Multiple Uses of Machine Translation and Computerised Translation Tools

John Hutchins
web: http://www.hutchinsweb.me.uk

Abstract

For many years MT systems and tools were used principally for the production of good-quality translations: either MT in combination with controlled (restricted) input and/or with human post-editing; or computer-based translation tools by translators. Since 1990 the situation has changed. Corporate use of MT with human assistance has continued to expand (particularly in the area of localisation) and the use of translation aids has increased (particularly with the coming of translation memories). But the main change has been the ever expanding use of unrevised MT output, such as online translation services (Babel Fish, Google, etc.), applications in information extraction, document retrieval, intelligence analysis, electronic mail, and much more.

1. Traditional uses

Machine translation (MT) has a long history – it is 60 years since Warren Weaver's memorandum of July 1949 launched research on the topic. For most of that history – at least 40 years – it was assumed that there were only two ways of using MT systems. The first was to use MT to produce publishable translations, generally with human editing assistance ('dissemination'). The second was to offer the rough unedited MT versions to readers able to extract some idea of the content ('assimilation'). In neither case were translators directly involved – MT was not seen as a computer aid for translators.

The first MT systems operated on the traditional large-scale mainframe computers in large companies and government organizations. The outputs of these systems were then revised (post-edited) by human translators or editors familiar with both source and target languages. There was opposition from translators (particularly those with the task of post-editing) but the advantages of fast and consistent output has made large-scale MT cost-effective. In order to improve the quality of the raw MT output many large companies included methods of 'controlling' the input language (by restricting vocabulary and syntactic structures) – by such means, the problems of disambiguation and alternative interpretations of structure could be minimised and the quality of the output could be improved. Companies such as Xerox used MT systems with a 'controlled language' from the late 1970s – many companies followed their example, and the Smart Corporation specialises to this day in setting up 'controlled language' MT systems for large companies in North America. In a few cases, it was possible to develop systems specifically for the particular 'sublanguage' of the texts to be translated (as in the Météo system for weather forecasts). Indeed, nearly all systems operating in large organisations are in some way 'adapted' to the subject areas they operate in: earth moving machines (Caterpillar), job applications (JobBank in Canada), health reports (Global Health Intelligence Network), patents (Japan Patent Information Office), health and social affairs (Pan American Health Organization), police data (ProLingua), software (SAP), and many more. These large-scale applications of MT continue to expand and develop, and they will do so into the foreseeable future.

Included in such expansion will undoubtedly be the further application of MT to the localisation of products. Localization became a specialist application of MT and translation memories in the early 1990s. Initially stimulated by the need of software producers to market versions of their systems in other languages, simultaneously or very closely following the launch of the version in the original language (usually English), localisation has become a necessity in the global markets of today. Given the time pressures, the many languages to be translated into, MT seemed the obvious solution. In addition, the documentation (e.g. software manuals) was both internally repetitive and changed little from one product to another and from one edition to the next. It was possible to use translation memories and to develop 'controlled' terminologies for MT systems. The process involves more than just translation of texts. Localisation means

the adaptation of products (and their documentation) to particular cultural conditions, ranging from correct expression of dates (day-month-year vs. month-day-year), times (12-hour vs. 14-hour), address conventions and abbreviations, to the reformatting (re-paragraphing) and re-arranging of complete texts to suit expectations of recipients.

