Recent developments in the use and application of machine translation

John Hutchins

Until the middle of the 1990s there were just two basic types of machine translation system. The first was the oldest and the traditional large-scale system mounted on mainframe computers in large companies. The purpose was to use MT in order to produce publishable translations. The results of MT systems were thus 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 early 190s – 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: Blench 2006), health reports (Global Health Intelligence Network: Blench 2007), patents (Japan Patent Information Office, Pan American Health Organization), police data (ProLingua), and many more. These large-scale applications of MT continue to expand and develop to the present day, and they are certain to do so into the foreseeable future.

Included in such expansion will undoubtedly be the application of MT to the localisation of products. Localization became a specialist application of MT and translation memories in the early 1990s. [For a survey see Esselink 2003]. 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 utilization of MT before the mid 1990s was software on personal computer (PC) systems. [For a survey see Hutchins 2003]. The first such systems appeared in the early 1980s soon after the marketing of PCs. They were followed by many companies marketing PCs – including most of the Japanese manufacturers of PCs – and covering an increasingly wider range of language pairs and on an increasingly wide range of operating systems. While 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 is clear that MT is sued 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 used edit to a higher quality. On the other hand, it seems more likely that users are wanting 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 sent 'raw' (unedited) MT translations as if they were as good as human translations. This must only be the case where users are 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 use systems rarely after initial enthusiasm, once they learn how poor the quality of MT output can be.

Mainframe, client-server and PC systems are overwhelmingly ‘general purpose’ systems, i.e. they are built to deal with texts in any subject domain. Of course, ‘enterprise’ systems (particularly controlled language systems) are over time focussed on particular subject areas, and adaptation to new areas is offered by most large MT systems (such as Systran). A few PC-based systems are available for texts in specific subject areas. Examples are the English/Japanese Transer systems for medical texts and patents. On the whole, however, PC systems deal with specific subjects by the availability of subject glossaries, which can be ranked in preference by users. For some systems the range of dictionaries is very wide, embracing most engineering topics, computer science, business and marketing, law, sports, cookery, music, etc.

Special devices, online MT

From the middle of the 1990s onwards, these two basic types of MT systems have been joined by a range of other types. First should be mentioned the obvious development from PC systems the numerous 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 clearly 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 users unfamiliar with the pronunciation of the phrases which may be output.

While many of these automated phrase-books and dictionaries are purchased on special-purpose devices, there are an increasing number of manufacturers of software for mobile telephones. This software is seen as obvious extension of their text facilities. Text messages can be translated and sent in other languages. The range of languages is not so far very wide, limited on the whole to the ‘commercially dominant’ languages: English, French, German, and Spanish. It can be predicted that software for mobile telephones will eventually supersede software for special-purpose devices, particularly as more of them provide direct access to online MT services.

This has been the second major change since the middle of the 1990s: the availability of free MT services on the Internet (Gaspari and Hutchins 2007). 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 (based on the Intergraph DP/Translator), initially restricted to selected forums, but which proved highly popular, and in 1994 Globalink offered an online subscription service – texts were submitted on line and translations returned by email. A similar service was provided by Systran Express. However, it was 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 next few years, the Babelfish service was joined by FreeTranslation (using the Intergraph system), Gist-in-Time, ProMT, PARS, and many others; in most cases, these were online versions of already existing PC-based (or mainframe) systems. The great attraction of these 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, although the evidence for this has not been shown, or encourage the use of the fee-based ‘valued-added’ post-editing services offered to users by some providers (e.g. FreeTranslation). 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 was something completely new – despite the availability of PC translation software. Attracted by the possibilities, many users ‘tested’ the service 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 (Somers 2007a). 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 of much of MT, where they are capable of judging it. 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; and some users must have found that online MT enabled them to read texts which they would have previously had to pass over.

