A fact-aligned corpus of numerical expressions

Sandra Williams, Richard Power

Department of Computing
The Open University, Milton Keynes MK7 6AA, UK
s.h.williams@open.ac.uk, r.power@open.ac.uk

Abstract

We describe a corpus of numerical expressions, developed as part of the NUMGEN project. The corpus contains newspaper articles and scientific papers in which exactly the same numerical facts are presented many times (both within and across texts). Some annotations of numerical facts are original: for example, numbers are automatically classified as round or non-round by an algorithm derived from Jansen and Pollmann (2001); also, numerical hedges such as 'about' or 'a little under' are marked up and classified semantically using arithmetical relations. Through explicit alignment of phrases describing the same fact, the corpus can support research on the influence of various contextual factors (e.g., document position, intended readership) on the way in which numerical facts are expressed. As an example we present results from an investigation showing that when a fact is mentioned more than once in a text, there is a clear tendency for precision to increase from first to subsequent mentions, and for mathematical level either to remain constant or to increase.

1. Introduction

1.1. Why collect the corpus?

Our aim in building the NUMGEN[1] corpus was to create a resource for NLP that would provide empirical evidence on linguistic and mathematical variations in numerical expressions. These expressions are extremely common in the factual documents that many Natural Language Processing (NLP) applications generate or analyse (de Marneffe and Manning, 2008; Reiter et al., 2005; Hallett et al., 2007; Gatt et al., 2009), and have been highlighted as a key problem in communicating information to the public (Paulos, 1988).

In our own research area, Natural Language Generation (NLG), numerical expressions have received surprisingly little attention, even though communicating numerical information is an important problem since input data is wholly or partially numerical in nearly every NLG system. For example, SkillSum and GIRL (Williams and Reiter, 2008) generated feedback on basic-skills tests, but variations in the presentation of numerical data were limited to a choice between number words and digits e.g.:

> "You scored seventeen."

as opposed to:

> "You scored 17."

The CLEF answer-renderer system for generating answers to queries posed to a medical database (Hallett et al., 2007) was limited to expressing whole numbers when it could have expressed results more interestingly (and, perhaps, more usefully) as proportions:

> "Your query has returned 965 patients between 30 and 70 years of age who had a clinical diagnosis of malignant neoplasm of breast and underwent surgery. This chart displays the distribution of patients in five age groups according to their gender and time of haematoma after surgery.
> — In the 30-39 years age group there were 163 patients (2 men and 161 women): 151 patients did not have haematoma after surgery, 12 patients had haematoma after surgery.
> — In the 40-49 years … "

Some recent NLG systems summarise numerical time-series data e.g., SumTime (Reiter et al., 2005) summarises data from weather prediction systems for oil rig personnel, and BabyTalk-Doc (Portet et al., 2007) summarises data from medical monitors (such as blood-pressure monitors) for clinicians, but both of these describe numerical data in the formulaic language of professionals, e.g.:

> "1.0-1.5 mainly SW swell falling 1.0 or less mainly SSW swell by afternoon" (SumTime)

> "toe/core temperature gap rises for 7 minutes to 2.4" (BabyTalk-Doc)

Both systems would require much greater flexibility to generate comprehensible numerical descriptions for non-professionals.

There has been no research on deeper issues such as whether to generate a vague phrase (*over ten*) or a precise one (*exactly twelve*), nor any empirical investigation of the range of possible choices, or of differences among authors or genres. Yet this is an important area where infelicities are common[2], and readers may differ widely in their levels of numeracy (Paulos, 1988).

We suspect that in Natural Language Understanding (NLU) research the situation is much the same. Indeed we are not aware of any systems that can recognise that two phrases ('25.9 per cent', 'more than a quarter') describe the same

[1] Generating intelligent descriptions of numerical quantities for people with different levels of numeracy. ESRC Small Grant RES-000-22-2760

[2] In academic sources that will be nameless, we recently found '$33\frac{1}{3}$%' and '591,000.0 people'.

numerical fact[3] — an issue of some importance in information extraction, when fact-alignment would allow a system to check for consistency among statements, or to select the most precise value from a set of alternative formulations.

