State of the Art Reports

Natural Language Translation

Computer-based translation systems and tools

John Hutchins

Machine translation (MT) is still better known for its failures than for its successes, and labours under misconceptions and prejudices from the ALPAC report of more than thirty years ago. The idea of developing fully automatic general-purpose systems capable of near-human translation quality has been long abandoned. The aim of MT research and related activities is to produce aids and tools for professional and non-professional translators which exploit the potentials of computers to support human skills and intelligence. This research is now taking place in the context of rapid growth in the use of MT systems and translation tools, and is thus inevitably more oriented towards specific needs than some of the more idealistic research of previous decades. The following brief survey emphasises European developments.

The recent growth of MT

The traditional MT user has been the large multinational company, requiring technical documentation and operating manuals in a range of languages. The system runs on a mainframe and produces 'raw' output of variable quality for revision (post-editing) by translators. A successful alternative has been the pre-editing of input texts (typically with a controlled 'regularised' language) to minimise the expensive editing processes. Both these types of MT use are continuing to expand rapidly. There are now millions of pages of translation produced every year (See the reports in MT News International no.6, September 1993, and no.12, October 1995).

Although MT software for personal computers began to appear in the early 1980s (with the Weidner MicroCAT system becoming particularly successful), it has been during the current decade that sales of these systems have shown a dramatic rise. There are now estimated to be some 1000 different MT packages on sale (when each language pair is counted separately.) The products of one vendor (Globalink) are present in at least 6000 stores in North America alone; and in Japan one system (Korya Eiwa from Catena, for English-Japanese translation) is said to have sold over 100,000 copies in its first year on the market. Nearly all the Japanese computer and software companies seem to have a product (Fujitsu, Toshiba, NTT, Brother, Catena, Matsushita, Mitsubishi, Sharp, Sanyo, Hitachi, NEC, Panasonic, Kodensha, Nova, Oki, etc.), mainly for Japanese and English in both directions. Outside Japan, products come mostly from independent companies set up to develop and market a range of translation software products (e.g. AppTek, CITAC, EJ Bilingual, LEC, Neocor, PC-Translator, and Globalink).

Though it is difficult to establish how much of the software purchased is regularly used (some cynics claim that only a very small proportion is tried out more than once), there is no doubting the growing volume of 'occasional' translation, i.e. by people from all backgrounds wanting gists of foreign text in their own language, or wanting to communicate in writing with others in other languages, however poor the quality. It is this latent market for low-quality translation, untapped until very recently, which is now being discovered.

Vendors of older mainframe systems (Systran, Fujitsu, Metal, Logos) are being obliged to compete by downsizing their systems; many have done so with success, managing to retain most features of their mainframe products in the PC-based versions. Systran Pro, for example, is a Windows-based version of the successful system developed since the 1960s for clients worldwide in a large range of languages; its large dictionary databases offer clear advantages over most newer cheaper products. Systran Pro is available for the language pairs (both directions) English-French, English-German, English-Spanish and for English to Italian and Japanese to English. The publishing company Langenscheidt is now marketing a PC version of Metal for German-English translation, and shortly for other languages. The competition is clearly intensifying.

At the same time, many MT vendors are providing network-based translation services for on-demand translation, with human revision as optional extras. In some cases these are client-server arrangements for regular users; in other cases, the service is provided on a trial basis, enabling companies to discover whether MT is worthwhile for their particular circumstances and in what form. Such services are provided, for example, by Systran, Logos, Globalink, Fujitsu, NEC, and MTSU (Singapore).

A further sign of the influence of Internet is the growing number of MT software products for translating Web pages. Japanese companies led the way: nearly all the companies mentioned above have a product on this lucrative market; they were followed quickly elsewhere (e.g. Systran, Globalink, Intergraph).

The most dramatic change of all has probably been the use of MT for electronic mail. Two years ago, CompuServe introduced a trial service based on Intergraph's Transcend system for users of the MacCIM Support Forum. Six months later, the World Community Forum began to use MT for translating conversational e-mail. Usage has rocketed. Most recently, Compuserve introduced its own translation service for longer documents either as unedited 'raw' MT or with optional human editing. Soon CompuServe will offer MT as a standard for all its e-mail.

The use is not simple curiosity, although that is how it often begins. CompuServe records a high percentage of repeat large-volume users for its service, about 85% for unedited MT -- a much higher percentage than might have been expected. It seems that most is used for assimilation of information, where poorer quality is acceptable. The crucial point is that customers are prepared to pay for the product -- and CompuServe is inundated with complaints if the MT service goes down! It should be remembered that France was in fact the first location for a networked MT service; this was the networking of Systran on the Minitel system; it too proved popular for many different and unexpected purposes, but globally its impact has been less than CompuServe's service.

