Paper presented at Translating and the Computer 18 (London, 14-15 November 1996)

RECENT TRENDS IN MACHINE TRANSLATION

Muriel Vasconcellos
International Association for Machine Translation [1]
and
L. Chris Miller
Multilingual Computing Solutions, Inc. [2]

Abstract

In the last two years machine translation (MT) has embarked on a voyage into the future, spurred by the presence of personal computers on individual desktops throughout the world and, more recently, universal access to electronic text on-line. This impressive growth has led to many new trends, including major changes in the profile of the user.

Apace with this trajectory has come better communication and increased collaboration between all the groups concerned--MT researchers, developers, users, and watchers. The International Association for Machine Translation (IAMT), together with its three regional associations created in 1991, has fostered this convergence by creating opportunities--workshops, conferences, publications--through which to share the latest information in this dynamically growing field.

1 New Dimensions in MT Service Delivery

1.1 Overview

The dreams of yesteryear's visionaries are finally coming true. Machine translation (MT) has launched on an unparalleled surge of growth--a historic shift in the way it is being used and a phenomenal increase in the number of people who rely on it. We now have MT software that is viable, affordable, and runs on virtually any 1990s desktop. Today there are more than 500 vendors of MT software for the personal computer around the world, and among them they put out well over 1,000 products. [3] One of the vendors, Globalink, sells its extensive line of software in at least 6,000 stores in North America alone, and at present Europe is its fastest-growing market.

The ubiquity of the desktop computer with access to the Internet has given momentum to an unprecedented growth in MT usership. We now have MT on-line, accessible through e-mail, client-server arrangements, Internet service providers, and a growing number of other sites on the Internet. The on-line phenomenon is changing our whole way of thinking about machine translation. Together, these two developments--the abundance of low-cost MT in shrink-wrapped boxes, coupled with MT on-line--are turning machine translation into an everyday commodity that is within the reach of virtually anyone with a late-model personal computer.

The sudden shift in MT use and the dramatic increase in its usership have also brought a sea change in the profile of the user. Because of its widespread availability, MT has been forced, appropriately or not, to graduate from the days when a system's caretakers had to nurture it constantly in order for it to perform acceptably. It now stands on its own, and, by and large, its new users must fend for themselves, whether by

customizing the system and/or learning how to postedit, or accepting the output as it is.

While all these changes are taking place in MT use, other exciting trends are also redrawing the entire map of the field of machine translation itself. Many languages, especially the more challenging ones, are being tackled and added to the vendors' repertoires. In the new game of "plug-and-play," MT engines are now being made interfaceable with a variety of other software functions. Speech translation is making steady progress. Off-the-shelf tools are speeding up research and development. Creative partnerships are being forged between and within the commercial and academic communities. Systems of different philosophies are being joined together. Indeed, on all fronts MT research is accelerating its ongoing march toward distant horizons. It's safe to say that never in the history of this field has so much happened within such a short period.

1.2 The On-line Phenomenon

We have all witnessed the explosive expansion of the World Wide Web, the Internet service providers, and, most recently, the intranets. Not many of us, however, are aware of the extent to which machine translation is being swept along in this tide. Already on-line access is causing MT use to grow at an unprecedented rate.

As of September 1996, low-cost machine translation in one form or another was available at some 30 on-line sites in cyberspace.[4] It comes in a variety of forms and modalities.

(

"Raw" (unpostedited) MT can already be accessed in the following ways:

through software installed on the individual PC that machine- translates pages from a Web site;
from a vendor directly (Systran, Logos, etc.);
from a client-server arrangement set up for the purpose (for example, the U.S. intelligence community's Open Source Information System and, coming soon, Intelink-SCI, which serves an estimated 100,000 visitors [Bostad 1996]);
from an Internet service provider (currently CompuServe's Document Translation Service);
from several commercial translation services.

MT vendors are also currently gearing up for intranets.

)

Postedited output, in turn, can be obtained from a growing number of commercial translation bureaus as well as CompuServe's Document Translation Service--and soon, undoubtedly, from other Internet service providers.

(

On-line purchase is yet another way to go. As with many other kinds of software, the vendors make it easy to order an MT package on-line. In fact, we predict that within a few years the shrink-wrapped box will have yielded almost entirely to on-line sale/purchase arrangements.