The second use ('dissemination') was initially rather reluctantly conceded by MT researchers. With the coming of MT software on microcomputers or personal computer (PC) systems the situation changed. Although intended for professional translators for the production of publishable translations (e.g. the systems in the early 1980s from ALPS and Weidner), they were soon followed by systems (many from leading Japanese manufacturers of PCs) which were clearly intended both for translators and for non-translators ('occasional translators') mainly interested in the 'assimilation' function. Such PC systems now cover an increasingly wider range of language pairs and run on a wide range of operating systems. As long as desktop PCs continue to be manufactured and used, this method of delivering MT will continue. What has always been uncertain is how purchasers have been using these systems. In the case of large-scale (mainframe) 'enterprise' systems it has always been clear that MT is used to produce drafts which are then edited by bilingual personnel. This may also be the case for PC systems, i.e. it may be that they have been and are used to create 'drafts' which are then edited to a higher quality. On the other hand, it seems more likely that users are 'occasional translators' who want just to get some idea of the contents (the basic 'message') of foreign texts and are not concerned about the quality of translations. This usage is generally referred to 'assimilation' (in contrast to the other aim: 'dissemination'). We know (anecdotally) that some users of PC MT systems have trusted them too much and have used 'raw' (unedited) MT translations as if they were as good as human translations – probably by users unfamiliar with the target language and unaware of the problems of translation by computer. However, it is an unfortunate fact that we do not know in any detail how PC systems have been and are being used. We know that sales of systems continue to be high enough for manufacturers to remain in business over many years, but it is suspected by many observers that purchasers rarely use systems after their initial enthusiasm, once they learn how poor the quality of MT output can be.

The MT engines of both mainframe (client-server) and PC systems are overwhelmingly 'general purpose' systems, i.e. they are built to deal with texts in any subject domain. As mentioned, 'enterprise' systems (particularly controlled language systems) usually concentrate on particular subject areas. By contrast there are few PC-based subject-specific systems: exceptions are versions of the English/Japanese Transer system for medical texts and for patents. On the whole, however, PC systems deal with specific subjects by making available subject glossaries, which can be ranked in preference by users. For some PC systems the range of dictionaries is very wide, embracing most engineering topics, computer science, business and marketing, law, sports, cookery, music, etc. How much they are used in practice is of course unknown.

2. Aids for translators

For most of MT history, translators have been wary of the impact of computers in their work. They obviously did not want to be 'slaves' to mainframe MT output – post-editing what they could do more quickly and accurately than the machines. Many saw MT as a threat to their jobs – little knowing the inherent limitations of MT. During the 1980s and 1990s the situation changed. Translators were offered an increasing range of computer aids. First came text-related glossaries and concordances, word processing on increasingly affordable microcomputers, then terminological resources on computer databases, access to Internet resources, and finally (most significantly of all) translation memories. The idea of storing and retrieving already existing translations arose in the late 1970s and early 1980s, but did not come to fruition until the availability of large electronic textual databases and with facilitating bilingual text alignment. The first commercial translation memory systems came in the early 1990s (Trados, Transit, Déjà Vu, WordFast, etc.) All translators are now aware of their value as cost-effective aids, and they are increasingly asking for systems which go further than simple phrase and word matching – more MT-like facilities in other words. With this growing interest, researchers are devoting more efforts to the real computer-based needs of translators. As just two examples there are the TransSearch and TransType systems: the first a sophisticated text concordancer, the second exploiting translation memories by predicting the words a translator may select when translating a text similar to ones already translated.

3. Special devices, online MT

From the middle of the 1990s onwards, mainframe and PC translation systems have been joined by a range of other types. First should be mentioned the obvious further miniaturisation of software: the numerous commercial systems for hand-held devices. There are a bewildering variety of "pocket translators" in the marketplace. Many, such as the Ectaco range of special devices, are in effect computerized versions of the familiar phrase-book or pocket dictionary, and they are marketed primarily to the tourist and business traveller. The dictionary sizes are often quite small, and where they include phrases, they are obviously limited. However, they are sold in large numbers and for a very wide range of language pairs. As with PC systems, there is no indication of how successful in actual use they may be – it cannot be much different from the 'success' of traditional printed phrase books. (Users may be able to ask their way to the bus station, for example, but they may not be able to understand the answer.) Recently, since early in this decade, many of these hand-held devices have included voice output of phrases, an obvious attraction for those unfamiliar with pronunciation in the target language.

With the widespread and growing use of mobile telephones, there are an increasing number of manufacturers providing translation software for these devices. MT is an obvious extension of their text facilities. The range of languages is so far not very wide, limited on the whole to the 'commercially dominant' languages: English, French, German, and Spanish. In some cases, the translation software is built-in. But now, more frequently, the translation software is accessed from Internet database servers – which can therefore provide large dictionaries and some linguistic processing. The next obvious development is the use of mobile devices as terminals for online MT services.