With cheaper PC software and wider access to the Internet, there has undoubtedly been an unprecedented growth in the use of MT, and primarily among non-professional translators. It should be remembered that in multinational companies the users of MT (the post-editors) are not always professionally qualified translators, many have been specially trained to deal with MT output.

In Europe the growth has been slower than elsewhere; PC-based systems are not yet being purchased on the scale apparent in North America or Japan. In Europe, MT systems are used mainly by large translation services and by multinational companies, e.g. the software group SAP is translating some 8 millions words a year using Metal and Logos; Ericsson is a large user of Logos for translating its manuals; and the European Commission has seen a rapid growth in the use of Systran, now some 200,000 pages a year - mainly by non-linguist staff wanting translations for information purposes or drafts for writing documents in non-native languages.

In general, however, there continues to be widespread opposition among many professional translators to fully automatic systems. What they and most translation agencies and smaller companies prefer is the translator workstation approach.

Translator workstations

For professional translators, the attraction of the workstation is the integration of tools from simple word processing aids (spelling and grammar checkers) to full automatic translation. The translator can choose to make use of whichever tool seems most appropriate for the task in hand. The vendors of these systems always stress that translators do not have to change their work patterns; the systems aim to increase productivity with translator-oriented tools which are easy to use and fully compatible with existing word processing systems. The four most widely used workstations all originate from Europe: Trados' Translation Workbench, IBM's TranslationManager, STAR's Transit, Eurolang's Optimizer. In facilities and functions, each offer similar ranges: multilingual split-screen word processing, terminology recognition, retrieval and management, translation memory (pre-translation based on existing texts), alignment software for users to create their own bilingual text databases, retention of original text formatting, and support for very wide range of European languages, both as source and target languages. Integration to MT systems is now provided by three of the workstations. In the case of Trados access is provided to the Transcend software from Intergraph; IBM Translation Manager links up with Systran; and Eurolang Optimizer with Logos.

In addition Europe has been the centre for most of the background and current research on workstations. The European Commission supported the major ESPRIT-funded TWB project (1989-94) involving 10 members from companies and universities, and it has supported TRANSLEARN, an LRE project for an interactive corpus-based translation drafting tool (a prototype translation memory system) based on EU regulations and directives from the CELEX (European Union law) database. A third project (DB-MAT), this time funded by Volkswagen, is investigating the use of a domain knowledge base integrated with a linguistic database as a translation tool; the languages are German and Bulgarian.

The Translation Service of the Commission itself is now developing its own workstation: EURAMIS. The aim is to optimize the efficiency of the translation resources already available, to create a database of translated EU documents (as a 'translation memory'), and to provide easy access to MT systems. It will allow individual translators to develop their own tailor-made resources and facilities, with tools for text corpus management, glossary construction, and text alignment. A particular emphasis will be on the integration of MT and translation tools, including the mutual enrichment of Systran dictionaries and Eurodicautom lexical databases.

MT in Europe

Most of the cheaper PC-based MT software originates from Japan and the United States. In comparison, there have been surprisingly few MT systems developed and manufactured by European organisations. Two come from the former Soviet Union: the successful Stylus system for Russian-English, English-Russian, and German-Russian systems and the PARS systems for Russian and Ukrainian to and from English.

Mention has been made already of the Langenscheidt T1 system, developed from Metal jointly with the Gesellschaft fuer Multilinguale Systeme, which succeeds Siemens as Metal agent for German versions -- software and rights to the Dutch-French version are handled by the LANT company in Belgium. Also from Germany is the Personal Translator, a joint product of IBM and von Rheinbaben & Busch, based on the LMT (Logic-Programming based Machine Translation) transfer-based system under development since 1985. LMT itself is available as an MT component for the IBM TranslationManager. Both Langenscheidt T1 and the Personal Translator are intended primarily for the non-professional translator, competing therefore with Globalink and similar products.

Other PC-based systems from Europe include: Hypertrans for translating between Italian and English; the Al-Nakil system for Arabic, French and English; the Winger system for Danish-English, French-English and English-Spanish, now also marketed in North America; and the TranSmart system for Finnish-English from Kielikone Ltd.

As the methods and techniques of natural language processing become more familiar outside the research laboratories, many companies have been developing their own translation tools. Although a world-wide trend (e.g. PAHO and Smart in the US, and NHK, JICST and CSK in Japan), it is a distinctive feature of European MT activity. Both Winger and TranSmart were initially built for specific customers. In the case of TranSmart, this was developed originally as a workstation for Nokia Telecommunications. Subsequently, versions were installed at other Finnish companies and the system is now being marketed more widely. A similar story applies to GSI-Erli. This large language engineering company developed an integrated in-house translation system combining an MT engine and various translation aids and tools on a common platform AlethTrad. Recently it has been making the system available in customised versions for outside clients.

