

Machine Translation: Potential for Progress

Ross Smith provides a status report on a controversial and much-maligned area of technological research and development

Machine Translation (MT) is the name commonly given to the discipline of creating and using computer programs to transform text in one language into text in another directly, without any human intervention. It is usually distinguished from Computer Assisted Translation (CAT), the more modest discipline of writing computer programs to help human translators in their task. The most common CAT system is called Translation Memory (TM); this uses existing translations stored in data bases to facilitate the translation of highly repetitive materials.

Machine Translation does not often make the headlines, though it can happen. In the non-English speaking world, at least, the last time the spotlight briefly fell on this normally unexciting activity was in September 1998, at the height of the Clinton-Lewinsky scandal. The report on the case by the Independent Council, Kenneth Starr, containing an account of the President's intimate dealings with Miss Lewinsky was placed on the Internet and to satisfy the morbid curiosity of surfers everywhere was immediately translated into the world's major languages using the free translation software available on a number of WWW sites. The results, not surprisingly, were laughable. A capable human being would have had a hard enough time translating such a potent combination of technical and colloquial English: the MT applications were quite out of their depth. The silliest gaffes appeared in newspapers and everyone agreed that the on-line MT programs were useless; the subject was soon forgotten.

Receiving bad press is nothing new to MT. One reason is undoubtedly a defensive reaction against what is seen as an encroachment by computers on a singularly human area, language, that which sets us apart from the beasts. Chess playing computer programs were derided in a similar fashion at first, as they vainly purported to be able to challenge humans at the quintessential mind game. As the number of humans capable of beating them declined, however, ridicule turned to grudging respect and then, with the defeat of the World Champion Kasparov at the hands of IBM's celebrated Deep Blue program, to a conflicting combination of disbelief, awe and resentment. Except at IBM, evidently. Would "human" chess ever recover? One Grand Master immediately affirmed that in the decisive match Kasparov had played like a frightened novice and Deep Blue could not beat any of the world's top 50 players in normal circumstances, but the inescapable truth was that the widely acknowledged best player of all time lost to a computer program. In MT terms that would be like a computer generating a flawless translation of Hamlet into Japanese; one can imagine the consternation that would produce in the translation profession.

For all its complexity, however, chess is vastly simpler than human language (the basic building blocks are limited to 32; there are unvarying rules; there is only one overriding purpose) and in the field of Artificial Intelligence (AI), to which they both belong, MT is progressing much more slowly than chess playing. Accordingly, it can still be laughed at, and is. Anyone wishing to write a critical piece on the subject simply needs to feed a certain type of text into an MT engine on the Internet and wait for something silly to come out. Writing containing ambiguous

terms, abbreviation and ellipsis is particularly common in English, where brevity is often sought at the cost of clarity, and this is very difficult for translation engines.

Strengths and weaknesses of MT

In any event, nobody thinks nowadays, as researchers did back in the 1950s, that high quality machine translation of complex writing is an achievable goal in the short-to-medium term. Original expectations were excessive, coming in the wake of Chomsky's generative grammar theory and the notion of universal constants in language. The slow progress since then, plus such well-publicised failures as the one described above, has made researchers wary and their writing abounds in caveats. A recent paper by Doug Arnold of Essex University, a leading British specialist, is eloquently called 'Why Translation is Difficult for Computers' (Arnold 2000) and paints a pretty desolate picture; for his part, John Hutchins, one of the world's best-known MT experts and President of the European Association for Machine Translation (EAMT), reckons that 'there is little sign that basic general-purpose MT engines are going to show significant advances in translation quality for many years to come' (1999 Singapore).

Nevertheless, machine translation can be quite impressive in the right circumstances. Efforts now focus on the creation of very practical, useable tools which can generate decent versions of texts which comply with certain syntactical and lexical prerequisites (particularly where so-called controlled languages or sublanguages are involved). The idea is to minimise ellipsis and ambiguity, seeking a clear word order which can be more easily 'mapped' from one language to another. (1) *'The man that I saw who was packing some toys'* is easier for a computer than (2) *'The man I saw packing toys'* even if it sounds stilted. For instance, the MT engine which is available to PricewaterhouseCoopers employees around the world via the firm's intranet, called Systranet and produced by market leader Systran, gave the following result for a translation into French of the two illustrative sentences quoted above:

(1) *L'homme que j'ai vu qui emballait quelques jouets* (comprehensible)

(2) *L'homme j'ai vu des jouets d'emballage* (incomprehensible)

In fact, for the MT engine the only 'easy' part of (2) is the initial subject, *'The man'*, because after that the omission of the relative pronoun and the ambiguity of *'packing'* (is it a verb participle or adjectival gerund?) cause major complications. In any event, this shows that within the restraints described above, output of reasonable quality can be obtained which requires only light post-editing.