CompuServe has been a pioneer in providing MT on-line. It currently offers three types of fully automatic service to its members and crunches some 32 million words (128,000 pages) annually (Flanagan 1996). Its first MT service was inaugurated on the MacCIM Support Forum two years ago. Six months later the experimental World Community Forum began machine-translating conversational messages between people around the world. Today the Forum translates into four languages simultaneously for an enrolled membership of more than 75,000 users. The most recent addition to the provider's array of MT capabilities is the CompuServe Document Translation Service (CDTS), which turns out "raw" and postedited translations of larger documents. Each month the demand for MT at these three sites has increased. CompuServe will

soon be offering machine translation as a standard option for e-mail and also for on-line chat (ibid.).

This growing use of MT on-line cannot be dismissed as casual curiosity. Unlike software purchased off the shelf, for which no direct measurements are possible, on-line access is documented automatically, and therefore patterns can be discerned. For example, the records for CompuServe's production translation service show a number of repeat large-volume users. The statistics (ibid.) reveal that about 85% of the requests are for raw MT--a much larger percentage than had been anticipated. What could not be determined automatically was whether the raw translation was being used for gisting purposes only or whether it was being postedited for further use. To discover more about its subsequent fate, Flanagan conducted a market survey which revealed that the CDTS is used mostly for business and technical purposes where assimilation-quality MT is sufficient (ibid.). The bottom line is that the customer is willing to pay for this service.

In the World Community Forum, although there is no direct evidence of the extent to which the machine translations are being relied on, at least one fact can be reported: the Forum's sysop is inundated with complaints on the rare occasions when the MT system goes down.[5]

So far, CompuServe uses MT for only three of its more than 3,000 on-line services. Expansion to these other services creates a market that is mind-boggling. With e-mail alone, which has a current volume of more than 40,000,000 messages a month, the potential for growth is incalculable. If we then venture outward to the larger space of the World Wide Web, the downloading of Web pages raises a prospect that is literally beyond imagination. No one knows how large the Web really is, but it was recently reported that LYCOS, a system that collects pages from the Web and indexes them, has a database containing 166,000,000 of the most "popular" pages, which is evidently the tip of the iceberg.

The volume of text on the Internet is reportedly 10 times larger this year that it was in 1995. Matching this explosion in overall size is a tandem increase in languages other than English. Non-English text represented 90% of total volume in 1996, and, despite the 10-fold overall growth, still reached a level of 80% this year. In other words, the non-English segment was actually 20 times larger. At the same time, studies have shown that such readers prefer to have a text in their own language, regardless of how awkward and flawed, than attempt to understand English.

In these circumstances, only a fully automatic process capable of handling very large volumes of text with near-real-time turnaround can provide the translation capacity required by on-line markets. Flanagan (ibid.) also points out that the on-line culture favors rapid and shallow assimilation of information. For these reasons, MT is an ideal fit.

(

2 The New User/Consumer Profile

Now that we have seen the new trends in MT from the point of view of the general public, we should look at the perspective of the user and the end consumer.

Acceptance. Although Flanagan and a few other forward-looking pundits (Church and Hovy 1993) predicted that "crummy" machine translation would find its true niche in cyberspace, most MT-watchers have been skeptical. Computer store MT packages already have a return rate of 20%--the highest in the software industry. Bearing in mind its low acceptance even for straightforward texts, the skeptics felt certain that electronic messages were too slangy, unstructured, and fraught with spelling and grammatical errors to ever catch on as an effective application for machine translation. Users would be dismayed by the results, and

with little positive reinforcement they would quickly abandon the translation option. However, Flanagan's study of reactions to MT on CompuServe showed that only 25% of the users have dropped out. Those who have stayed in for the long term have usually gone through a series of phases as they continued to use it: amazement, dismay, reconsideration, and, finally, pragmatism (Flanagan 1996).

Purpose of translation--a new typology. Traditionally MT usage has been classified according to its purpose. It is considered to be either translation for *dissemination*, or translation for information purposes only, also known as "gisting" or *assimilation.* At this point we would like to add a third category and at least two types of each. Type 1 represents the more direct use of MT in one of its natural niches, while type 2 is a further development that requires greater human intervention at some point in the process.

