11.4 "New Directions in Machine Translation"

Talk given by Dr. Richard Sharman from IBM on 19 November 1992 at King's College, London.

Dr. Sharman first explained that in the 70's and early 80's great progress had been made in Speech Recognition techniques using statistical methods based on hidden Markov Models. These advances led to the use of statistical modelling based purely on a large corpus of spoken material. These methods are now used by all modern speech recognition systems.

In 1988 a paper was published by IBM Research in the 'Coling' 1988 Conference, on "Stochastic Modelling in MT" in which the use of these statistical techniques was proposed for automating Machine Translation.

IBM then started a project using these techniques based on the English and French versions of the Canadian Parliament Hansard.

In 1989 a prototype IBM system was getting 40% of short sentences correct. In 1991 the figure was up to 60%. Recent work has concentrated on translating long sentences.

In 1991 the Defence Advanced Projects Research Agency (DARPA) funded a number of major systems developing these techniques for MT, on a competitive basis. System trials are now run every six months and the better systems get more financial support. These trials are also performed by a Systran system for comparative purposes.

These trials have themselves developed and changed with experience and the last one was based on text from the Wall Street Journal, about mergers and acquisitions. Despite the fact that the IBM system was based on the Canadian Hansard parallel text it still performed extremely well.

The basic idea behind these new methods is that by automatically comparing the words in a sufficient number of translated sentences in a pair of languages it is possible to estimate quite accurately what the translation of individual words and phrases would be in other randomly chosen sentences (with the same vocabulary).

IBM extracted, automatically by computer analysis, 20-30 million paired sentences from the Canadian Hansard, comprising 120-140 million words. Dr. Sharman outlined some of the problems encountered but stressed that the work was carried out automatically by computer program without any other human intervention.

After finding paired sentences the next problem was to obtain individual word alignments. This was achieved statistically by counting the number of times individual pairs of words were found in sentence pairs, and then computing the most likely word alignment of each sentence pair.

Put simply, translation of other sentences using the same vocabulary is then carried out by computing the most likely word sequence in the target language, given the actually occurring word sequence in the source language.

It is hoped that with greater experience, and with more powerful, perhaps specially designed, computer processors the system will one day be able to handle speech to speech translation in real time.

After his talk Dr. Sharman was bombarded with questions from an obviously very interested audience.

One questioner wondered whether some human intervention based on linguistics and/or existing translation expertise wouldn't increase the efficiency of the system. Dr. Sharman replied that they had indeed tried to do this experimentally but objective evaluation showed that results were usually not as good as leaving the system to work completely automatically.

He pointed out that they felt that a big advantage of the system was that given a sufficient body of translated material they could rapidly develop a translation system for any pair of languages. He thought that this largely automatic system by-passed the advantage claimed for using an interlingua to reduce the number of language pairs involved to a minimum. With their system the number of language pairs was largely immaterial.

Note. Dr.Sharman will be pleased to supply further information, including copies of his slides, on application to him at IBM United Kingdom Laboratories Limited, Hursley Park, Winchester, Hampshire SO21 2JN, UK.

J.D.W.