Computers, Language and Speech

A joint Royal Society/British Academy discussion meeting took place on 22–23 September 1999, organized by **Professor Karen Spärck Jones FBA, Professor Gerald Gazdar FBA** *and* **Professor Roger Needham FRS**, *who report on the event.*

There has recently been a rapid development in the use of statistical techniques in both written and spoken language processing. This implies that there are important issues to address about the interaction between formal symbolic theories of language, on which language processing has been primarily based, and the new statistical approaches. The meeting held at the Royal Society focused on this interaction, and was particularly timely because there is an increasing demand for natural language processing systems that are able to cope with bulky, changing or untidy material, for instance systems that can select relevant content from text streams or summarise audio news broadcasts. Statistical methods for extracting patterns from data can help here, and improvements in information technology mean that there are now the powerful machine resources needed to apply these.

The central question addressed in the meeting was how best to combine rule-based and statistics-based approaches to natural language. More specifically, it provided an opportunity for the text and speech communities to exchange their respective findings and ideas. For the text community, the issue is how to enhance text interpretation and generation, which have been mainly done using symbolic, rule-based approaches, with corpus-based strategies that would suggest, for instance, that for a particular type of text one syntactic analysis rule is more likely to be applicable than another. For the speech community, the question is how to enrich speech recognition or production, hitherto predominantly statistics-based, with prior knowledge of a symbolic kind, for example incorporating linguistic models of syllabic structure and stress into statistical models for word recognition.

Some of the papers illustrated the interaction between statistical data and model rules for speech processing, whether in recognition or synthesis. Others were concerned with text or transcribed speech. At the same time, the papers addressed many different language levels from the components of words, through intermediate units like phrases or sentences, to discourse units like whole dialogue turns, to extended text, and even to the real world domains that underlie linguistic expressions – for example, the real world of airports, airlines, flights, meals, dates and times that underpins automated phone enquiry and booking systems that operate in this domain.

Several papers started from the use of statistical data and pushed this past words to capture larger unit regularities and hence higher-level language structure, for example conventional relationships between turns in a dialogue; others also started from the data, but attempted to leverage pattern capture by exploiting independent linguistic features, constraints or rules, for instance about word pronunciation. But the complementary strategy – starting from the rule end but modifying and developing an initial model in the light of observed usage – was also represented.

The papers illustrated a wide range of techniques for capturing statistical regularities and for representing language structure, both as exhibited in discourse and embodied in resources like grammars and dictionaries, in a way suited to linking data and rules. Again, just as the papers attacked different language levels, they also addressed different subtasks within the scope of a comprehensive language processing system, for instance from word recognition in interpretation to style constraints in text generation. They also illustrated the role of statistically-motivated approaches for some application tasks, like translation.

The meeting's practical implications were illustrated by the fact that several papers touched on the 'unknown word' problem. New compound words, or new names, for example complex company names, are a challenge for language processing systems. But procedures that rely on memory, i.e. past data, are bound to fail here, so appropriate ways of invoking rules are essential.

Finally, a number of papers addressed the inputs and outputs for work in this whole area, namely the general requirements for systematically described corpus data as input, and the evaluation of the results of data analysis, both from a methodological point of view and as illustrations

of the performance that language processors exploiting statistical resources can currently achieve.

The forum and format of the meeting, both prestigious and neutral, were well suited to an international event cutting across institutional affiliations. The fact that the meeting was organized jointly with the Royal Society enhanced the interdisciplinary character of the proceedings and was helpful in particular in reaching scientists and engineers, to complement the Academy's linguists and phoneticians. Attendance was good and included many younger researchers, as the organizers had hoped.

It is, of course, difficult to say whether a meeting held only a few months ago, in an area in which media-friendly 'discoveries' and 'breakthroughs' do not figure, and where the papers have only very recently been published, has had an effect. But the discussion that followed each paper was extensive and invariably constructive. And the feedback from those who attended was uniformly positive: one participant commented that it was the only such event at which he had felt moved to attend every paper, and the only one at which he had no cause to regret any of his attendance. More generally, the time was clearly right for such a meeting and the comments made by participants suggest that it will encourage relevant research, both by those who were there and by those who seek out and read the published papers.

The proceedings of the meeting (including reports of the discussion) were published in April 2000 by the Royal Society, in the Philosophical Transactions: Mathematical, Physical and Engineering Sciences Series A, Volume 358, Issue 1769.