This has indeed been one of the most significant changes since the middle of the 1990s: the availability of free MT services on the Internet. Online MT services appeared in the early 1990s but they were not free. In 1988 Systran in France offered a subscription to its translation software using the French postal services Minitel network. At about the same time, Fujitsu made its Atlas English-Japanese and Japanese-English systems available through the online service Niftyserve. Then in 1992 CompuServe launched its MT service, initially restricted to selected forums, but which proved highly popular, and in 1994 Globalink also offered an online subscription service – texts were submitted online and translations returned by email. A similar service was provided by Systran Express. However, it was undoubtedly the launch of AltaVista's Babelfish service in 1997 (based on the various Systran MT systems) that caused the greatest publicity. Not only was it free but results were (virtually) immediate. Within the decade, the Babelfish service has been joined by FreeTranslation (using the Intergraph system), Gist-in-Time, ProMT, PARS, Microsoft Windows Live Translator, and many others; in most cases, these are online versions of already existing PC-based or mainframe systems, and primarily rule-based. The exception has been the latest entrant, the Google Translate, based on the latest developments in statistical MT – the coverage of languages is expanding rapidly beyond competitors, and the text resources are vast. The great attraction of online MT services was (and is) that they are free to users (even if not to providers) – it is evidently the expectation of the developers is that free online use will lead to sales of PC translation software (when available), although the evidence for this has not been shown; or that it will encourage the use of the fee-based 'valued-added' post-editing services offered to users (e.g. by FreeTranslation). Whether any of this has in fact happened is not known.

While online MT has undoubtedly raised the profile of MT for the general public, there have, of course, been drawbacks. To most users 'discovering' online MT services the idea of automatic translation has been (and is) something completely new – despite the availability of PC translation software. Attracted by the possibilities, many users 'tested' the services by inputting for translation sentences containing idiomatic phrases, ambiguous words and complex structures, and even proverbs and deliberately opaque sayings. A favourite method of 'evaluation' was back translation ('to-and-fro' translation), into another language and then back into the original – a method which might appear valid to the uninitiated but which is not satisfactory. Not surprisingly, they found often that the results were unintelligible, they found that MT was liable to much 'faulty' and 'inaccurate' results, that MT suffered from many limitations – all well-known to company users and to purchasers of PC software. Numerous commentators have enjoyed finding fault with online MT and, by implication with MT itself. Users have undoubtedly been gravely disappointed by the poor quality; there is no doubt that the less knowledge users have of the language of the original texts the more value they attach to the MT output.

However, we know very little (indeed almost nothing) about who uses online MT and what for. We do not know their ages, backgrounds, knowledge of languages, we do not know how many translate only

into their native language, how many use online MT to translate into an unknown foreign language, how many are translators using MT as rough drafts, how many use the subject glossaries available, and so forth. Almost all that we do know are the surprising facts that translation of web pages is very much a minor use (no more than about 15% at best), that the average length of texts submitted is just 20 words, and that more that 50% of submissions are one- or two-word phrases. It had been anticipated that longer texts would be submitted – the general maximum length of 150 words is clearly no impediment – and that much of the translation would be of web pages. The surprisingly low submission of texts longer than a few words seems to suggest that online MT is being used primarily for dictionary consultation – despite the availability of many free online dictionaries – and perhaps therefore by people with some familiarity with foreign languages. Whatever ways people are using them, overall usage of online MT continues to increase exponentially (e.g. FreeTranslation from 50,000 in 1999 to 3.4 million in 2006; the totals for Babelfish are much higher).

The translation of web pages – a facility provided by PC systems before online MT services came – has complications in addition to the obvious problems of rendering the often colloquial and culture-dependent nature of the texts. Many web pages include text in graphic format, which no MT system can deal with, and therefore often much of the webpage will be untranslated. This may account for the low usage of webpage translation on online MT systems. It is thus all the more surprising that so many website developers and owners recommend users to online MT services for translation of their web pages. It is clear that they do not appreciate the poor results of any MT version, nor are they aware of consequent negative impacts on their company or products.