If we consider numerical expressions of proportion, the range of possible linguistic variation becomes vast, since a proportion such as 0.259 can also be expressed through a variety of mathematical forms: *just over a quarter*, *around one in four*, *25%*, and so on. Apart from some noteable research on numerical hedge phrases such as *more than*, *less than*, and *around* (Dubois, 1987), variations in numerical proportion expressions have been largely ignored in linguistics.

1.2. What is unique about the corpus?

The NumGen corpus is *fact-aligned*, which means that where two or more phrases express the same numerical fact, they are assigned the same fact identifier (`factID`). Such numerical facts occur both within a single text, and across texts written by different authors for different audiences. This cross-linking is akin to the semantic alignment of concepts in paraphrase corpora, e.g., Barzilay and McKeown's (2001) machine-learning alignment of pairs of paraphrases ("burst into tears", "cried") and ("comfort", "console") in their corpus of multiple translations of five classic novels into English. Our alignment differs in that it applies only to numerical expressions. Furthermore, texts on the same topic in our corpus are not strictly parallel (in the sense that not all content overlaps and ordering of information varies); they are linked only through references to the same numerical facts.

Another novelty of the NumGen corpus is that each set of texts represents a wide range of linguistic settings for numerical expressions, from the formality of scientific articles to the relative informality and high readability of popular science magazines and 'tabloid' newspaper articles. They also cover an assortment of mathematical forms of varying degrees of technical difficulty, ranging from proportional changes in risk over time to simple integers.

2. Corpus collection and annotation

The corpus was collected by searching the Internet for articles published on the same date that described the same numerical fact (or constellation of facts). Articles were downloaded from scientific journals, popular science magazines, and newspapers,[4] see table 1. Currently, the corpus consists of around 55,000 words in 110 articles on ten topics, with annotations of around 2,000 numerical expressions, about 400 of them hedged.

Corpus annotation was semi-automatic. The texts were first subjected to an automatic sentence splitter, along with a simple program for locating numerical expressions; the results were then hand-corrected and further annotated in a spreadsheet. Finally, another simple program converted the annotations to XML.

We give below an extract from a corpus article on A-Level results in the UK[5] followed by part of the annotated version.

> **A-level results show record number of A grades**
> Record numbers of teenagers have received top A-levels grades. More than a quarter of papers were marked A as results in the so-called gold standard examination reach a new high.
> The overall pass rate also rose beyond 97 per cent for the first time — the 28th straight increase — fuelling claims that A-levels are now almost impossible to fail.
> [...] Applications to university have already increased by nine per cent this year.
> According to figures released today by the Joint Council for Qualifications, 25.9 per cent of A-level papers were awarded an A grade this summer, compared to 25.3 per cent 12 months earlier — and just 12 per cent in 1990.
> *(Daily Telegraph, 14th August 2008)*

```
<SENTENCE id=3>
  <NUMEX
     id=001
     factID="ALevels02"
     type=fraction
     format=words
     hedge="more than"
     hedgeSem=">"
     Vg=0.25
     round = "y"
     Va=0.259>
    More than a quarter
  </NUMEX>
  of papers were marked A as results
  in the so-called gold standard
  examination reach a new high.
</SENTENCE>
.....
<SENTENCE id=8>
  According to figures released today
  by the Joint Council for
  Qualifications,
  <NUMEX
    id=008
    factID="ALevels02"
    type=percentage
    format=digits
    Vg=0.259
    round = "n"
    Va=0.259>
   25.9 per cent
  </NUMEX>
  of A-Level papers were awarded an
  A-grade this summer, compared to
  <NUMEX
    id=009
    type=percentage
    format=digits
    Vg=0.253
    round = "n">
```

[3] By a 'numerical fact' we mean a statement that describes an object, or set, or situation, through a numerical property (here it is the number of papers graded A in the 2008 UK A-Level exam, as a proportion of the number of papers taken)

[4] Unfortunately this wide range of sources has led to copyright problems in making the corpus available to the research community. We are in the process of writing to copyright holders to obtain permission to release their materials for research purposes.

