Machine Translation in Europe and North America: brief account of current status and future prospects

John Hutchins

PROFILE

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

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 means for accessing foreign-language information, 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.

Keywords: Machine translation, Europe, America

The recent growth of MT

Since its beginnings in the 1950s and 1960s, the traditional use of MT is the production of translations of technical documentation, e.g. for multinational companies. Systems produce 'raw' versions of variable quality which have then to be revised ('post-edited') by translators or by subject experts knowing the original language. Post-editing can be expensive, and many companies using MT adopt a cost-effective alternative, the pre-editing of input texts (typically with a controlled 'regularized' language) with the aim of minimizing incorrect MT output and reduce (or eliminating) 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.

Although MT software for personal computers began to appear in the early 1980s, sales were relatively low until the mid 1990s. The quality is not good enough for professional translation, but it is found to be adequate for individual 'occasional' users, e.g. for identifying the main content of foreign texts or for communicating in other languages.

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 - many developed initially by European companies. The most widely used currently are: SDL, Transit, Déà Vu, MultiTrans, 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 out-

put is not at all 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 (many free), such as the well-known 'Babelfish' on AltaVista - and now also available on Yahoo. Others include FreeTranslation, Google Translator, Tarjim, WorldLingo, and many more are being added both for specific language pairs and for the 'major' languages (English, French, German, Spanish, Arabic, Japanese, Korean, Chinese).

MT in Europe and North America

PC-based MT software is available from a large number of European and North American vendors and covering virtually all European language pairs. Here we can mention only the most notable (for a full listing see the *Compendium of translation software* at http://www.hutchinsweb.me.uk/ Compendium.htm). Nearly all cover the major European languages (English, French, German, Italian, Spanish), and many of them also translate from less common Languages (Greek, Polish, Russian, Hungarian, Turkish, etc.) and from and into Arabic, Chinese, Japanese, Korean, etc. In addition, there are many systems specifically designed for particular language pairs: English-German (Personal Translator PT), English-Italian (PeTra), English-Finnish (TranSmart), Arabic-English (Al-Mutarjim Al-Arabey, Al-Nakil, Al-Wafi); French-German (FB-Active), German-Russian (PROMT), Russian-Ukrainian (PARS), Portuguese-Spanish and other languages (Falatudo), Catalan-Spanish (interNOSTRUM), etc.

Most of the systems mentioned above are available in different versions such as 'corporate' or 'enterprise' for large companies; 'professional' for independent professional translators; and 'home' or 'personal' for occasional users, 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 corporate clients. In the United States, the PAHO (Pan American Health Organization) developed on-site systems for English and Spanish in the early 1980s, followed later by English-Portuguese; the Smart Corporation continues to develop customized systems for most European languages for large corporate clients; and European providers of custom-built systems include ESTeam and Xplanation n.v., the latter 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.). Major users in Europe are companies such as SAP and Siemens, and in particular the European Commission.

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. Related to this activity is the development of software for the localisation of websites. With the growth of the Internet, many companies offer information about their products and services, which increasingly needs to be made available in other languages. The information has to be updated regularly, and software such as

IBM Websphere has been developed specifically for translating webpages as and when required.

Automatic translation of news websites is growing in both Europe and North America. Most companies involved apply customized versions of MT software supplied by the major vendors such as Systran.

In contrast to the situation in Japan and other Asian countries, the application of MT to patents has been relatively neglected. There are only two systems specifically for translating patents: the PaTrans developed for LingTech A/S to translate English patents into Danish; and APTrans designed for generating multilingual patent claims from controlled English language input.

MT research

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) in what is now called rule-based machine translation (RBMT). Currently, the dominant paradigms of MT research are corpus-based. In statistical machine translation (SMT), words and 'phrases' (sequences of two or three words) from 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 exclusively on statistical correlations, EBMT applies both statistical techniques and linguistics-based methods similar to those of earlier RBMT approaches.

Significant 'by-products' of this corpus-based research have been developments of aids for translators, not just improvements in translation memories, their creation and exploitation, but also systems for error detection and correction and for automatic text prediction, i.e. suggestions for text completion to aid human translators who frequently translate similar technical documents.

