[From: Gerd Willée, Bernhard Schröder, Hans-Christian Schmitz (eds.) *Computerlinguistik: was geht, was kommt? Computational linguistics: achievements and perspectives. Festschrift für Winfried Lenders* (Sankt Augustin: Gardez! Verlag, 2002), p. 159-162]

Machine translation today and tomorrow

John Hutchins

The field of machine translation (MT) was the pioneer research area in computational linguistics during the 1950s and 1960s. When it began, the assumed goal was the automatic translation of all kinds of documents at a quality equalling that of the best human translators. It became apparent very soon that this goal was impossible in the foreseeable future. Human revision of MT output was essential if the results were to be published in any form. At the same time, however, it was found that for many purposes the crude (unedited) MT output could be useful to those who wanted to get a general idea of the content of a text in an unknown language as quickly as possible. For many years, however, this latter use of MT (i.e. as a tool of assimilation, for information gathering and monitoring) was largely ignored. It was assumed that MT should be devoted only to the production of human-quality translations (for dissemination).

Many large organizations have large volumes of technical and administrative documentation that have to be translated into many languages. For many years, MT with human assistance has been a cost-effective option for multinational corporations and other multilingual bodies (e.g. the European Union). MT systems produce rough translations which are then revised (post-edited) by translators. But post-editing to an acceptable quality can be expensive, and many organizations reduce costs and improve MT output by the use of 'controlled' languages, i.e. by reducing (or even eliminating) lexical ambiguity and simplifying complex sentence structures – which may itself enhance the comprehensibility of the original texts. In this way, translation processes are closely linked to technical writing and integrated in the whole documentation workflow, making possible further savings in time and costs.

At the same time as organizations have made effective use of MT systems, human translators have been greatly assisted by computer-based translation support tools, e.g. for terminology management, for creating in-house dictionaries and glossaries, for indexing and concordances, for post-editing facilities, and above all (since the end of the 1980s) for storing and searching databases of previously translated texts ('translation memories'). Most commonly these tools are combined in translator workstations – which often incorporate full MT systems as well. Indeed, the converse is now true: MT systems designed for large organizations are including translation memories and other translation tools. As far as systems for dissemination (publishable translations) are concerned the old distinctions between human-assisted MT and computer-aided translation are being blurred, and in the near future may be irrelevant.

It is widely agreed that where translation has to be of publishable quality, both human translation and MT have their roles. Machine translation is demonstrably cost-effective for large scale and/or rapid translation of technical documentation and software localization materials. In these and many other situations, the costs of MT plus essential human preparation and revision or the costs of using computerized translation tools (workstations, translation memories, 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 highly specialized technical subjects.

However, translation does not have to be always of publishable quality. Speed and accessibility may be more important. From the beginnings of MT, unrevised translations from MT systems have been found useful for low-circulation technical reports, administrative memoranda, intelligence activities, personal correspondence, indeed whenever a document is to be read by just one or two people interested only in the essential message and unconcerned about stylistic quality or even exact terminology. The range of options has expanded significantly since the early 1990s, with the increasing use and rapid development of personal computers and the Internet.

More powerful PCs have encouraged the marketing of translation software for the general public. As general-purpose systems, the quality is inevitably poor. Input texts often contain high proportions of non-technical, colloquial language of the kind which MT systems have always found most problematic. Quality is usually not good enough for professional translators (although some use the output for drafts), but it is found adequate for individual 'occasional' users, e.g. for gists of foreign texts in their own language, for communicating with others in unknown languages, and for translating Web pages and electronic mail

It is the coming of online translation on the Internet, however, that has brought the most significant changes, with potentially far-reaching implications for the future. Exposure to information in many languages has created a rapidly growing demand, and this may well be MT's niche market: the real-time online supply of rough translations to support personal communication and information needs. The quality of the translations can be (and frequently is) ridiculed, but there is no doubt that the output is useful, particularly if the source language is not known at all and if the subject and context are familiar to some extent. The situation is unlikely to improve much (at least in the near future), but some quality improvements may come with specialization, i.e. by the development of systems designed for specific subject areas (as in the large-organization systems), or for specific document types (e.g. patents, letters), or even for specific language registers (e.g. email and text messaging). There are already stand-alone PC systems for medical translation and for patent documents, but the Internet would be the obvious home for such specialized MT systems. They will probably not be free (as many online translation services are now), but users will surely accept charges for better quality.

On the other hand, the ready availability of low-quality MT from Internet services and from commercial stand-alone software could well increase the demand for higher-quality human translation, particularly from those with no previous experience of translation. Some suppliers of online translation are already providing add-on human translation services (e.g. post-editing or full translation). Currently they are used mainly by organizations without their own translation services, but wider use may be expected in the future.

For Internet users, a desirable development would be integration with other language applications. What users are seeking is information, in whatever language it may have been written or stored – translation is just a means to that end. Many would welcome the seamless integration of translation with summarization, database mining, document retrieval, information

extraction, etc. There is already research on 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 on the Internet.

Perhaps most desired of all are systems capable of translating spoken language – not just for trained speakers in restricted domains (e.g. hotel booking and business negotiations, as in current research projects in Japan, USA and Germany), but for all speakers in all situations. Users will want reliable and accurate results – poor quality text can re-read and puzzled over, spoken output must be understood immediately. Automatic speech translation of open-ended communication will not come in the near future, and may never be possible, but in the medium term we may expect to have systems capable of translating the utterances of most speakers in well defined situations (banks, theatres, airports, rail stations, etc.)

At a more mundane level, the language coverage of all MT systems needs to be wider. Currently, most concentrate on the major commercial languages (English, French, German, Spanish, Japanese, Chinese, Korean); and many languages spoken by large populations in developing countries have been ignored by software companies and even by research groups. Equally, there is a real need for systems to deal with the kind of colloquial (often ill formed and badly spelled) language found in emails and chatrooms.

The traditional rule-based approaches found in current systems are probably not equal to these tasks on their own. In 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. Above all, the crucial problem of lexicon acquisition (always a bottleneck for MT) is receiving major attention by many research groups, in particular by exploiting the large lexical and text resources now available (e.g. from LDC, ELRA, and the Internet itself). These developments promise faster system development times, and wider deeper language coverage.

In time there will be fewer ‘pure’ MT systems (commercial, online, or otherwise) and more computer-based tools and applications where automatic translation is just one component – this will be the case particularly with specialized systems for specific users and specific domains. Integrated translation software will be the norm, available and accessible for anyone from their own computer (desktop, laptop, network-based, etc.) and from other equipment linked to networks (televisions, mobile telephones, hand-held devices, etc.). Most probably, software will no longer be bought for stand-alone computers (whether PCs or client-servers) but accessed from the Internet as and when required. Automatic translation will become an everyday and essential part of the global information society.