Dissemination--type 1 Domain-specific, one-to-many languages, publication quality, simple texts--mainly for localization.

Users: Corporations, for technical documentation.
Current status: Productivity improvements of up to 100% (average 35%-40%); many satisfied users.
Problems: Controlled input required, "hands-free" impossible; although productivity could be improved with better quality, this niche is rapidly being filled with low-cost systems, making linguistic investment less easy to justify.
What it takes: Customizability, filters, translation memory; other aids for translator/posteditor, finely tuned lexicon, easy to update.
Outlook: Steady growth, on-line access, more languages, improved quality.

Dissemination--type 2 General-purpose, one-to-one or few languages, publication quality, simple texts, wide range of complexity in texts.

Users: Translation agencies, in-house translation services, individuals; market difficult to specify, potentially very large.
Current status: Not growing as rapidly as other uses; willing translators show productivity of up to 65% (average 30%).
Problems: Heavy postediting, judgment calls are time-consuming, domain drift, hence need for improved quality; few systems perform well in this arena; linguistic development investment difficult to target.
What it takes: postediting aids; very large and sensitively coded lexicon(s), easy to update (better a combined dictionary than "topical glossaries"); parser and rule base; filters and translation memory also helpful.
Outlook: Slow but steady growth; on-line access will greatly expand use.

Assimilation--type 1 "Raw" MT for gisting, sometimes automated postediting; broad range of subjects.

Users: Large institutions serving individual users, serious *or* casual--e.g., governments, libraries; on-line services for databases, Web pages, e-mail.
Current status: Long-standing applications in governments; large on-line market poised to explode.
Problems: Quality tends to be poor.
What it takes: Very large and judiciously coded lexicon(s), easy to update (better a combined dictionary than "topical glossaries"); parser and rule base.
Outlook: MT will grow exponentially in this area, volume incalculable; opportunity to improve quality to gain larger market share; need for more language combinations; trend toward "plug-and-play."

Assimilation--type 2 Lightly postedited raw MT, "MT+" for "gisting-plus."

Users: Same as type 1 but user requires better quality.
Current status: Underutilized.
Problems: Lack of public awareness of this option; shortage of suitable posteditors, translators often not able to relax standards.
What it takes: Good quality; large and richly coded dictionaries.
Outlook: Potentially large area of growth; opportunity to develop automated postediting; better quality; more languages; "plug-and-play."

Conversation--type 1 Real-time, fully automatic translation of written natural dialogue. From the linguistic standpoint, the process combines the characteristics of dissemination and assimilation. *Dissemination* is the output of the writer/speaker, while *assimilation* assumes the perspective of the reader/listener, and *conversation* combines the two. Its main characteristic is that the interlocutors provide feedback to one another as they structure the information that they share.

Users: None as yet; it would be used for on-line chat, TV captioning, etc.
Current status: The category *conversation* is new.[6] There are few existing examples of machine-translated conversation outside the laboratory.
Problems: Because of the feedback factor, less-than-perfect expression is often sufficient to make the communication work (human beings tend to seek the least effort that is sufficient). The input is "flawed" with incomplete sentences and many other discourse-related characteristics. Solutions to these linguistic challenges will help in the translation of speech.
What it takes: Additional research on discourse, adaptation of dictionaries and rule bases to cover 1st person, nonstandard forms, incomplete sentences (fragments, false starts, etc.); capture of interlocutor feedback.
Outlook: Pressure of demand will lead to premature use; acceptance will vary; creates opportunity for research leading to across-the-board linguistic improvements,

especially for speech translation.

Conversation--type 2 Speech translation, with added problems of speech recognition, "noise," etc.

What it takes and *outlook*: Beyond the scope of the present report.

Participation of professional translators. Until recently a number of leading translation professionals have been vocal in their opposition to machine translation. For example, while some members of the American Translators Association were open to learning about MT and experimenting with it, most of them either lurked on the sidelines or took an active stance against MT.
)
On the other hand, since the late 1970s a few professional translators have worked as MT posteditors at such places as Wright-Patterson Air Force Base, Environment Canada, the Pan American Health Organization, the European Commission, Xerox Corporation, and several translation agencies--for example, Antler and LexiTech, to mention some well-known case histories. It is also true that some of the corporate MT operations have employed posteditors with no experience in translation.

