4 NOTES ON MR SALLIS' TALK

Precis of lecture given on 15/11/77 - PHILIP J. SALLIS.

A good deal of current work in Artificial Intellegence centres around the formalisation and construction of meaning structures as the product of communicable concepts - the communication being facilitated by natural language text.

The research describes two distinct endeavours:

First, by using an appropriate theory of Information Transfer, it attempts to establish an hypothesis of concept organization within text, showing relationships within what has been described as an Information Structure. The hypothesis also relates to the way in which this structure changes as it is operated on by the 'Thesaural Knowledge Store' of subject-related knowledge presumed to be held by an individual, and produces a permutative structure which represents what is known and interpreted regarding the text.

Second, an assuption is made about the nature of the concept-communication process for purposes of demonstrating the theory: it is not, however, regarded as a method of testing the thory as such. A model of this pre-supposed process is outlined in a transcript of the talk and is available from the author.

It is intended that a computer simulation of the model will be carried out and the whole process tested for validity. Testing and experimental design have not yet been fully defined, but comparison of output with that of human trialists has been suggested. Also the use of experts and non-experts for similar trials would seem appropriate. In any case, all and only the variables present in the model would be used by the trialists and output would necessarily have to be documented as an on-going event.

A full-text data-base using nil-redundancy and data sharing principle was described, as were methods of storage and retrieval for it. A paper pertaining to this aspect of the research is currently being prepared for publication.

Requests for papers to:

Mr. Philip J. Sallis,
Centre for Information Science,
The City University,
St. John Street,
LONDON EC1.

5 NOTES ON MR CLARKE'S TALK

The Resolution of Ambiguities by Computer

Douglas CLARKE
Department of Mathematics
Cranfield Institute of Technology
Cranfield, Bedford, England

Used in traditional mode, a digital computer can identify ambiguous sentences in input text, but cannot necessarily resolve them.

In on-line interactive (i.e. conversational) mode, the computer can identify and can also resolve ambiguous sentences presented to it by the user. In such person-with-computer dialogue, the computer analyses each sentence presented to it. If the sentence is ambiguous, the computer retorts with the possible meanings and asks which meaning is intended.

Such a process simulates aspects of person-with-person dialogue. Nevertheless, between persons, the one being informed can assume meanings not intended, perhaps not even realised, by the informant. The procedure described will recognise meanings possibly otherwise overlooked, and will ensure that only the intended meanings are passed on.

Some of the instances treated as ambiguous would nevertheless not be so regarded by a person reading the text, in virtue of contextual or pragmatic considerations. The avoid unnecessary dialogue to resolve such ambiguities, and to reduce possible user resistance, the incorporation of procedures to perform context and pragmatic analyses is proposed. These would be resorted to initially to resolve an ambiguity, with the facility nevertheless for the user to countermand any results thereby obtained. Should these procedures fail, resort is made to the basic question-and-answer procedure to resolve the ambiguity.

This procedure, with an example of text involving some ambiguous sentences, indicates a line of research with wide ranging applications, e.g. in machine translation, in computerised mailing systems, and in the presentation of problems to a computer in natural language.

NOTES ON PROFESSOR LOH'S TALK by BRUCE SHAAK

Proceedings of a meeting held on September 19th 1977 jointly sponsored by the BCS Natural Language Translation Specialist Group and the Association for Literary and Linguistic Computing. It took place at King's College, London, and was chaired by Mr David Wigg, Treasurer of the Specialist Group, in the absence of the Chairmen of the two bodies. About 30 people attended.

Professor Shiu-Chang Loh, FBCS, of the Chinese University of Hong Kong, entitled his talk "Machine Translation of Chinese Mathematical Articles".

Professor Loh began by demonstrating the enormity ofthe information explosion - about 1000 books were published in Europe in 1500; by 1950 the annual output was 120,000; currently world output is about 1000 new titles per day. This covers a large number of languages and needs a large number of translators. Today the EEC employs about 1500 and Japan trains about 4000 each year. All this costs a large amount of money.