Custom-built MT has become a speciality of Cap Volmac Lingware Services, a Dutch subsidiary of the Cap Gemini Sogeti Group. Over the years this software company has constructed controlled-language systems for textile and insurance companies, mainly from Dutch to English. Probably the best known success story for custom-built MT is the PaTrans system developed for LingTech A/S to translate English patents into Danish. The system is based on methods and experience gained from the Eurotra project of the European Commission.

The most distinctive feature of the European scene is the growth of companies providing software localisation, which are acquiring considerable experience in the use of translation aids and MT systems (e.g. Logos, Metal and XL8). A forum for the interchange of experience and the establishment of standards was set up in 1990: the Localisation Industry Standards Association, publishing a newsletter ('LISA Forum') and producing a CD-Rom directory of products, standards and methods ('LISA Showcase'). Ireland, as the main centre for these services, has its own Software Localisation Group and has recently set up a Localisation Resources Centre (with support from the Irish government and EU.)

MT research

As already indicated, nearly all the major computer software companies are showing interest in developing translation tools and systems, with Japanese and American companies in the vanguard; the growth in sales of PC-based products has revealed the huge potential market.

By comparison, academic research has declined relatively in the area of written language MT as such (i.e. as distinct from speech translation and multilingual applications in language engineering). Nevertheless, there is research on developing dialogue-based systems which combine computer-interactive authoring and translating into an unknown language usually within a restricted subject field, thus ensuring higher quality output. There is much interest in exploring new techniques in neural networks, parallel processing, and particularly in corpus-based approaches: statistical text analysis (alignment, etc.), statistics-based generation from example texts, hybrid systems combining traditional linguistic rules and statistical methods, and so forth. Above all, the crucial problem of lexicon acquisition (always a bottle-neck for MT) is receiving major attention by many academic research groups, and the large lexical and text resources (such as those available from the Linguistic Data Consortium and ELRA) are being widely and fruitfully exploited. University MT research groups are increasingly working jointly with commercial organisations to develop customer-specific systems, e.g. Carnegie-Mellon University and the Caterpillar Corporation, or to undertake basic research for companies. However, the main emphasis of MT research has shifted to applications within the context of multilingual tools for specific needs and to longer-term research on speech translation.

Since the ending of Eurotra, research funds from the European Union have been more widely focused on projects within the broad field of language engineering, which includes multilingual tools of all kinds as well as translation assistance in various contexts.

Practical implementation and collaboration with industrial partners is emphasised throughout, as well as the need for general-purpose and re-usable products. Very many of these multilingual projects involve translation of some kind, usually within a restricted subject field and often in controlled conditions. Here it is not possible to describe all those projects which involve multilinguality and translation. (For details see Joerg Schuetz in MT News International no.15, October 1996.)

Speech translation

The goal of automatic speech translation was always a distant vision for MT, until developments in speech technology began to make it a feasible objective from the 1980s. At first, research was on a small scale, e.g. the project at British Telecom to translate formulaic messages over the telephone. Later, in Japan a joint government and industry company ATR was established in 1986 near Osaka, and is now one of the main centres for automatic speech translation. The aim is to develop a speaker-independent real-time telephone translation system for Japanese to English and vice versa, initially for hotel reservation and conference registration transactions.

Other speech translation projects have been set up subsequently. The JANUS system is a research project at Carnegie-Mellon University and at Karlsruhe in Germany. The researchers are collaborating with ATR in a consortium (C-STAR), each developing speech recognition and synthesis modules for their own languages (English, German, Japanese); in January 1993 the consortium gave a successful public demonstration of telephone translation.

In May 1993 began the long-term VERBMOBIL project funded by the German Ministry for Research and Technology. VERBMOBIL is intended to be a portable aid for business negotiations as a supplement to users' own knowledge of the languages (German, Japanese, English). Numerous German university groups are involved in fundamental research on dialogue linguistics, speech recognition and MT design; a prototype is nearing completion, and a demonstration product is targeted for early in the next century.

Speech translation is probably at present the most innovative area of computer-based translation research, and it is attracting most funding and the most publicity. However, few experienced observers expect dramatic developments in this area in the near future -- the development of MT has taken many years to reach the present stage: widespread practical use in multinational companies, a wide range of PC-based products of variable quality and application, growing use on networks and for electronic mail; and researchers know that there is still much to be done to improve quality.

This article has been updated from one originally appearing in ELRA Newsletter vol.1 no.4, December 1996

Copyright (c)1996 John Hutchins, University of East Anglia, UK

For later information please see the author's personal website at http://ourworld.compuserve.com/homepages/wjhutchins and also the EAMT website at http://www.eamt.org

© Copyright British Computer Society