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) and 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 the 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 the online MT services came – has complications in addition to the obvious problems of satisfactorily and intelligibly 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 from the low usage of webpage translation on online MT systems. However, it is all the more surprising that so many website developers and owners recommend users to online MT services for translation of their web pages (Gaspari and Somers 2007). 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. Localization became a specialist application of MT and translation memories in the early 1990s. 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. As mentioned above, localisation is a major application of MT. 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 chatroom and social networking sites. Some tentative attempts were made to deal with chatroom conversation (Condon and Miller (2002) 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 – however, perhaps all users expect everything to be in (some variant of) English...

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 primarily 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. Some of the difficulties of spoken language translation may be overcome by adding visual clues to reduce ambiguities, i.e. as multimodal systems to aid dialogue communication (e.g. Costantini et al. 2002, Burger et al. 2003). Complexities of speech translation are, however, generally 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, one-way communication, e.g. from a 'doctor' or 'medical professional' (nurse, paramedic, pharmacist, etc.) asking the 'patient' a question, which might 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 (Rayner and Bouillon, 2002), and/or with interactive multimodal assistance (Seligman and Dillinger 2006) 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: the 'patients' are likely to be regular users and could be more familiar with a language-specific device than the medical professional – and might also use it in other than health-related situations. [For a survey of possibilities see Somers 2006, 2007b]. 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 of English phrases. 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. Adaptation to medical domains is being planned.

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/English, Arabic/English and Italian/English. A welcome feature of this activity is the collaborative efforts and the exchange of resources by research groups [see IWSLT 2005, 2006, 2007]. 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 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.

Rapid development, open source, hybrid systems

As mentioned already, the rapid development of systems is becoming recognised as important for MT applications. One of the advantages of statistical machine translation – 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 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 (cf. Hutchins 1986). 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 for Dutch, French, German, Indonesian, Italian, Spanish, etc. (http://sourceforge.net/projects/gpltrans/); but it is to be expected that much 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 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/Spanish and English/Japanese. These are followed by (in no particular order) English/French, English/German, English/Italian, English/Chinese, English/Korean, and French/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/English and Arabic/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. In terms of population these are not ‘minor’ languages – 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).

Minorities, immigrants

The categorization 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 (e.g. SALTMIL at LREC in 2006) 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 [For an overview see Somers 2003]. Before MT can be contemplated, these resources must be created – and the Internet may help to some extent with glossaries and bilingual corpora. There is in addition, the question whether the communication needs of immigrants and minorities are best met with MT or with lower-level technologies, as indicated above with reference to spoken language translation.

One specific target of MT for immigrants or minorities has been the translation of captions (or subtitles) for television programmes. 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 (Piperidis et al. 2004). 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 (Turcato et al 2000), 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 (Seo et al. 2001). 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 Morrissey et al. (2007) have experimented with hybrid statistical and example-based methods. Experiments have mainly been from English text into American Sign Language (Huenerfauth 2007) and British Sign Language (Marshall and Sáfár 2003, Morrissey et al. 2007), but also there is also a report of one from sign language into English (Stein et al. 2007), where there are references to systems for other languages (Chinese and Spanish). We may expect more in future.

Information retrieval and extraction

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 early 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 share in retrieving 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 images and to spoken ‘documents’, e.g. the experiments by Flank (2000) and by Etzioni et al. (2007) on multilingual image retrieval, and by Meng et al. (2001) for retrieving Chinese broadcast stories which are ‘similar’ to a given input English text (not just a query).

Information extraction has similar close links to MT, strengthened likewise by the growing statistical orientation of MT. Many 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. 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 their transliteration has become an increasingly active field in the last few years. (For a summary of the issues see Condon and Miller 2006)

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 (Saggion 2006, Siddharthan 2005). The major problems are the unreliability of MT (incorrect translations, distorted syntax, etc.) and the imperfections of current summarization systems which are based on the detection of sentences important as indicators of content (paragraph-initial sentences, sentences containing lexical clues, particular 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 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 summarisation, this is a difficult task; but the possibility of multilingual question-answering is attracting more attention in recent years (see the proceedings of the *Workshop on Multilingual Question Answering (MLQA06)*, held in 2006 as part of the EACL conference in Trento, Italy.)

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 (Senez 1995). 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 – it may explain to some extent the frequency of translation of short phrases.