The last two years have seen an impressive upsurge of interest in MT on the part of professional translators. Many are asking for information on MT, and Flanagan reports (1996) that she has received "hundreds" of resumes from translators seeking employment with CompuServe's Document Translation Services. As we suggested earlier, of the 85% raw MT provided by the CDTS, it is possible that some is being postedited.

Corporate/institutional vs. individual user. Major changes in the user profile are also seen in the shift from corporate or institutional users to individuals. As long as MT programs ran on mainframes, minicomputers, or sophisticated workstations like the Sun, the high cost of hardware and software, not to mention overhead and support personnel, limited MT to the corporate or institutional setting. Even today, in fact, some MT vendors still concentrate on the corporate user. This was corporate-user thinking that led the authors of the 1991 Ovum report (Engelien and McBryde 1991) to forecast that by the year 2000 some 400 MT units would be being sold annually in Europe and the United States, each at a price of about US$150,000. As it turned out, the phenomenal growth in the PC market over the subsequent two years proved their numbers to be off by a factor of *250* in terms of units sold, and by a factor of as much as *2,500* in terms of price (Vasconcellos 1993).

The corporate user calculates the costs involved and monitors productivity carefully--often with highly detailed statistics. The output is usually postedited because more often than not the MT system has been enlisted to support or produce a marketable product, because otherwise the high cost is difficult to justify. Translating "raw" information has not proven to be commercially profitable and has therefore had to be subsidized by the public sector. Postediting is the most expensive of all the factors of production, and therefore the one that has been studied the most. In addition, months are usually spent on customizing the dictionaries in order to improve translation quality and thereby reduce the amount of postediting required. Corporate and institutional users also make an effort to demonstrate savings over alternative approaches, usually human translation. Case histories are studied to learn from the experiences of others.

The first commercial MT product to run on a microcomputer, or *personal computer (PC)* as we know it today, was MicroCAT, introduced by Weidner in 1983. While the software was still quite expensive, nevertheless it opened up MT to a whole new world of users. Now, of course, there is an infinite variety of off-the-shelf PC packages selling at bargain-basement prices--many of them for less than US$100 per unit.

Since the hardware is usually already in place and the software is eminently affordable, there is no longer any great need to monitor costs or productivity.

Also, many more types of applications are possible, attracting a much broader-based user population. For a number of applications the user does not need much knowledge of the languages in question. In fact, MT is often used to determine whether or not a text is worth translating, or to scan a text for keywords in order to decide on its further disposition. The latter has been particularly successful in law offices. Whereas for the corporate user the translation of retrieved information has been expensive and difficult to justify, for the individual consumer MT is just another low-cost software package. There are many reports of PCMT meeting users' needs in both business and the home--in the first case for repetitive business correspondence, documents, books, patents, technical manuals, software, etc.; in the latter, for personal letters, recipes, tourist travel, homework, and a much bigger "etc.," since the possibilities are far from having been fully explored. And now with MT on-line the possibilities are beyond imagination: technical support, forum messages, chat, e-mail, material downloaded from databases--the sky's the limit.

Mass customization. The "in" concept these days is to offer mass-produced products that are at the same time customized or customizable to suit the needs of individual users--a trend that has been hastened by the lower cost of producing software on CDs. An example is Globalink's newest engine, Telegraph, which bundles a number of language pairs in one package and also plugs-and-plays with a variety of other products. As corporate MT evolved over the years, it became reasonably flexible and adaptable to users' needs. However, such tailoring was costly and often required frequent contact with the vendor. Now, however, the products come with a much wider range of options; the dictionaries are more easily updated; topical glossaries are available for a large range of subject areas; some products, such as Globalink's Telegraph, allow the user to update linguistic rules; and, as we saw above, the translation engine can be plug-and-played with other software such as pre- and postprocessors, grammar and style checkers, translation memory, dictionaries and thesauri on CD-ROM, on-line databases, etc. The options now available allow for highly individualized customization.