One of the most often-cited success stories in MT is 'Météo', an automatic English-French translation system used for weather bulletins in bilingual Canada. Developed at the University of Montreal, it generates the French reports used by airlines and shipping companies, among others, with a minimal risk of error. The linguistic input is limited to a very specific number of specialised terms and constructions, i.e. it is a 'sublanguage'. The following example is illustrative:

METRO TORONTO.

TODAY ... MAINLY CLOUDY AND COLD WITH OCCASIONAL

FLURRIES. BRISK WESTERLY WINDS TO 50 KM/H. HIGH NEAR

MINUS 7.

TONIGHT ... VARIABLE CLOUDINESS. ISOLATED FLURRIES. DIMINISHING WINDS. LOW NEAR MINUS 15.

FRIDAY ... VARIABLE CLOUDINESS. HIGH NEAR MINUS 6.

LE GRAND TORONTO.

AUJOURD HUI ... GENERALEMENT NUAGEUX ET FROID AVEC QUELQUES AVERSES DE NIEGE. VENTS VIFS D'OUEST A 50 KM/H. MAXIMUM D'ENVIRON MOINS 7.

CETTE NUIT ... CIEL VARIABLE. AVERSES DE NIEGE EPARSEES. AFFAIBLISSEMENT DES VENTS. MINIMUM D'ENVIRON MOINS 15.

VENDREDI ... CIEL VARIABLE. MAXIMUM D'ENVIRON MOINS 6.

(example in Arnold 1994)

This type of prose, with very simple grammar and a restricted vocabulary, is what MT handles best.

Who uses MT, and why

In addition to the Canadian meteorological authorities, a wide range of private companies and public institutions use machine translation and computer assisted translation systems. In fact, one simple argument in favour of MT's validity is the sheer number of major businesses and bodies using it: surely so many corporate heavyweights cannot all be wrong. Companies making large use of MT and CAT include Xerox (which develops its own systems), Ericsson, Osram, Océ Technologies, SAP, Ford, Rover, General Motors, Aérospatiale and Berlitz (Hutchins 1999 Beijing). Technology and manufacturing companies such as these produce numerous manuals, user guides, technical specifications lists, etc., which are ideal fodder for MT and CAT. Translation memory systems can be particularly useful in this context. For instance, when a new version or upgrade of a product is placed on the market the documentation describing its technical features and how it should be used will probably be almost the same as for the product's predecessor: using translation memory, all the existing similar or identical materials can be taken advantage of without any effort on the part of the 'human' translator. The program seeks matching translation units in its memory and pulls them out for a quick check and approval by the translator, or even translates entire swathes of documentation by itself (when matches between current text and existing translated text are 100% and the translator has full confidence in the system). For the uninitiated, 'translation units' are pairs of matching sentences or phrases delimited by typical markers such as full stops, question marks and carriage returns.

The alternative to this would be the extraordinarily laborious process of comparing the new documentation with the old, detecting re-usable text, cutting and pasting it into the new documents and making necessary modifications. This brings to light what is seen as one big advantage of CAT for the human translator, this being that it frees him/her from the most boring and toilsome part of the translation process. Although some technical translators insist on seeing MT as a threat, in fact it is more of an aid, since it handles the most mechanical aspects of the work while the

human translator continues to be essential to translate more complex text and to review the MT output.

The rewards for private businesses of using computer assisted translation techniques are clear. Savings are achieved by reducing translators' fees and consistency is raised between different generations of products and related documents, and also, if appropriate, between different geographical locations (translation memories can easily be sent by network or email from one country to another and reused, even inverting the source and target languages if desired). A collateral effect is also to make in-house translators feel happier, since the computer takes over much of the donkey-work.

The rewards for producers of MT and CAT software are also clear, to judge by the number of them: at the latest count, there are about 80 companies around the world operating in this field (Hutchins 2000).

On a public level, one of the most important users of computerised translation resources is the Translation Service of the European Commission, which has been using the Systran MT engine for a number of years and also uses both its own TM software and the Trados Workbench TM suite to improve productivity, as well as quality and coherence. The automatic translation system is available not only to the translators but also to all Commission officials with a PC linked to the internal network. Considerable use is apparently made of the highly customised MT facility, mainly for producing fast translations of short texts with standardised terminology (mail, minutes of meetings, etc.), browsing texts in languages the user does not know and making drafts of user-authored documents in languages other than the user's mother tongue (Blatt 1998).

