International MT conference at Aston

by Margret Grindrod

The second British Computer Society conference on Machine and Machine-Aided Translation, held April 7 to 9 at Aston University in Birmingham, drew participants from all over the world.

Over fifty people took part, from the fields of computing, languages, linguistics, mathematics and lexicography, and included almost equal proportions of people from industry and academia.

Over the three days of the conference, topics discussed ranged from term banks to on-screen parsing, automatic abstracting to MT trials.

The opening address, given by Professor D E Ager of Aston University's Modern Languages department, was followed by a short presentation by Veronica Lawson on recent developments in machine translation.

The second paper, given by Dr Crist-Jan Doedens of BSO Research in the Netherlands, concerned on-screen ATN (augmented transition networks) parsing for natural languages. In ATN parsing, the grammar and parsing strategy are represented in large maps or charts, and thanks to the new high-resolution graphics and window techniques available, these can be shown on-screen. The user can allow parsing to proceed, occasionally freezing it to inspect it.

Dr Doedens' colleague from BSO, K Schubert, presented the next paper, entitled 'Interlingual Terminologies and Compounds in the DLT (Distributed Language Translation) Project'. The DLT project has adopted a modified form of Esperanto as the basis for a pivot language between source and target language pairs. The use of an interlingua, as it is called, should prevent source language ambiguities finding their way into the target language output, but often it is difficult to fulfill both the main requirements of resemblance to internationally known forms, and suitability for handling by AI (artificial intelligence) processes.

Dr P Hancox of Aston University considered the applicability of a limited MT system to indexing for libraries. He described his attempts to develop a system, using the PRECIS index language, for translating strings such as are entered in bibliographies, into French.

Moving ahead to assessment of existing, commercially available MT systems, Mr E Macklovitch, of the Translation Bureau of the Secretary of State of Canada, described a four month trial of the English to French version of *MicroCat*, produced by Weidner. During the trial, human and machine-aided translation processes were compared, and it was found that human translators who were specialists in the subject of the text to be translated were considerably faster, and that *MicroCat* did not increase productivity. This appeared to be largely due to the fact that a considerable amount of time was taken up with entering the source text into the computer, and that the raw MT output required a great deal of post-editing. After the trial, Mr Macklovitch investigated the types and frequencies of errors produced by *MicroCat*, distinguishing some that could be corrected by the user, and some which would necessitate changes being made to the system's linguistic model.

Later in the afternoon of the first day, Dr R Leermakers and J Rous, of Phillips Research in the Netherlands, presented a paper on the Rosetta project. The Rosetta method uses isomorphic grammars, where the two grammars for two different languages are attuned to one another so that the process of constructing a sentence in one language can be parallel to the process of constructing the translation of that sentence in the target language. There are four levels of analysis (morphology, surface syntax, deep syntax, semantics), and for each analytical component, there exists a comparable generative component.

In the last presentation of day one, Hilary Horsfall spoke about the interactive English-Japanese machine translation system being developed at the Centre for Computational Linguistics at UMIST. Most developers of MT systems recognise the need for some kind of human editing during the MT process, and most favour post-editing. The UMIST system, on the other hand, relies on pre-editing, in the form of interaction between the machine and the author of the source language text. Where the machine spots an ambiguity, it presents the author with a menu of possible interpretations, and requests the correct one. The reason for the choice of pre-editing is that the system is designed to be used by the (English-speaking) authors of computer software manuals for translating those manuals into Japanese, in a situation where there is a lack of sufficient Japanese-speaking post-editors. At the moment, the system is implemented in Prolog, with some routines for character handling being written in C. The module for analysis of the English text is based on Lexical Functional Grammar, incorporating the feature system of GPSG (Generalised Phrase Structure Grammar).

Day two opened with a paper by E Luctkens and M Fermont of INFODOC, Belgium, on a prototype English-French MT system. Like several other MT projects, this one concentrated on texts relating to data processing. As in the BSO project described by Dr Doedens, the

analysis of the source language text is based on ATNs. In order to cut down execution time, the analysis ends as soon as an acceptable structure for the whole sentence has been discovered. This phase is followed by structural and lexical transfer, and syntactic and morphological generation. The programs used had taken advantage of the recursive possibilities in PASCAL, but the researchers found that this wasted machine time and memory, and intend to avoid recursion in future.