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 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. 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.

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.

However, 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; the translation of electronic mail; 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 automatic 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 commercially and on the Internet.

While all-purpose MT systems will continue to be developed and marketed it seems probable that in future years

there will be many computer-based tools and applications where automatic translation is just one component. Integrated translation software would then be available 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.

Existing systems have been developed for well-written scientific and technical documents and assume human post-editing. Internet usage demands systems specifically for the kind of colloquial (often ill formed and badly spelled) messages found in emails and chat rooms. 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 as the basis of future systems for this application.

Corpus-based methods promise more rapid development of systems, as well as overcoming the inevitable deficiencies of human-produced rule-based approaches. Although SMT research now dominates MT research, the great majority of commercial systems are RBMT systems. Few SMT systems have reached public operational status. 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.

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 (now on his website) (2000 to the present) and of the Machine Translation Archive (http://www.mt-archive.info) (2004 to the present); Editor of Early years in Machine Translation: Memoirs and Biographies of Pioneers (Amsterdam: John Benjamins, 2000).

欧州と北米における機械翻訳
―現状と将来予想についての簡潔な説明―

MTの最近の進展

・80年代からPC版のMTソフトが販売されだした。
・翻訳家、翻訳業者、中小企業では、コンピュータベースの機械翻訳支援システム、なかんずく「翻訳メモリ」が利用されている。
・機械翻訳の機能だけでなく、技術文書のオーサリングシステム、印刷、出版と組み合わせた統合翻訳システムに発展している。
・インターネットの普及で、Webページの翻訳ニーズが高まり、フリーな翻訳サービスが出現した。

欧州と北米におけるMT

・欧米では、多くのMT販売業者が存在している。一覧は、http://www.hutchinsweb.me.uk/Compendium.htm　に掲載している。
・対象言語は、欧米の主流の言語（英仏独伊西）、非主流の言語（ギリシャ語、ロシア語、トルコ語等）、アラビア語、中国語、日本語、韓国語等である。
・ユーザ分類では、大企業版、翻訳専門家版、個人版に分かれる。
・特定顧客向けのMT開発もある。北米ではPAHO（パンアメリカンヘルス機構）の英語-スペイン語、英語-ポルトガル語があり、欧州では、制限言語の業者もいる。
・大口利用者としては、DARPA、USAF、XEROX、FORD、GM、SAP、シーメンス、ECである。
・日本や他のアジアと対照的に、特許機械翻訳業者は2社しかない。

MT研究

・90年代半ばまで、辞書と文法によるルールベースMTが中心であったが、現在の中心的なパラダイムは、対訳コーパスを用いたコーパスベースである。
・統計的MTは、対訳コーパスを統計処理して語と語の頻度、句と句の頻度情報を利用する。
・コーパスベースは人手翻訳者の校正支援としても利用されている。
・用例ベースMTは翻訳単位が語彙や短い句を対象とするコーパスベースに比較してより大きな単位である句や節を対象にしている。
・話し言葉のMTは大衆の興味を惹きつけているが、オープンな環境での普及には懐疑的な意見が多い。
・欧州では、EUへの新規加入国の言語や方言も対象になっている。

MTと人手翻訳

・MTは、大量で迅速、リアルタイムの翻訳に適している。特に繰り返しの多いマニュアルのローカライゼーションに適している。
・人手翻訳は、高品質な翻訳で、文学、法律、一回限りの高度に専門化した分野のドキュメントのような錬られた文章が主対象である。
・MTの新しい応用としては、外国語で書いている著者のためのドラフト版の作成、TV字幕のリアルタイム翻訳、データベースやWebページのリアルタイム翻訳、eメール等である。

将来のMT

・インターネットが大きな刺激となっている。多言語翻訳によるクロス検索、分析、要約とMTの結合、あるいはユビキタスな環境でのMTである。
・話し言葉の翻訳は非文が多く、ルールベースMTでは対応できないのでコーパスベース方式が中心である。
・MT商品の主流はルールベース翻訳である。
・統計的MTの商品は殆どない。LangWeaver社はこの分野のリーダであったが、最近Google翻訳サービスが始まった。独自に開発した統計的MTを用いている。