A recent development is systems designed for website localization. As mentioned above, localization became a specialist application of MT and translation memories in the early 1990s. The extension into website localization was an obvious move – which came, however, not until after 2000. The most significant development has been the introduction of specialised systems, notably IBM Websphere, which is designed for Internet service providers and for large corporations to supply and edit translations of their own web pages localised to their specific domain, as well for cross-language communication with customers and for providing 'gist' translations internally.

The limitations of MT when dealing with colloquial and elliptical 'normal' language – as opposed to the formal written texts of books and magazines – is highlighted by its problems with electronic mail. Just as most PC systems have provided facilities for translating web pages, many seek to embrace email text as well – with what success or user satisfaction is unknown. Few researchers have focused specifically on this type of text; they have been mainly in Japan and Korea; and even fewer have marketed such systems. An exception is Translution, which offers online translation of emails for companies. Subscriptions vary according to the level of service, and whether web-based or located on a client-server system.

Even more challenging perhaps is the language of social networking sites. Some tentative attempts have been made which highlight illustrate the similarities of such texts with spoken language and the similarities of their shared problems. But the huge possibilities of devising MT for social networking in general appear to have not yet been tackled.

4. Speech translation

As mentioned earlier, an increasing number of phrase-book systems offer voice output. This facility is also increasingly available for PC based translation software – it seems that Globalink in 1995 was the earliest – and it seems quite likely that it will be an additional feature for online MT sometime in the future. But automatic speech synthesis of text-to-text translation is not at all the same as genuine 'speech-to-speech translation', the focus of research efforts in Japan (ATR), the United States (Carnegie-Mellon University), Germany (Verbmobil project) and Italy (ITC-irst, NESPOLE) for many years since the late 1980s. The research in speech translation is beset with numerous problems, not just variability of voice input but also the nature of spoken language. By contrast with written language, spoken language is colloquial, elliptical, context-dependent, interpersonal, and primarily in the form of dialogues. MT has focused on well-formed, technical and scientific language and has tended to neglect informal modes of communication. Speech translation therefore represents a radical departure from traditional MT. Complexities of speech translation can, however, be reduced by restricting communication to relatively narrow domains – a favourite for many researchers has been business communication, booking of hotel rooms, negotiating dates of meetings, etc. From these long-term projects no commercial systems have appeared yet. There are, however, other areas of speech translation which do have working (but not yet commercial) systems. These are communication in patient-doctor and other health

consultations, communication by soldiers in military (field) operations, and communication in the tourism domain.

The potentialities of health-communication applications are obvious, particularly for communication involving immigrant and other 'minority' languages. However, there are different views of the most effective and most appropriate methods. In some cases, communication can be one-way, e.g. a 'doctor' or 'medical professional' (nurse, paramedic, pharmacist, etc.) asks the 'patient' a question, which can be answered nonverbally or by a simple "yes" or "no". In other cases, communication may be two-way or interactive, e.g. patient and doctor consulting a screen displaying possible 'health' conditions. Or communication may be via a 'phrasebook'-type system with voice input to locate phrases and spoken output of the translated phrase and/or with interactive multimodal assistance. Nearly all systems are currently somewhat inflexible and limited to specific narrow domains. Speech translation itself may be only one factor in successful health-related consultation since cultural and environmental issues are also involved; and whether medical personnel should be the initiators and 'in control' is another issue. However, before even such issues of usability and appropriateness can be resolved, the robustness of speech translation even in highly constrained domains has to be satisfactory – the weakest point is still automatic speech recognition, even though domain-specific translation itself is also still inadequate.