Another public institution that uses MT in a very big way is the Pan American Health Organisation, which over the past 20 years has developed its own specific systems for translating from English to Spanish and vice versa. These very successful systems, for general translation though evidently slanted towards health matters, are now also being licensed to external users. By the PAHO's own figures, Spanam and Engspan, as the MT applications are called, handle around 80% of translation volume in these two languages.

If two countries had to be singled out for their interest in MT these would perhaps be Finland and Japan. They share two common traits: a love of high technology and a language which is little known beyond their borders. In Finland, Nokia Telecommunications developed its own system which was later implemented in other Finnish companies and is now being marketed more widely. Nokia's great local rival, Ericsson, makes considerable use of the Logos computerised resources for translating its manuals. In Japan, specifically tailored MT engines exist for translating abstracts of Japanese scientific and technical articles into English, for translating Japanese stock market reports into English, and for translating English news articles into Japanese. Commercial English-Japanese systems abound and almost all Japanese computer and technology companies (Fujitsu, Toshiba, Sharp, etc.) apparently make and/or use a product, mainly for Japanese and English in both directions (Hutchins 1999 Singapore, 2000).

In the above cases of MT and CAT applications used by both private and public organisations, there are three overriding objectives: saving money, increasing speed and maintaining consistency, usually in that order. It is clear that these objectives are being achieved. Some organisations have been using computerised

language resources for many years, others have invested more recently. These institutions are evidently gaining something of value, otherwise they simply would not devote time or money to this area. Machine translation can additionally be used when there is a need for a rough translation but the need is not strong enough to make it worthwhile paying for a 'human' translation, or simply when no 'human' translator is available (for instance, one major user of the PricewaterhouseCoopers/Systran facility is PwC's office in Mauritius, for translation into French). If machine translation's self-declared limitations are accepted, therefore, the contempt with which this field has historically been treated now seems quite out of place.

Who does not use MT

The pragmatism and pursuit of tangible goals that have prevailed in the field of machine translation for the last two decades have effectively buried its more fantastic or romantic aspects. The notion of efficient automated translation between virtually any languages at the press of a button, bringing together the most diverse members of the human race, has returned to the field of science fiction. For a time a lot of effort was devoted among MT researchers to the creation of 'interlinguas', that is, intermediate sets of universal codes or symbols capable of expressing vocabulary and grammatical structures in any language, and therefore useable as the middle stage in translation between any language pair. This differs from the traditional rule-based 'transfer' systems which contain the grammatical rules pertaining to only the source and target languages. However, enthusiasm seems to have tailed off as the enormous difficulty of the enterprise has become apparent. In machine translation, as everywhere else, money reigns and research efforts mostly focus on potentially profitable projects. Almost all investment is centred on systems to and from English and the world's other big languages. The current trend in fact is to move away from the exquisitely Platonic 'interlingua' approach and towards the much cruder 'example-based' and 'corpora-based' methods, in which equivalent phrases and sentences in the source and target languages are lined up from existing translation memories or enormous bilingual corpora available on the Internet and searched, on a hit-and-miss basis, until the necessary translation is found (or not). The creation of such corpora requires a lot of work, which needs to be paid for, and accordingly they are not plentiful and exist only in the world's major languages. This approach has become viable now that computer processing has reached such spectacular speed. The hope in the MT community is that it can be combined with the rule-based methods to produce better quality translation, rather like the manner in which a chess program applies its knowledge of chess rules and then examines hundreds of thousands of moves which are possible in theory but senseless in practice before hitting on a feasible combination.

The languages of poorer countries are evidently not interesting from a commercial viewpoint and their own authorities have few resources to invest in natural language processing of any kind. While a few systems do exist (Somers 1997), therefore, it must be said that at present there is a tremendous bias in MT and CAT towards the languages of international commerce. This may change one day, if anyone is capable of inventing a true interlingua.

