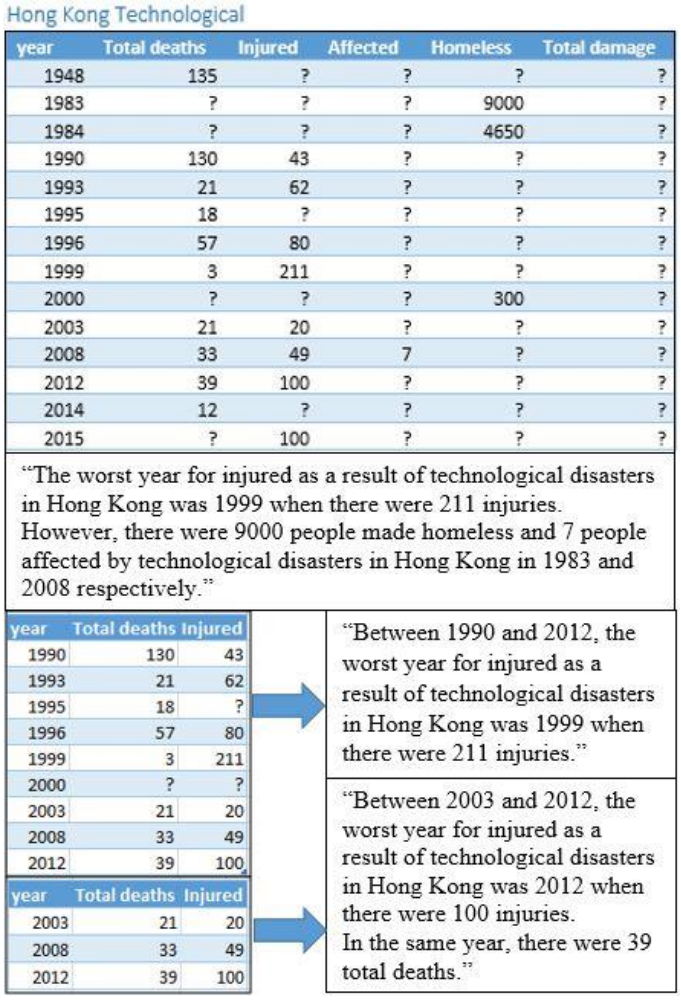

Figure 2 – The full dataset for technological events in Hong Kong, the "optimal" block with missing data, and the "optimal" block without missing data (selected by the algorithm). This figure also shows the corresponding generated texts (also produced by the algorithm) for each data block.

"There are years missing in the sequence…we don't have any period that would give us something to talk about a decade, and definitely not the most recent."

These sentences are produced in the same way the sentences about columns are produced – giving the time period, and the maximum values for each column. Like the definition of the current year, the current decade is also classed as being incomplete, and should not be used. A separate sentence is added to report the current year so far.

5 Pilot Evaluation

An experiment was designed to judge the output texts generated by the algorithm using SimpleNLG (Gatt & Reiter, 2009). Six datasets with varying degrees of missing data were chosen, and three texts were generated for each dataset. One text was generated using the entire dataset, another with the "optimal" block, and the third with the largest block containing no missing data.

The text structure was kept the same for all outputs to minimise any unwanted bias in the writing style. All texts report only the worst figures.

The datasets were ordered alphabetically, and the order in which the texts were shown were randomly generated by numbering them (1 for full dataset text, 2 for no missing data block text, and 3 for the "optimal" algorithm output), and a random sequence generator to create the order.

Participants were asked to choose which of the three texts they thought was the most appropriate in describing the dataset. 20 participants were recruited using social media. A space was also left for participants to leave comments.

We hypothesised that the text produced by our algorithm would be preferred over the control texts. Although not statistically significant, we found participants preferred texts generated by the full dataset (40.8% against 32.5% (optimal) and 26.7% (full)). Comments left by participants (detailed in section 6) allow improvements to be made.

6 Future Work

6.1 Missing Data

Text with missing data should be highlighted as having missing data by adding phrases such as "of the data recorded" or "of the data available". This ensures the reader of the text is aware that not all data was available when generating this text. One participant did not think this was clear and remarked "*for me, appropriate [text] would be 'the worst year recorded*'".

Secondly, business rules can be added to improve plausibility and confidence. If there are a large number of deaths but no one injured or affected, the number of deaths may be too high, or there may be data missing from the other columns. This was considered by one participant who commented that it was "*easy to find the outliers*" and that "*simply stating the biggest number could lead to false information".*

Additionally, a weighting factor will be added to the scoring function to model the importance of missing data. For instance:

$$Score(block) = \#DataElements(block) - Weight * \#MissingData(block)$$

6.2 Text

Multiple participants commented on the language used, particularly conjunctives. One participant said "*there are some texts where the connective 'however' does not seem to fit well",* while another pointed out "*'however' shouldn't be used as it refers to the same year thus making this statement confusing"*. Care will be taken to resolve this. Also, reporting "*large figures as $158230000 in so many digits"* was confusing for participants, so presentation of such values will be made more appropriate.

Participants felt the time period should also be made explicit for the full dataset as one participant noted: "*I would never find 'the worst year ever' without a date range to be appropriate"*. Therefore, the date range will be added for all texts.

The content of the text could be ordered by importance. Importance could be measured by how important the information is e.g. if a death toll is particularly large. The importance could be investigated by revisiting the interview with the journalists from the experiment, or run a corpus analysis and look at the frequency of words in the text.

7 Conclusion

Knowledge has been gathered from domain experts and used to design and create an algorithm. While the pilot evaluation had mixed results, the feedback is crucial in taking steps to improve the algorithm.

References

Daniel, F., Casati, F., Palpanas, T., Chayka, O. and Cappiello, C. (2008*). Enabling Better Decisions Through Quality-Aware Reports In Business Intelligence Applications.*

Ericsson, K. A. (2006). Protocol Analysis and Expert Thought: Concurrent Verbalizations of Thinking during Experts' Performance on Representative Tasks. In: *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge: Cambridge University Press. p223-242.

Gatt, A and Reiter, E (2009). *SimpleNLG: A realisation engine for practical applications*. Proceedings of ENLG-2009

Guha-Sapir, D., Below, R., and Hoyois, Ph. *EM-DAT: International Disaster Database.* Available: www.emdat.be. Last accessed 10th May 2017.

McVeigh, K. (2015). *Top universities fail to record sexual violence against students, The Guardian,* 24th May. Available: https://www.theguardian.com/education/2015/may/24/top-universities-fail-record-sexual-violence-against-students-russell-group. Last accessed 10th May 2017.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.

Trussler, M and Soroka, S. (2014). Consumer Demand for Cynical and Negative News Frames. *The International Journal of Press/Politics*. 19 (3), p360-379.

Van der Kaa, H. & Krahmer, E. (2014). Journalist versus news consumer: The perceived credibility of machine written news. *Proceedings of the Computation + Journalism conference*.