In the military field, the MT team at Carnegie-Mellon University developed a speech translation system (DIPLOMAT) which can be quickly adapted to new languages, i.e. languages where the US Army is deployed (Serbo-Croat, Haitian Creole, Korean). The system was based on an example-based MT approach; spoken language was matched against phrases (examples) in the database and the translations output by a speech synthesis module. An evaluation 'in the field' concluded that the speech components were satisfactory but the MT component was not adequate – translation was far too slow in practice, and a feedback ('back translation') module enabling users to check the appropriateness of the translation introduced additional errors. Further development was not pursued. However, in the same domain, another system on a hand-held PDA device has been more successful it seems. This device (Phraselator, from VoxTec, initially funded by DARPA) contains a database of phrases in the foreign language which the English-speaking user can select from a screen. Output is not synthesised speech but pre-recorded by native speakers. The device has been used by the US Army in various operations in Croatia, Iraq, Indonesia, including civilian emergency situations (e.g. the tsunami relief in 2005), by the US navy, by law enforcement officers, etc. A wide range of languages is now covered and the device and its software are now more widely available commercially.

One of the most obvious applications of speech translation is the assistance of tourists in foreign countries. Many of the organisations mentioned earlier are involved in developing systems (ATR in Japan, ITC-irst in Italy, and Carnegie-Mellon University in the USA). Many groups are utilizing the BTEC corpus of Japanese/English tourism and travel example expressions; but most have extended investigation to Chinese and English, Arabic and English and Italian and English. A welcome feature of this activity is the collaborative efforts and the exchange of resources by research groups, e.g. at the International Workshops on Spoken Language Translation since 2005. In many cases, translation is restricted to 'standard' phrases extracted from corpora of dialogues and interactions in tourist situations. However, in recent years, researchers have moved to systems capable of dealing with 'spontaneous speech', i.e. something more like real-life applications. Despite the amount of research in an apparently highly-restricted domain it is clear that commercially viable products still lie some way in the future. In the meantime, for some years yet, the market will see only the voice-output phrase-book devices and systems mentioned above.

5. Rapid development, open source, hybrid systems

One of the advantages of statistical machine translation (SMT) – the current focus of most MT research – is claimed to be the rapid production of systems in new language pairs. Researchers do not need to know the languages involved as long as they have confidence in the reliability of the corpora which they work with. This is in contrast to the slower development of rule-based MT (RBMT) systems which require careful lexical and grammatical analyses by researchers familiar with both source and target languages. Nearly all commercially available MT systems (whether for mainframe, client-server, or PC) are rule-based systems, the result of many years of development. Statistical MT has only recently appeared on the marketplace. The LanguageWeaver company, an offshoot of the research group at the University of Southern California, began marketing SMT systems in 2002. It began with Arabic-English and has now added many other language pairs. (Many

users of these systems are US government agencies involved in information gathering and analysis operations – see below.)

Increasingly, resources for statistical MT (components, algorithms, etc.) are widely available as 'open source' materials. The Apertium system from Spain has been the basis for freely-available MT systems for Spanish, Portuguese, Galician, Catalan, etc. There are other open source translation systems (less widely used), such as GPL Trans, but it is to be expected that more will be available in the coming years.

Many researchers believe that the future for MT lies in the development of hybrid systems combining the best of the statistical and rule-based approaches. In the meantime, however, until a viable framework for hybrid MT appears, experiments are being made with multi-engine systems and with adopting statistical techniques with rule-based (and example-based) systems. The multi-engine approach involves the translation of a given text by two or more different MT architectures (SMT and RBMT, for example) and the integration of outputs for the selection of the 'best' output – for which statistical techniques can be used. The idea is attractive and quality improvements have been achieved, but it is difficult to see this approach as a feasible economic method for large-scale or commercial MT. An example of appending statistical techniques to rule-based MT is the experiment (by a number of researchers in Spain, Japan, and Canada) of 'statistical post-editing'. In essence, the method involves the submission (for correction and improvement) of the output of an RBMT system to a 'language model' of the kind found in SMT systems. One advantage of the approach is that the deficiencies of RBMT for less-resourced languages may be overcome.

