Small Grant Ref. RES-000-22-2760: NumGen*

Sandra Williams
The Open University, Milton Keynes, MK7 6AA, U.K.

30 July 2009

Non-Technical Summary

NumGen was a one-year scoping study that investigated how to express numerical quantities, especially proportions (fractions, percentages and ratios), for different audiences.

Numerical quantities are extremely common in all kinds of documents. Pick up any newspaper and you will find it packed with them - “Red meat increases the risk of cancer by 67 percent” or “More than a quarter of students were awarded A grades”. It is surprising, then, that in Natural Language Generation (the study of computer applications that automatically generate documents) numerical quantities have received little attention beyond the decision of whether to output digits (such as 27) or number words (such as twenty-seven).

Another surprise was the lack of research on numerical proportions in Linguistics. What kinds of variations might we expect in proportion expressions? Suppose we asked 50 people to write a proportion in the sentence ‘......... of people believe they are smarter than average’ given data 983/1000? How many different proportion phrases would we get?

It is important to know more about how to express numerical information because different users have different needs. For example, not all users are numerate. In fact, a 2003 UK Government study found that nearly half of adults have problems understanding mathematical concepts such as percentages.

So what has NumGen achieved over its one-year lifetime? Here is a list of major activities and outputs:

- We studied variations in the way that people write numerical quantities by collecting sets of texts written by different people about the

[*] NumGen: Generating intelligent descriptions of numerical quantities for people with different levels of numeracy, a project funded by the Economic and Social Research Council (ESRC), 2008-9. Award holder: Sandra Williams.

same numerical facts. For example, we collected a set of articles from different newspapers about the 2008 A-Level results. Each article mentioned certain numerical facts such as the overall pass rate and the proportions of students who had been awarded A and B grades. We built up a collection of 110 articles on ten topics. For each article, we annotated numerical quantities with their type (decimal, whole number, fraction, etc.), value, units (e.g., kilograms), and hedging phrases (such as *more than*, *a little under*, *around*). The collection contains nearly two thousand numerical quantities. We hope to make this annotated collection available to other researchers but need to obtain permission from the publishers who hold the copyright.

- From our collection of annotated texts, or corpus, we noticed that many numerical quantities were rounded and expressed in simpler mathematical forms in the early parts of texts; for example, 25.9

- From the Mathematics Curriculum for Schools, we devised a scale of complexity for different numerical expressions. Our idea is essentially that children are taught simple mathematical concepts first (e.g., how to count from one to ten), then they are taught progressively harder concepts that build on simpler ones. Our scale corresponds to Key Stages in the Mathematics Curriculum and it is used in a Natural Language Generation system to choose a mathematical form that is appropriate for a user's level of numeracy.

- From a further study of our corpus, and from pragmatics literature on round numbers, we saw that numerical facts given in texts can be precise or approximate and that numerical hedging phrases (e.g., *something around* and *exactly*) tell the reader explicitly whether a number is precise or approximate. We investigated the relationship between hedging and rounding in numerical expressions in our corpus because it is so fundamental to an understanding of how to express numerical quantities. Our results demonstrated that round numbers tend to be hedged more often than precise numbers and that hedged numbers tend to be rounded more often than non-hedged ones. To find out whether people can recognise when a number is approximate or precise, we circulated a questionnaire to over 100 people asking them to judge whether sentences containing round and non-round numbers were approximate or precise. Our results demonstrated that they can do this reliably. We wrote an article about the corpus study and user study and submitted it to the Journal of Quantitative Linguistics.

- Following the above investigations, it became clear that the three important features needed to generate a proportion are (1) numerical type (fraction, percentage, or ratio), (2) level of precision, or degree of rounding, and (3) type of hedging phrase (approximate, greater than, less than, or exact). We expressed the problem as a computational model in a formalism known as constraint logic programming. When the model is given a proportion between 0 and 1 as input, it comes up with sets of solutions giving mathematical type, a value, and type of hedging phrase. We proved that the model covered all instances of fractions and percentages in the corpus and also that the solutions produced by the model correspond to ones that humans would choose. To do this we persuaded 50 people to complete a questionnaire in which they wrote numerical expressions in sentences. We submitted an article about the model and user study to Computational Linguistics.

- A further study with mathematics tutors is in progress to find out whether they express numerical information for people with poor numeracy in ways that correspond to the output of our computational model.

- NumGen has been presented at an international conference and at a seminar at the University of Aberdeen. In September, we will present the project at Macquarie University in Sydney.

Finally, we believe that the ESRC funding has allowed us to identify a new area of research.