What these examples of MT applications illustrate is that MT is being used not for ‘pure’ translation but to aid bilingual communication in an ever-widening range of situations; and MT is becoming just one component of multilingual, multimodal document (text) and image (video) extraction and analysis systems. The future scope of MT and its applications seems to be without limit.

============

References[1]

[1] Most items are available through the Machine Translation Archive (http://www.mt-archive.info)

Blench, Michael (2006): 'Automated translation system Job Bank.' In: *AMTA 2006: 7th Conference of the Association for Machine Translation in the Americas, "Visions for the Future of Machine Translation"*, August 8-12, 2006, Cambridge, Massachusetts, USA.

Blench, Michael (2007): 'Global Public Health Intelligence Network (GPHIN).' In: *MT Summit XI,* 10-14 September 2007, Copenhagen, Denmark. *Proceedings*; pp.45-49.

Burger, Susanne, Erica Costantini, and Fabio Pianesi (2003): **'**Communicative strategies and patterns of multimodal integration in a speech-to-speech translation system.' In: *MT Summit IX*, New Orleans, USA, 23-27 September 2003; pp.32-39.

Condon, Sherri, and Keith Miller (2002): 'Sharing problems and solutions for machine translation of spoken and written interaction.' In: *ACL 2002: Workshop on Speech-to-Speech Translation*, 11 July 2002, Philadelphia, p.93-100.

Condon, Sherri and Keith J.Miller (2006): 'Name transliteration: current methods, applications, and evaluation in name transliteration and name translation.' Tutorial at *AMTA 2006 conference*, August 8, 2006, Cambridge, Massachusetts, USA; 46pp.

Esselink, Bert (2003): 'Localisation and translation.' In: Harold Somers (ed.) *Computers and translation: a translator's guide* (Amsterdam: John Benjamins, 2003); pp.67-86.

Etzioni, Oren, Kobi Reiter, Stephen Soderland, and Marcus Sammer (2007): 'Lexical translation with application to image searching on the web.' In: *MT Summit XI*, 10-14 September 2007, Copenhagen, Denmark. *Proceedings*; pp.175-182.

Gaspari, Federico and John Hutchins (2007): 'Online and free! Ten years of online machine translation: origins, developments, current use and future prospects.' In: *MT Summit XI*, 10-14 September 2007, Copenhagen, Denmark. *Proceedings*; pp.199-206.

Gaspari, Federico and Harold Somers (2007): 'Using free online MT in multilingual websites.' Tutorial at *MT Summit XI,* 10 September 2007, Copenhagen, Denmark.

Harrison, Ann (2005): 'Machines not lost in translation.' *Wired News*, March 9.

Huenerfauth, Matt (2005): 'American Sign Language generation: multimodal NLG with multiple linguistic channels.' In: *ACL-2005*: *Student Research Workshop,* University of Michigan, Ann Arbor, June 2005; pp. 37-42.

Hutchins, John (2003): 'Commercial systems: the state of the art'. In: Harold Somers (ed.) *Computers and translation: a translator's guide* (Amsterdam; John Benjaomins, 2003); pp.161-174.

IWSLT (2005, 2006, 2007): *International Workshops on Spoken Language Translation*: 2nd, 24-25 October 2005, Pittsburgh, USA; 3rd, 27-28 November 2006, Kyoto, Japan; 4th, 15-16 October, Trento, Italy

Marshall, Ian and Éva Sáfár (2003): 'A prototype text to British Sign Language (BSL) translation system.' In: *ACL-2003*: 41st Annual meeting of the Association for Computational Linguistics, July 7-12, 2003, Sapporo, Japan. 4pp.

Meng, Helen, Berlin Chen, Sanjeev Khudanpur, Gina-Anne Levow, Wai-Kit Lo, Douglas Oard, Patrick Schone, Karen Tang, Hsin-Min Wang, and Jianqiang Wang (2001): 'Mandarin-English information (MEI): investigating translingual speech retrieval.' In: *HLT-2001: Proceedings of the First International Conference on Human Language Technology Research,* San Diego, CA, March 18-21, 2001; 7pp.