The languages most often in demand and available commercially are those from and to English. The most frequent pairs (for online MT services and apparently for PC systems) are English to/from Spanish and English to/from Japanese. These are followed by English to/from French, English to/from German, English to/from Italian, English to/from Chinese, English to/from Korean, and French to/from German. Other European languages such as Czech, Polish, Bulgarian, Romanian, Latvian, Lithuanian, Estonian, and Finnish are more rarely found on the market. Until the middle of the 1990s, Arabic to/from English and Arabic to/from French were also rare, but this situation has changed for obvious political reasons. Other Asian languages have also been relatively neglected: Malay, Indonesian, Thai, Vietnam and even major languages of India: Hindu, Urdu, Bengali, Punjabi, Tamil, etc. And African languages (except Arabic dialects) are virtually invisible. Many are among the world's most spoken languages. The reason is a combination of low commercial viability and lack of language resources (whether for rule-based lexicons and grammars or for statistical MT corpora).

6. Minorities, immigrants

The categorization of a language as a 'minority language' is determined geographically. In the UK, world languages such as Hindi, Punjabi and Bengali are minor, because the major language is English. In Spain, the languages Basque and Catalan are both 'minor' because the official language is Castilian Spanish. In the context of the European Union, languages such as Welsh, Irish, Estonian, Lithuanian are 'minor', whether official languages of a country or not. From a global point of view, 'minor' languages are those which are not 'commercially' or 'economically' significant. The language coverage of MT systems reflects this global perspective, and so the problems and needs of 'minority' languages were virtually ignored. Recently they have had more attention – in Spain with MT systems for Catalan, Basque, and Galician; in Eastern Europe with systems for Czech, Estonian, Latvian, Bulgarian, etc.; and in South and South East Asia with MT activity on Bengali, Tamil, Thai, Vietnamese, etc. This growing interest is reflected in the holding of workshops on minority-language MT. The problems for minority and immigrant languages are many and varied: there is often no word-processing software (indeed some languages lack scripts), no spellcheckers (sometime languages lack standard spelling conventions), no dictionaries (monolingual or bilingual), indeed a general lack of language resources (e.g. corpora of translations) and of qualified/experienced researchers. Before MT can be contemplated, these resources must be created – and the Internet may help to some extent with glossaries and bilingual corpora.

One specific target of MT for immigrants or minorities has been the translation of captions (or subtitles) for television programmes. Probably the most ambitious experiment is at the Institute for Language and Speech Processing (Athens) involving speech recognition, English text analysis and caption generation in English, Greek and French. Usually, however, captions in foreign languages are generated from caption texts produced as a normal service for the deaf or hearing impaired by television companies. A group at Simon Fraser University in Canada has investigated the translation of English television

captions into Spanish and Portuguese, and a group at the Electronics and Telecommunications Research Institute in Korea are developing CaptionEye/EK, an MT system for translation English television captions into Korean. In both cases, translation is based on pattern matching of short phrases (in systems of the example-based MT type.)

Apart from minorities and immigrants, there are other 'disadvantaged' members of society now beginning to be helped by MT-related systems. In recent years, researchers have looked at 'translating' into sign languages for the deaf. The problems go, of course, beyond those encountered with text translation. The most obvious one is that signs are made by complex combinations of face, hand and body movements which have to be notated for translation, and have to be mimicked by a computer-generated avatar. In most cases, conventional rule-based approaches are adopted, but there have also been experiments with hybrid statistical and example-based methods. Most research has focussed on translation of English text into American Sign Language and into British Sign Language, but also there are also reports involving German sign language.

7. Information retrieval, information extraction, and other applications

Translation is rarely an isolated activity; it is usually a means for accessing, acquiring and imparting information. This is clearly the case with many examples already mentioned: translation in health-related communication, translation of patents and technical documentation, translation of television subtitles, etc. MT systems are therefore often integrated with (combined or linked with) various other NLP activities: information retrieval, information extraction and analysis, question answering, summarisation, technical authoring.