MT and the Internet

The Internet has provided developers of machine translation systems with a whole new world of opportunities. On the Internet, users demand pure machine

translation systems, capable of providing reasonable quality fast in translations of email, chat, web pages and so on. There is no place for CAT because users do not want computerised assistance for their own linguistic efforts but a finished product which they can read and understand. Compuserve, one of the world's biggest supplier of Internet services, has been offering free automatic translation of emails and chat for around five years and says that the service is very successful, with high repeat rates. Another pioneer in on-line translation was the Babelfish application offered free by the Altavista web portal using Systran technology, which is still going strong. Transparent Language, a company specialising in language technology which runs a very popular on-line translation service with 18 language pairs (www.freetranslation.com), reports that in a typical one-hour period over 2,000 translation requests are received. An impression of the volume of MT and similar services available on-line can be obtained at the www.translate-free.com site, which lists 37 different translation facilities (though in fact some are merely bilingual dictionaries). Two of the biggest and most veteran players, Systran (<http://www.systransoft.com>) and Logos (<http://www.logos-usa.com>), offer free machine translation and multilingual dictionary searches under their own brands on the Web as a means of capturing customers requiring something more powerful and customised.

To gain an idea of what these programs can actually do, the following text was selected at random from the Internet and fed into two free on-line MT services offered by Systran (<http://web.systranet.com/systran/net>) and Microsoft (<http://officeupdate.lhsl.com>). The translation to French was then back-translated into English using the same engine.

Source text:

English Today will interest everyone concerned with or fascinated by the English language: teachers and advanced students of English as a first or second language; linguists, writers, broadcasters and journalists; and anyone with a broad, general interest in English.

On-line MT service:	Translation into French	Translation back into English
Systranet (Systran engine)	L'anglais aujourd'hui intéressera chacun concerné par ou fasciné par l'anglais: professeurs et étudiants avancés de l'anglais comme première ou deuxième langue; linguistes, auteurs, animateurs et journalistes; et n'importe qui avec un large, général intérêt en anglais.	English today will interest each one concerned by or fascinated by English: professors and advanced students of English like first or second language; linguists, authors, organizers and journalists; and no matter who with broad, general English interest.
Microsoft (Lernout &Hauspie engine)	Anglais Aujourd'hui intéressera tout le monde intéressé avec ou fasciner par la langue anglaise: professeurs et étudiants avancés d'anglais comme une premier ou deuxième langue; linguistes, écrivains, speakers et journalistes; et n'importe qui avec un intérêt général, général en anglais.	English Today will interest everybody concerned person with or to fascinate by the English language: professors and English advanced students like a first or second language; linguists, writers, broadcasters and journalists; and that with a general interest, general in English.

The translations into French are considerably more accurate than the back-translation into English, and are certainly sufficient to convey the essential content of the original text. It is worth adding that most MT applications can be instructed not to translate certain words, such as the name of the publication in the above example, to provide a more polished product.

MT and English

Considering the status of English as the leading language for international communication in so many fields and the fact that 80% of Internet linguistic content is reputed to be in English, it is reasonable to assume that MT engines around the world are being used largely to translate from and into English. This assumption is borne out by the available data: there are virtually no commercial MT systems, whether on or off the Internet, in which translation to or from English is not an option (Hutchins 2000) and according to Systran, English is either the source or target language in almost all the translation requests made using its Systranet MT facility (which caters for 20 language pairs), with around 63% of translation being into English and 37% from English. Figures for the utilisation of Transparent Language's free on-line translation service (18 language pairs) indicate that by far the most popular combinations are English-Spanish and Spanish-English, with translation from English being around double the volume of translation into English in overall terms. The presence of English is in fact so overwhelming that when reference is made to 'machine translation' what is actually being referred to, almost without exception, is automated translation into and from the English language.

The use of MT to translate e-mail, chat and site content over the Internet and messages sent over corporate networks may be expected to contribute to the forging of a simplified international version of English. The commencement of such a phenomenon can be observed now, though whether it will eventually crystallise in a simplified, standardised Global English cannot yet be predicted with any accuracy. Except in the event of extremely unlikely geopolitical upheavals on an enormous scale, English will continue to be a dominant linguistic presence for some time; what we cannot know is the form which that presence will take. Anyone who has worked in an international or multinational company will be familiar with the typical communications in a kind of English between non-native speakers of different nationalities, whether spoken or written. The language is simplified: auxiliary verbs vanish, irregular endings become regular, prepositions are almost always wrong, yet communication of the most important information is somehow achieved. Something similar happens in machine translation, as long as the input is not too poor. If this situation continues long enough, one can imagine a kind of 'business pidgin' eventually emerging to which MT will make a contribution, and this contribution may grow as the programs improve and as more and more people have access to on-line facilities.