Morrissey, Sara, Andy Way, Daniel Stein, Jan Bungeroth, and Hermann Ney (2007): 'Combining data-driven MT systems for improved sign language translation.' In: *MT Summit XI,* 10-14 September 2007, Copenhagen, Denmark. *Proceedings*; pp.329-336.

Piperidis, Stelios, Iason Demiros, Prokopis Prokopidis, Peter Vanroose, Anja Hoethker, Walter Daelemans, Elsa Sklavounou, Manos Konstantinou, and Yannis Karavidis: 'Multimodal multilingual resources in the subtitling process.' In: *LREC-2004: Fourth International Conference on Language Resources and Evaluation*, Proceedings, Lisbon, Portugal, 26-28 May 2004; pp.205-208.

Rayner, Manny and Pierrette Bouillon (2002): 'A flexible speech to speech phrasebook translator'. In: *ACL-2002 workshop "Speech-to-speech translation"*,11 July 2002, Philadelphia, USA; pp. 69-76.

Saggion, Horacio (2006): 'Multilingual multidocument summarization tools and evaluation.' In: *LREC-2006: Fifth International Conference on Language Resources and Evaluation*. Proceedings, Genoa, Italy, 22-28 May 2006; pp.1312-1317.

SALTMIL (2006): LREC-2006, *5th SALTMIL Workshop on Minority Languages: "Strategies for developing machine translation for minority languages"*, 23 May 2006.

Seligman, Mark and Mike Dillinger (2006): ‘Usability issues in an interactive speech-to-speech translation system for healthcare’. In: *HLT-NAACL 2006: Proceedings of the Workshop on Medical Speech Translation*, 9 June 2006, New York, NY, USA; pp.1-8.

Senez, Dorothy (1995): ‘The use of machine translation in the Commission.’ In: *MT Summit V* Proceedings, Luxembourg, July 10-13, 1995; 11pp.

Seo, Young-Ae, Yoon-Hyung Roh, Ki-Young Lee, and Sang-Kyu Park (2001): ‘CaptionEye/EK: English-to-Korean caption translation system using the sentence pattern.’ In: *MT Summit VIII: Machine Translation in the Information Age*, Proceedings, Santiago de Compostela, Spain, 18-22 September 2001; pp.325-329.

Siddharthan, Advaith and Kathleen McKeown (2005): ‘Improving multilingual summarization: using redundancy in the input to correct MT errors.’ In: *HLT-EMNLP-2005: Proceedings of Human Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, October 2005; pp. 33-40.

Somers, Harold (2003): Translation technologies and minority languages. In Somers, Harold (ed) *Computers and translation: a translator’s guide* (Amsterdam: John Benjamins), p.87-103.

Somers, Harold (2006): ‘Language engineering and the pathway to healthcare: a user-oriented view.’ In: *HLT-NAACL 2006: Proceedings of the Workshop on Medical Speech Translation*, 9 June 2006, New York, NY, USA; pp.32-39.

Somers, Harold (2007a): ‘Machine translation and the World Wide Web.’ In: Khurshid Ahmad, Christopher Brewster, and Mark Stevenson (eds.) *Words and intelligence II: essays in honor of Yorick Wilks* (Dordrecht: Springer); pp.209-233.

Somers, Harold (2007b): ‘Theoretical and methodological issues regarding the use of language technologies for patients with limited English proficiency.’ In: *TMI-2007: Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, Skövde [Sweden], 7-9 September 2007; pp.206-213.

Stein, Daniel, Philippe Dreuw, Hermann Ney, Sara Morrissey, and Andy Way (2007): ‘Hand in hand: automatic sign language to English translation.’ In: *TMI-2007: Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, Skövde [Sweden], 7-9 September 2007; pp.214-220.

Turcato, Davide, Fred Popowich, Paul McFetridge, Devlan Nicholson, and Janine Toole (2000): ‘Pre-processing closed captions for machine translation’ In: *ANLP/NAACL 2000 workshop: Embedded machine translation systems*, May 4, 2000, Seattle, Washington, USA; pp. 38-45.