Computer-based Translation in Europe and North America, and its Future Prospects

John Hutchins
WJHutchins@compuserve.com

Abstract: The aim of using computers for translation is not to emulate or rival human translation but to produce rough translations which can serve as drafts for published translations, as gists for information gathering, and as cross-language communication aids. The field of machine translation (MT) covers the usage, research and development of computer aids and systems ranging from production systems for large corporations to Internet aids for individuals in their own homes.

Keywords: Machine translation, Europe, America

The recent growth of MT

The traditional use of MT is the production of translations of technical documentation, e.g. for multinational companies. The system produces 'raw' output of variable quality which has then to be revised (post-edited) by translators. Post-editing can be expensive, and a successful cost-effective option is the pre-editing of input texts (typically with a controlled 'regularized' language) to minimize incorrect MT output and reduce editing processes. An important development of this usage, now expanding rapidly (with millions of translated pages every year), is the integration of translation with technical authoring, printing and publishing processes.

Although MT software for personal computers began to appear in the early 1980s, sales were relatively low until the mid 1990s. There are now estimated to be some 1000 different MT packages on sale (when each language pair is counted separately.) Quality is not good enough for professional translators, but it is found to be adequate for individual 'occasional' users, e.g. for gists of foreign texts in their own language, or for communicating with others in other languages. The quality may be poor but the demand is great.

Professional translators, translation agencies and smaller companies prefer computer-based translation tools, and in particular translator workstations, often referred to by their most distinctive component as 'translation memory' systems. The most widely used currently are: Trados Translation Workbench, Transit, Déjà Vu, SDLX, MultiTrans, Logoport, LogiTerm, Wordfast, and ProMemoria. Each offer similar ranges of facilities and functions: multilingual split-screen word processing; terminology recognition, retrieval and management; creation and use of translation memories (bilingual text corpora of previous translations and their originals); and support for all European and many Asian languages, both as source and target languages. Finally, and not least, workstations provide access to fully automatic translation if and when required.

The Internet has produced a rapidly growing demand for real-time on-line translation. The need is for fast acquisition of foreign-language information, and top quality output is not essential. Many PC-based systems are marketed for the translation of Web pages and of electronic mail, and there is great and increasing usage of MT services (often free), such as the well-known 'Babelfish' on AltaVista.

At the same time, the Internet is providing the means for more rapid delivery of quality translations to individuals and to small companies, and a number of MT system vendors now offer translation services, often with human post-editing.

MT in Europe and North America

PC-based MT software is available for many European language pairs. Here we can mention only the most notable (for a full listing see the "Compendium of translation software" available on the EAMT website: www.eamt.org). Many systems from Europe and North America cover all the major European languages (English, French, German, Italian, Spanish), e.g. Systran, IdiomaX, LogoMedia, Reverso, Transcend. There are also systems for specific pairs: T1 and Personal Translator PT (English-German), PeTra (English-Italian), TranSmart (English-Finnish), ProMT (English-Russian, German-Russian), PARS (English-Russian, Russian-Ukrainian.)

Among the wide range of languages provided by Systran are Greek and Arabic; among those by LogoMedia are Polish, Ukrainian, Turkish, Persian and Arabic. Systems specifically for Arabic and English include: TranSphere, Al-Mutarjim Al-Arabey, Al-Nakil, Al-Wafi. Finally, in addition to these systems for translating mainly to and from English, there is a rapidly growing range of systems for translation between non-English European languages, e.g. French and German, Spanish and Portuguese, Catalan and Spanish, etc.

Most of the systems mentioned above are available in different versions for large enterprises, for independent professional translators, and for occasional (home) use, e.g. for translating Web pages and emails.

Apart from commercial systems there continue to be custom-built systems for company-internal use or for clients. In the United States, the PAHO (Pan American Health Organization) developed systems for English and Spanish in the early 1980s; the Smart Corporation develops systems for most European languages for large corporations. In Europe both Winger and TranSmart were initially built for particular customers; the PaTrans system was developed specifically for LingTech A/S to translate English patents into Danish. European providers of custom-built systems include ESTeam, Xplanation n.v. and Cap Volmac Lingware Services, the latter two specializing in controlled-language systems.

Many large translation services and multinational companies use MT systems for translating large volumes of texts, e.g. in the United States government institutions (DARPA, USAF, etc.) and large corporations (Xerox, Ford, General Motors, etc.), in Europe companies such as SAP and Siemens, and in particular the European Commission as a major user of MT and translation aids. One of the most distinctive features of the European scene are translation companies providing localisation of documentation and products – these companies have acquired considerable experience in the use of translation aids and MT systems.

Many companies have websites offering information about their products and services, and increasingly these are being made available in other languages using software specifically developed for translating webpages on the fly, e.g. IBM WebSphere.

MT research

There is much interest in exploring new techniques in neural networks, parallel processing, and particularly in corpus-based approaches: statistical text analysis (alignment, etc.), example-based machine translation, hybrid systems combining traditional linguistic rules and statistical methods, and so forth.

Until the mid 1990s, most MT research was still based on the implementation of lexical and grammar rules (with translation via an interlingua or at least ‘deep structure’ representations): rule-based machine translation (RBMT). Currently, the dominant paradigms of MT research are corpus-based. In statistical machine translation (SMT), words and ‘phrases’ (word sequences) of a bilingual corpus (of original texts and their translations) are aligned as the basis for a ‘translation model’ of word-word (and phrase-phrase) frequencies. Translation involves the selection of the most probable words in the target language for each input word and the determination of the most probable sequence of the selected words (on the basis of a monolingual ‘language model’). Example-based machine translation (EBMT) involves similar alignment of bilingual data, but here the translation units are larger than individual words or short word sequences; input sentences are matched against phrases or clauses (examples) in the corpus, then equivalent phrases in the target language are extracted, and adapted and combined in acceptable output sentences. Both methods make substantial use of large bilingual corpora, but where SMT is based primarily on statistical correlations, EBMT also applies linguistics-based methods similar to those of earlier RBMT approaches.