A most unusual approach to parsing was demonstrated by Professor Geoffrey Sampson of Leeds University. He pointed out that parsing systems that relied on the complete grammaticality of the source text might not be sufficient to cope with natural language input, which might often not be well-formed. In simulated annealing, a new AI technique borrowed from thermodynamics, the system chooses one approximate possible parse, and generates local modifications until the best fit is achieved.

At the conference exhibition which followed, Professor Sampson provided a demonstration of this technique on a BBC micro. Other stands at the exhibition included Donald Clarke's system for constructing Chinese characters, and David Candeland's micro version of the British Term Bank project.

'Do-It-Yourself Lexicography' was the subject of Professor Frank Knowles' contribution.

Mr W Williams, of the Cambridge Language Research Unit, proposed what he called the 'cognitive linguistic unit', drawing on Masterman's concept of the 'breath group', as a new unit of language for machine translation.

Professor W Moskovitch of the Hebrew University of Jerusalem, Israel, described the 'state of the art' in MT in the Soviet Union, where it is accorded priority status. Professor Moskovitch reported that progress appears to be about five years behind that of the west, and that at the moment MT is pursued for the prestige rather than the prospect of any immediate benefits in the form of cheaper or faster translations.

Improvement in the quality of machine translations was the subject of Donald Clarke's paper. Rather than rely solely on its internal dictionary, the machine could be made to extract more meaning, both from the general context and from scanning for references in the preceding text, with the aim of producing more accurate translations.

P. Fermont (left) and Dr C. Doedens

A different application of the techniques common in MT was described by Ms K McCarthy. At the School of Information Technology and Computing, Wolverhampton Polytechnic, a project is under way to develop an automatic précis system for use in television subtitling for the deaf. Although viewers can normally cope with 120 spoken words per minute, subtitles for the deaf can generally only be read at a rate of 90 words per minute. For this reason, some form of précis is necessary. In this project, a low-level syntactic parser is employed, and then a Prolog knowledge-based system analyses the parsed text, and implements four types of text reduction rules to produce the précis. At the moment, the parser can cope with 75% of a sample text.

Donald Clarke

The last paper of the day, on microcosmic modelling and automatic abstracting, was presented by Mrs B Sharp, also of Wolverhampton Polytechnic. Like Donald Clarke, she focussed on the importance of viewing the text above the sentence level, this time for the purposes of producing automatic abstracts of a more cohesive nature.

The third and final day of the conference began with a talk given by Dr R Sharman of the Speech Research Group at the IBM (UK) Science Centre on the *Epistle* text critiquing system.

The last four papers were concerned with the fields of lexicography and terminology. Jeremy Clear, of Birmingham University, described the work carried out by Collins Publishers and Birmingham University on the construction of the COBUILD lexical database. A large corpus of both spoken and written general modern English has been built up to serve as the basis for compiling a dictionary of English for foreign learners. Computer-aided analysis of the corpus should ensure that where explanatory examples are given in the dictionary, the focus is not on rare, striking examples, but on ones which are common in naturally occurring language.

Dr W Teubert, of the Institut für Deutsche Sprache in Mannheim, West Germany, gave a paper on similar work being carried out by the Institut in setting up LEDA (the LExicographical DAta base for German), which includes inflectional and syntactic information not included in existing dictionaries for German. The large number of different verb forms presented a problem over what to include, and the team found that rules given in existing dictionaries for predicting all possible verb forms were inadequate for the construction of a 'Vollformgenerator'.

Dr W. Teubert

Richard Candeland, of UMIST's Centre for Computational Linguistics, discussed term banks as an aid to machine translation and machine-aided translation, and referred to the British Term Bank, a prototype of which had been demonstrated the previous day.

The final paper, concerning a research project investigating the potential of the computer in retrieving terminological information, was presented by Agnes Kukulska-Hulme, of Aston University's Department of Modern Languages, who had also played a major part in organising the conference.

Two speakers were, unfortunately, unable to attend. They were X Huang, of New Mexico State University, USA, who was to have spoken on 'A bi-directional Chinese grammar in a MT system', and M Nagao, of Kyoto University, Japan, who was to have presented a review of the Japanese national MT project. However, the conference provided a wide range of topics for discussion, and provided the basis for much fruitful exchange of ideas, which will doubtless further the development of MT-related research.