Attempts at mechanised translation have been increasing considerably over the last 30-40 years. Serious attempts in the USA took place during the 50's and 60's and millions of dollars were spent without any real success. In 1967 ALPAC (Automatic Language Processing Advisory Committee) published a report admitting defeat and stopping financial support for further research into machine translation since "... there seemed no possibility of tangible success in the near future."

Professor Loh and his team were fortunately late in receiving this report and in 1969 they began work on a Chinese to English translation system. It was a small team with small aims; the objectives were limited to dealing initially only with simple sentences and considerable pre-editing was accepted. At first a lot of pre-edit indicators were needed but the current system only requires the insertion of about 16 separate indicators.

Indicators are needed to assist the system with certain difficult points. Chinese verbs have only the present form, the active and passive voice are not grammatically distinguishable, nouns are the same in the singular and the plural and Chinese has no articles. Professor Loh has found that the use of articles in English is very difficult to deal with. Even English experts have different opinions on the subject.

The team aimed first at scientific texts and found that of the 49,000 possible Chinese symbols a list of about 2,000 would be sufficient for most scientific texts. (Newspapers and regular usage employ about 3,500). Once the Chinese text has been pre-edited by someone who knows only Chinese but has had special training in the system requirements, the Chinese symbols and indicators (special characters) are given numeric codes. The codes used for the symbols are the internationally accepted telegraphic four-digit codes. The codes can usually be committed to memory by the coder after a reasonable length of time. (NB Current developments include a Chinese-character keyboard which will enable the coding to be carried out automatically).

One of the problems of mathematical texts is the presence of formulae. These are dealt with by coding a spacing size indicator which makes it possible to insert the formulae in the English output text after processing. After pre-editing and coding the data is then punched up for computer processing.

The initial passes carry out validation and attempt to identify the symbol groupings; Chinese symbols are read in groups of 1 to 5 to give "words" in the English sense. The identification works on the "largest match principal" Once this has been done the system attempts to use the grammatical information from the dictionary to resolve the sentence structure. Error reports are produced during these phases where the system suspects mis-codings or requires more information (e.g. indicators wrong or no dictionary entry).

When the errors have been dealt with and the sentence structure is clear to the system the translation and output phases take place, producing printed target-language output. The output is proof-read by experts in the relevant subject. The experts do not have to be able to check against the original. This is done regularly by language experts, but as a separate exercise to ensure that the translation quality remains high.

The system was used initially on Chinese journals in Physics and then work began on Mathematics. The main difficulty is that of maintaining an up-to-date dictionary. In fact Professor Loh's team had to compile their own for Mathematics since none was available. In addition science experts dealing with new subjects frequently invent new words or new meanings according to their own rules.

It is difficult to compare the speed of the computer translation system against manual translation, but it is relevant to note that the "official" translation of Acta Mathematica Sinica appears about 18 months after the original. The version produced using CULT (Professor Loh's system) is issued about one month after the original. The quarterly journal has been translated by CULT regularly over the last three years. The processing speed was stated to be about ten pages per minute on the current ICL 1904A.

At present there are about 100 subscribers to the CULT edition and thus the exercise is not paying its way. However, it is hoped that work will soon commence on Chinese journals in other fields - in particular, medicine (acupuncture) and geology (seismology). Other developments include attempts to translate from English into Chinese using similar principles and, as far as the sentence structure is concerned, similar programming procedures.

Mr Goshawke (Specialist Group Secretary) thanked Professor Loh on behalf of those present for an interesting and entertaining talk. He had heard Professor Loh lecture at the recent Commission of the European Communities Conference in Luxembourg and it appeared that his was the only system presented there which was actually in operation. He hoped that Professor Loh would be able to extend his system to deal with other languages, in particular those of the EEC.

BRUCE SHAAK