Although most MT researchers are aiming still for autonomous translation systems, where human intervention is minimal, there are also many researching dialogue-based and computer-interactive systems, including the use of controlled or ‘regularized’ input – with the aim of ensuring higher quality output.

The most innovative area of current research is automatic translation of spoken language. The main centres are ATR in Japan, the Carnegie-Mellon University (USA), the University of Karlsruhe (Germany), all collaborating in a project (C-STAR consortium) to develop speaker-independent real-time telephone translation systems for Japanese, English and German – initially for hotel reservation and conference registration transactions. Until recently, there was also in Germany the government-funded Verbmobil project to develop a portable aid for business negotiations (German, Japanese, English). Speech translation attracts much publicity, but few observers expect dramatic developments in the near future. While we can envisage MT of speech in highly constrained domains (e.g. telephone enquiries, banking transactions, computer input) it seems unlikely that automatic speech translation will extend to open-ended interpersonal communication.

The planned accession of states in Central and Eastern Europe to the European Union has stimulated research on MT and translation tools for languages such as Czech, Polish, Hungarian, Slovenian, Estonian and Bulgarian - not just for supporting translation of treaty and other legal documents but also for enabling public access to information resources. Mention should also be made of research on systems for 'minority' languages in Europe, such as Basque, Catalan and Galician in

Spain and immigrant languages such as Hindi, Bengali and Gujarati in the United Kingdom.

Finally, the Internet has demonstrated an urgent need to replace the existing systems, developed for well-written scientific and technical documents and assuming human post-editing, by systems and translation aids which are developed specifically to deal with the kind of colloquial (often ill formed and badly spelled) messages found in emails and chat rooms, where there is no possibility of any human revision. The old linguistics rule-based (RBMT) approaches are probably not equal to the task on their own, and we may expect corpus-based methods making use of the voluminous data available on the Internet itself to form the basis of future systems for this application.

MT and human translation

Machine translation is demonstrably cost-effective for large scale and/or rapid translation of (boring) technical documentation, (highly repetitive) software localization manuals, and many other situations where the costs of MT plus essential human preparation and revision, or the costs of using computerized translation tools (workstations, etc.), are significantly less than those of traditional human translation with no computer aids.

By contrast, the human translator is (and will remain) unrivalled for non-repetitive linguistically sophisticated texts (e.g. in literature and law), and even for one-off texts in specific highly specialized technical subjects. Indeed, it is probable that the ready availability of low-quality MT output from Internet services will create a demand for high-quality human translations from people who have previously had no exposure to translation facilities.

For the translation of those texts where the quality of output is much less important, machine translation is often an ideal or even the only solution. For example, to produce translations of scientific and technical documents that may be read by only one person who wants to merely find out general background information and/or specific data, MT will increasingly be the only answer. And there are new applications where human translation has never featured: the production of draft versions for authors writing in a foreign language; the real-time translation of television subtitles; the translation of information from databases; the on-line translation of Web pages, etc.

MT in the future

The Internet will drive changes in the nature and application of MT. What users of Internet services are seeking is information, in whatever language it may have been written or stored – translation is just a means to that end. Users will want seamless integration of information retrieval, extraction and summarization systems with translation. There is now increasingly active research in such areas as cross-lingual information retrieval, multilingual summarization, multilingual text generation from databases, and so forth and, before many years, there may well be systems available on the market and the Internet.

In future years there will be fewer 'pure' MT systems (commercial, on-line, or otherwise) and more computer-based tools and applications where automatic translation is just one component. Integrated translation software would then be the norm not only for the large corporation but also for anyone from their own computer (whether desktop, laptop, or network-based, etc.) and from any device (television, mobile telephone, PDA, etc.) accessing services on computer networks.

Current commercially available MT systems are still predominantly based on methods of the older RBMT research, i.e. they utilize elaborate lexical and grammar rules and large dictionaries. The acquisition of lexical data is a major impediment to the development of systems for new language pairs and the improvement of existing systems. Corpus-based methods promise more rapid development of systems, as well as overcoming the inevitable deficiencies of human-produced rules. However, as yet few SMT systems are in the marketplace, the leader has been Language Weaver offering translation systems for Arabic, Chinese, French, German, Persian, Romanian, Spanish, etc. to and from English. Most recently, the online 'Google Translate' service has began offering its own internally-developed SMT system for Arabic, Chinese, Japanese and Korean into English – using the resources of Google's massive text databases.

W. John Hutchins is the author of articles and books on linguistics, information retrieval, and in particular machine translation - many available from his website. He is active in the European Association for Machine Translation (president 1995-2004) and the International Association for Machine Translation (president, 1999-2001).

Principal works: "Machine Translation: Past, Present, Future" (Chichester: Ellis Horwood, 1986); "An Introduction to Machine Translation" [with Harold Somers] (London: Academic Press, 1992); Editor of "MT News International" (1991-1997); Compiler of "Compendium of Translation Software" (2000 to the present) and of the "Machine Translation Archive" (2004 to the present); Editor of "Early years in Machine Translation: Memoirs and Biographies of Pioneers" (Amsterdam: John Benjamins, 2000).