If near-perfect machine translation one day becomes widely available and if it is fast and easy to use, then the importance of having a *lingua franca* will diminish radically. The Babelfish era will have arrived, in which the need to learn English will become as obsolete as the need to learn long division after the pocket calculator had been perfected. The ultimate machine translation engine is unlikely to take the form of a small fish which one inserts in the ear (like the Babelfish made famous in the satirical science-fiction series 'The Hitchhiker's Guide to the Galaxy'); rather, word-processing and speech-processing applications, whether on desktop, laptop or palmtop machines, will come equipped with an MT button,

providing instantaneous translation of the selected text into the language set by the user. English may continue to rule the linguistic roost, perhaps alongside Spanish and Chinese, but learning it will no longer be an indispensable prerequisite to participation in the international community.

All this may sound utopian from the viewpoint of natural language processing and it certainly contrasts with the rather forlorn views currently prevalent among researchers. Their pessimism stems essentially from how slowly MT seems to be progressing: 'overall it has to be admitted that at present the actual translations produced do not represent major advances on those made by the MT systems of the 1970s', comments John Hutchins (1999 Singapore). For his part, Doug Arnold succinctly lists four decisive limitations of computers, namely their inability to: (i) perform vaguely specified tasks; (ii) learn things (as opposed to being told them); (iii) perform common sense reasoning; (iv) deal with some problems where there is a large number of potential solutions (*combinatorial explosion*), ominously adding that: 'The bad news, from an MT perspective, is that each of these limitations is relevant' (2000).

The good news, however, may be that these limitations will not *always* be relevant, or even applicable. Yves Champollion, a professional translator and polyglot descendant of the famous Egyptologist, as well as being the creator of a CAT application called Wordfast, takes a quite different view. He argues as follows: 'MT software is still in its infancy. And, given the pace of development in the computer industry, we may see, sooner than expected, an MT solution that provides decent translation. All it takes is a resourceful computer and better MT software. The hardware is here. The software will inevitable follow.' Yves Champollion considers that advances in areas such as neural networking and cybernetics, together with the compilation of large knowledge bases in the form of dictionaries, glossaries and translation memories, will soon have an impact on MT capabilities (Champollion 2001). For Champollion, the future of 'human' translation lies in proof-reading computer output.

The general feeling among academics specialising in machine translation, though it may not be explicitly expressed, is that they have got about as far as they can with the current techniques. As mentioned above, new approaches based on large bilingual corpora have provided some encouragement, but the potential of such methods in themselves is evidently limited. Essentially, a technical or methodological breakthrough is needed for MT research to pick up speed again, which could come from inside the discipline of machine translation itself or from another related area of Artificial Intelligence. Sheer computing power could bring opportunities that are impossible at present, as occurred with chess programs. Research on neural network methods, which are basically intended to make computers function in the same way as human brains, might throw up currently unimaginable solutions.

Machine translation is a relatively young technology derived from the more mature fields of computing and linguistics, and considering the amazing progress in IT over the last 20 years and the fact that theoretical linguistics is thriving as never before, the potential for progress is enormous. In any event, whoever is right about the future of MT it is clear that the rewards of developing a program capable of producing reliable translations to a reasonable standard would be so immense that both academic researchers and commercial developers can be expected to continue their efforts in the foreseeable future.

REFERENCES:

John Hutchins (University of East Anglia):

'Retrospect and prospect in computer-based translation', Proceedings from the Singapore MT Summit, 1999;

'The development and use of machine translation systems and computer-based translation tools', Proceedings from the International Symposium on Machine Translation and Computer Language Information Processing, Beijing, 1999

'Compendium of Translation Software, commercial machine translation systems and computer-aided translation support tools, EAMT, 2000

(The above papers by John Hutchins are available at:

<http://ourworld.compuserve.com/homepages/WJHutchins/>)

D.J. Arnold (University of Essex):

'Machine Translation: An Introductory Guide' Blackwells-NCC, London, 1994

'Why Translation is Difficult for Computers', 2000 (to appear in H.L. Somers (ed) *Computers and Translation: a handbook for translators*, John Benjamins)

Harold Somers (UMIST):

'Machine Translation and Minority Languages', ASLIB Translating & the Computer Conference, 1997

Yves Champollion (Champollion & Partners):

'Machine Translation and the Future of the Translation Industry', Translation Journal Vol. 5 No.1, 2001

Achim Blatt (European Commission Translation Service)

'Workflow using linguistic technology at the Translation Service of the European Commission', 1998 EAMT Workshop Proceedings, Geneva, 1998