Data Quality in Reasoning

Stephanie Inglis, Ehud Reiter and Somayajulu Sripada

Department of Computing Science
University of Aberdeen, Aberdeen, Scotland
{r01si14, e.reiter, yaji.sripada}@abdn.ac.uk

Abstract. Decision making using data is dependent on the quality of the data being used to make those decisions. Currently, data-to-text recommendation systems do not take this into consideration. Unsatisfactory recommendations are likely to cause further damage, which could have a detrimental effect economically or from a health and safety perspective. Highlighting quality issues in data-to-text systems will allow readers to consider this.

Keywords: Data quality, data-to-text, reasoning.

1 Introduction

In a world so reliant on big data, a large amount of decision making and reasoning is based on techniques such as data mining which is especially prevalent in policy making and resource allocation. Performance measuring variables are used to detect problematic areas in an organisation. Additional resources, such as a larger budget, can be given to the problematic area in an attempt to alleviate issues and boost productivity. This scenario can apply to a wide array of real world situations from increasing a company's profit, to improving road safety or creating prevention policies to limit damage from natural disasters. Due to ignorance or lack of awareness, the data used to make these decisions is frequently of lower quality than expected. Data-to-text systems are often used to summarise data or produce recommendations. When incorrect or incomplete data is used to make recommendations, their output could be damaging. Acting on poor recommendations may be detrimental, resulting in loss of money, or perhaps lives. Currently, the awareness and reporting of data quality issues within these systems is inadequate. This paper looks at the impact of using poor quality data to make decisions on the allocation of money to improve road safety issues.

2 Road Traffic Reports

Multiple studies have been carried out by collecting data from hospitals on traffic accident casualties, then comparing it with police reports to observe how many of these traffic casualties had been reported to the police. Casualties not reported to the police do not feature in road traffic statistics, making this data "missing" from official

reports such as the World Health Organisation's (WHO) Global Traffic Reports [2], and from datasets modelling road traffic incidents. Such models simulate road crashes and assist in decision making on how to mitigate dangerous scenarios.

For all studies investigated, less severe injuries were less likely to be reported to the police. Therefore, any datasets using police data is highly likely to have missing injury data. All studies had discrepancies between the injury severity on the police report, and the severity assigned by the hospital. These conflicts are not rectified. It is likely the hospital report is more accurate since medical staff are better qualified to assess this variable, but again it is the police data used in statistical analyses.

One study was conducted in Birmingham where data was collected at the Accident Hospital for the first 100 people who had been in a traffic accident for each month in 1970 [1]. All fatalities were reported, but reporting lessened with injury severity level. Ignoring entries with conflicts for simplicity, 255 serious injuries were reported while 64 (20%) were not, and 528 slight injuries were reported while 289 (35%) were not. This is typical of findings across studies, mostly because reporting to the police is not always legally required, such as for single vehicle crashes where no one else is affected by the incident (80% missing), or situations where only property was damaged.

The most striking figure in this study is that 98% of pedal cyclists who were injured with no other vehicle involved did not report a road accident to the police. Therefore, situations where police figures are used to make decisions on cyclist safety will almost surely be incorrect. The data implies that cycling is very safe, as only 3 people had an isolated cyclist incident, however the reality is that 112 people were injured.

3 Conclusion

Studies comparing police and hospital data about road traffic incidents show clearly that the data is not as accurate as it seems to be. The impact of using low quality data to make decisions can be very high depending on the domain, especially where people's lives are at risk. Being more aware of the quality of data being used in reasoning will allow more correct conclusions and will yield greater results in the longer term. Future work by the authors will seek to improve mindfulness of data quality in data-to-text systems by highlighting issues to readers, by dynamically describing data quality issues that are present in the outputs of textual systems alongside normal output.

References

1. Bull, J. P., & Roberts, B. J. (1973). Road accident statistics-A comparison of police and hospital information. *Accident Analysis and Prevention*, *5*(1), 45–53.
2. World Health Organisation Traffic Reports, http://www.who.int/violence_injury_prevention/publications/road_traffic/en/, last accessed 2018/05/30.