[5] Daily Telegraph, 14th August 2008

Topic	Journal Articles	Press Releases	Science Mags.	Newspapers / Internet
Puffin population	0	1	0	13
New planet	1	0	4	13
Exam results	0	0	0	13
Red meat	1	0	0	10
and cancer	0	0	0	0
Emigration	0	0	0	7
Economic forecast	0	1	0	10
Sunbeds and cancer	0	1	0	6
Arrests of women drunks	0	0	0	8
Obesity gene	2	0	0	6
Sale of a Monet	0	0	0	12
TOTALS	**4**	**3**	**4**	**99**

Table 1: Corpus Articles and Sources

```
      25.3 per cent
     </NUMEX>
     <NUMEX
       id=010
       type=cardinal
       format=digits
       units=months
      Vg=12
       round = "y">
      12 months
     </NUMEX>
     earlier.
   </SENTENCE>
```

Numerical expressions are shown annotated with `<NUMEX> ... </NUMEX>` tags with some attributes such as mathematical form (e.g., percentage, fraction, cardinal, ratio) and units similar to Grishman and Sundheim (1995). In addition, we created the following attributes which are unique to our corpus:

- Fact-alignment, e.g., `factID='ALevels02'`
- Given value, e.g., V_G=`0.25`
- Actual value, e.g., V_A=`0.259`
- Decision on roundness of V_G, e.g., `round='y'`
- Numerical hedges, e.g., `hedge='more than'`
- Numerical hedge semantics, e.g., `hedgeSem='>'`

Fact-alignment is the means by which we track instances of numerical facts within a text and across groups of texts. In the above fragment, the fact identifier `ALevels02` denotes the proportion of exam papers with A-grades, which is expressed as *More than a quarter* (sentence 4) and *25.9 per cent* (sentence 9). In total, we found 22 instances of this particular fact in 14 texts with linguistic variations: *25.9%, one in four, 25.9 per cent, 25.9 percent, more than a quarter, more than one in four, one in four*.

We annotated the values given in the text (V_G). Once we had identified numerical facts with more than one mention, we were in some cases able to judge whether one value (e.g., *a quarter*) was an approximation of another value specified elsewhere (*25.9 per cent*). If so, we assumed that the most precise one (*25.9 per cent*) was close the actual value (V_A). Of course, if a fact only occurred once, or if the given values were all the same, then it was not possible to estimate actual values in this way.

We also implemented an automatic decision procedure to determine whether the given number V_G was round, adapting a proposal by Jansen and Pollmann (2001) based on empirical studies of number frequencies in texts. Briefly, Jansen and Pollmann suggest that round numbers are simple multiples of so-called ‘favourite numbers’, which in decimal systems are defined as members of the set $10^N * M$, where N is any integer, and M is either 1, 2, 0.5, or 0.25. In other words, a favourite number is a power of ten, either left alone, or doubled, halved or quartered. A round number is then defined as a relatively small multiple of a favourite number; Jansen and Pollman suggest the set given by $K * F$ where F is a favourite number and K is an integer from 1-20. We have preferred to implement a rather stricter criterion in which M can take only the values 1 and 0.5. This is in part an arbitrary judgement (since there are degrees of roundness), but we find it counterintuitive for instance to admit 42.5 as a round number (it can be expressed as 17 * 2.5 where 2.5 is the favourite number $10^1 * 0.25$). Under our stricter definition, the favourite numbers (counting from 1) are 1, 5, 10, 50, 100, ... and the round numbers over 20 are accordingly 25, 30, 35, 40, and so forth.

Lastly, we annotated numerical hedges. Since these often indicate that V_G is approximate (or indeed precise in the case of the hedge *exactly*), they also contributed to our estimates of V_A. We created an attribute for hedge semantics (`hedgeSem`)which can have values ‘$>$’, ‘$<$’, ‘=’, or ‘$\approx$’.

3. Example of research results

The corpus has been the subject of a number of empirical studies, one of which was the effect of discourse position on numerical expressions (Williams and Power, 2009). Brows-

Observation	Frequency	Proportion	Significance
Equal Precision	26	0.30	<0.001
Unequal Precision	62	0.70	
Increasing Precision	56	0.90	
Decreasing Precision	6	0.10	<0.001
Equal Maths Level	57	0.65	
Unequal Maths Level	31	0.35	<0.010
Increasing Maths Level	25	0.81	
Decreasing Maths Level	6	0.19	<0.001

Table 2: Precision and Mathematical Level for first and subsequent mentions).

ing the corpus, we noticed that when the same numerical fact was referenced more than once within a particular text, the first mention (typically in the heading or the first paragraph) was often expressed in a relatively approximate, non-technical way, while subsequent mentions were more precise and technical. For instance, in the previous section we gave an example where 'more than a quarter' was mentioned first and '25.9%' subsequently. These phrases differ both in precision and mathematical form (simple fraction in the first case, more technical percentage in the second).