Multilingual access to information in documentary sources (articles, conferences, monographs, etc.) was a major interest in the earliest years of MT, but as information retrieval (IR) became more statistics-oriented and MT became more rule-based the reciprocal relations diminished. However, since the mid 1990s with the increasing interest in statistics-based MT the relations have revived, and 'cross-language information retrieval' (CLIR) is now a vigorous area of research with strong links to MT: both fields are concerned with the retrieval words and phrases in foreign languages which match (exactly or 'fuzzily') with words and phrases of input 'texts' (queries in IR, source texts in MT), and both combine linguistic resources (dictionaries, thesauri) and statistical techniques. There are extensions of CLIR to multilingual retrieval of images and spoken 'documents', to retrieval of broadcast stories which are 'similar' to a given input English text (not just a query).

Information extraction (or 'text mining') has had similar close historical links to MT, strengthened likewise by the growing statistical orientation of MT. Many commercial and government-funded (international and national) organisations have to scrutinize foreign-language documents for information relevant to their activities (from commercial and economic to surveillance, intelligence, and espionage). The scanning (skimming) of documents received – previously an onerous human task – is now routinely performed automatically. Searching can focus on single texts or multilingual collections of texts, or range over selected databases (e.g. via syndicated feeds) or the whole Internet. The cues for relevant information include not just keywords such as 'export', 'strategic', 'attack', etc. (and their foreign language equivalents), but also the names of persons, companies and organisations. Since the spelling of personal names can differ markedly from one language to another, the systems need to incorporate 'transliteration' facilities which can convert, say, a Japanese version of a politician's name into its (perhaps original) English form. The identification of names (or 'named entities') and the problems of transliteration have become increasingly active fields in the last few years.

Information analysis and summarisation is frequently the second stage after information extraction. These activities have also, until recently, been performed by human analysts. Now at least drafts can be obtained by statistical means – methods for summarisation have been researched since the 1960s. The development of working systems that combine MT and summarisation is apparently still something for the future. The major problems are the unreliability of MT (incorrect translations, distorted syntax, etc.) and the imperfections of current summarization systems (which seek 'indicative' contents in paragraph-initial sentences, sentences containing 'important' lexical clues, sentences including specific names, etc.) Combining MT and summarization would be a desirable development in many areas – not just for information gathering by government bodies but also for managers of large corporations and for most researchers with no knowledge of the original language. Such potential users of MT rarely want to read the whole of a document; what they want is to extract information for a specific need.

The field of question-answering has been an active research area in artificial intelligence for many years. The aim is to retrieve answers in text form from databases in response to (ideally) natural-language questions. Like summarization, this is a difficult task; but the possibility of multilingual question-answering is attracting more attention in recent years.

Finally, the impetus in large corporation to produce documentation in multiple languages in as short timescales as possible has led to the closer integration of the processes of authoring (technical writing) and translating. This is true not only where companies have decided to adopt 'controlled languages' for their documentation – as we have seen above – but also where writers make use of rough translations as aids. Surveys of the use of Systran at the European Commission have shown that much of its use is by administrators and other officials when writing documents in languages they are not fully fluent in – a draft translation from a text in their own language is used as the basis for writing in another. Perhaps this is what some users of online MT and of PC systems are doing; the translation systems are aids to writing in another relatively poorly known language.

This survey has not exhausted all the applications that have been envisaged for MT; we may mention suggestions for combining MT and photocopiers, MT and document scanners, MT and cameras (e.g. for reading menus and road signs), and finally – in a reversion to MT's origins – the use of MT techniques for decipherment.

What these examples of MT applications illustrate is that MT technology is being used not just for 'pure' translation but increasingly as an aid to bilingual communication in an ever-widening range of contexts and situations, and embedded in a multiplicity of multilingual, multimodal document (text) and image (video) extraction and analysis systems. Whenever there is a need for communication and contact across languages, there will be a potential use for MT – the applications seem unending.

8. References

This survey cannot list references to all the systems and applications mentioned. The main source is the Machine Translation Archive (http://www.mt-archive.info), see the 'index of applications'. For information about commercial systems see the Compendium of Translation Software (http://www.eamt.org/soft_comp.php).