(

3 International Association for Machine Translation

The growth in MT usage and the changes in the user profile have been paralleled by improved communication and increased collaboration between all concerned. The International Association for Machine Translation (IAMT), together with its three regional associations, has fostered this convergence by creating opportunities through which to share the latest information in the dynamic field of machine translation.

The idea behind the IAMT experiment, launched in 1991 at MT Summit III in Washington, D.C., was to open up regular communication between all who are interested in machine translation, from any perspective, throughout the world. Five years later, its principles have become fully embodied in its work.

IAMT is for everyone. IAMT and its three regional associations bring together users, developers, researchers, sponsors, and any other person or institution with an interest in machine translation. The fundamental aim is to enhance communication by sharing information--experience, knowledge, interests, concerns.

IAMT is worldwide. The work of IAMT is carried out through its three regional associations--the Asia-Pacific, American, and European associations for machine translation (AAMT, AMTA, and EAMT,

respectively)--which together cover all the continents of the world. The tripartite structure allows for different approaches and emphasis in keeping with local circumstances, while the international umbrella ensures that the advantages of membership are shared by all. Individuals, institutions, and corporations are free to join any of the three associations.

IAMT conferences and publications. In addition to organizing the biennial Machine Translation Summit, IAMT, through its regional associations, offers a number of workshops, conferences, and publications. Its growing shelf of publications includes the worldwide newsmagazine, *MT News International*; proceedings of various meetings; the *MT Yellow Book*, a locator directory published by AMTA; and tabulated information on MT systems, available both in hardcopy form and on-line.

This exchange of information, ideas, and perspectives has contributed to the new MT paradigm and will undoubtedly help to rationalize the unbridled growth that lies ahead.

NOTES

1. San Diego-based consultant on translation and machine translation; e-mail: 71024.123@compuserve.com; also president, International Association for Machine Translation, 655 Fifteenth Street, N.W., Suite 310, Washington, D.C. 20005 USA, tel/fax: +1-703-716-0912, e-mail: AMTAinfo@aol.com.

2. Washington, D.C.-based consultant on machine translation; e-mail: 70303.314@compuserve.com; also founding partner of Multilingual Computing Solutions, Inc., 1736 Kenyon Street, N.W. Washington, D.C. 20010, USA.

3. Except for quotations from Flanagan, the data in this paper on current MT products and their use has been provided by L. Chris Miller and Edith Westfall, Multilingual Computing Solutions, Washington, D.C. Their database lists over 1,800 commercial MT products, the criterion for a "product" being that it is boxed and/or priced separately. Separate boxing by a single company may correspond to different operating systems or different language combinations (pairs). Lately some vendors have introduced several language combinations on a single CD, which is counted as one product.

4. See Westfall (1996). While written a year earlier and subsequently modified by the magazine's editorial staff, this article gives an idea of the extent of the on-line phenomenon.

5. Personal experience of L. Chris Miller, who also serves as one of the sysops of the World Community Forum.

6. Proposed by E. Hovy at AMTA-96 (Montreal, October 1996).

REFERENCES

Bostad, Dale A. 1987. "Machine Translation for General Purposes," supplement to *Expanding MT Horizons*, proceedings of 2nd conference of the Association for Machine Translation in the Americas (Montreal, 2-5 Oct 1996), p. 5.

Church, K.W. and E.H. Hovy. 1993. Good Applications for Crummy Machine Translation. *Machine Translation* 8:239-258.

Engelien, B., and R. McBryde. 1991. *Natural Language Markets: Commercial Strategies*. London: OVUM, Ltd., 1991.

Flanagan, Mary. 1996. "Two Years Online: Experiences, Challenges, and Trends," *Expanding MT Horizons*, proceedings of 2nd conference of the Association for Machine Translation in the Americas (Montreal, 2-5 Oct 1996), pp. 192-197.

Vasconcellos, Muriel. 1993. "The Present State of Machine Translation Usage Technology, or: How Do I Use Thee? Let Me Count the Ways." In *MT Summit IV: Proceedings* (Kobe, 20-22 July 1993), pp. 35-45.

Westfall, Edith. "Translation Services Go Online!," *Multilingual Communications & Computing*, vol. 7, no. 2, pp. 15-19.