To test the validity of this observation, we extracted from the corpus the 88 instances in which the same fact was presented at least twice in the same text, and compared first and second mentions. Each pair was classified as showing equal, increasing, or decreasing precision by two judges, with agreement of 94% ($\kappa = 0.88$, Cohen's kappa). The results (table 2) showed a clear tendency for precision to increase, and for mathematical level either to remain the same or to increase (binomial tests).

How this result should be interpreted is an interesting question. It is well-known that newspaper articles standardly begin with a summary: however, if we equate summarisation with brevity, we find that paradoxically the less precise formulation is often *longer* (compare 'more than a quarter' with '25.9%'). Perhaps the less precise formulation is more memorable, or more useful for reasoning purposes. Whatever the explanation, the finding is clear and was made possible by the use of a fact-aligned corpus[6].

4. Conclusions

To our knowledge, the NUMGEN corpus is unique in containing multiple texts describing overlapping facts, with selected numerical facts linked across and within texts. By choosing texts on the same topics from varied publications (ranging from tabloid newspapers to scientific journals), we have shown that numerical expressions differ not only in their surface form (e.g., '12%' vs 'twelve percent') but also at a deeper semantic level, through features like the mathematical form, the distinction between actual and given value, and hedges representing arithmetical relations. These features have also been annotated, so that the corpus can be used for studying a fuller range of options (both deep and superficial) in describing numerical facts, and linking them to contextual factors such as document position and intended readership. An early study has been briefly described; the corpus will continue to be used in research on generating numerical expressions and also as data for research for automatic alignment of numerical expressions in NLU.

Acknowledgments

We would like to thank members of The Open University Natural Language Generation Group and the anonymous reviewers for commenting on this paper. This research was funded by the Economic and Social Research Council under Small Grant Ref. RES-000-22-2760.

5. References

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse, France, July. Association for Computational Linguistics.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*.

B.L. Dubois. 1987. Something of the order of around forty to forty-four. *Language in Society*, 16(4):527–541.

Albert Gatt, Francois Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur, and Somayajulu Sripada. 2009. From data to text in the Neonatal Intensive Care Unit: Using NLG technology for decision support and information management. *AI Communications*, 22:153–186.

Ralph Grishman and Beth Sundheim. 1995. Appendix c: Named entity task definition (v2.1). In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 317–332. Morgan Kaufmann Publishers, Inc.

Catalina Hallett, Donia Scott, and Richard Power. 2007. Composing queries through conceptual authoring. *Computational Linguistics*, 33(1):105–133.

C. J. M. Jansen and M. M. W. Pollmann. 2001. On round numbers: Pragmatic aspects of numerical expressions. *Journal of Quantitative Linguistics*, 8(3):187–201.

John Allen Paulos, editor. 1988. *Innumeracy: Mathematical illiteracy and its consequences*. New York: Hill and Wang.

[6] The corpus has also been used in two further studies, one on the relationship between hedging and rounding, the other on planning approximate expressions (both forthcoming).

François Portet, Ehud Reiter, Jim Hunter, and Somayajulu Sripada. 2007. Automatic generation of textual summaries from neonatal intensive care data. In *Proceedings of AIME 2007*, pages 227–236.

Ehud Reiter, Somayajulu Sripada, Jim Hunter, and Jin Yu. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167:137–169.

Sandra Williams and Richard Power. 2009. Precision and mathematical form in first and subsequent mentions of numerical facts and their relation to document structure. In *Proceedings of the 12th European Workshop on Natural Language Generation*, Athens.

Sandra Williams and Ehud Reiter. 2008. Generating basic skills reports for low-skilled readers. *Natural Language Engineering*, 14(4):495–525.