Machine Translation - The Road Ahead

Walter Hartmann
MTConsulting Co.
37 Rollingwood Drive
Pittsford, NY 14534
Tel: (716)586-6150
Fax: (716)218-9538
Email: wh@mtconsult.com

The application of MT on the Internet has certainly attracted much attention in recent years, and many observers see its future mostly in this arena of real-time raw translation. However, the need for high-volume, fast turn-around translation of publication quality has not abated. This paper will take stock of that particular use of MT and venture predictions as to its future.

Machine Translation - the Road Ahead

Given the amount of renewed public attention bestowed upon Machine Translation (MT) in the last few years, it may seem to the outsider that MT has received a refreshed lease on life, compliments of the Internet. From real-time MT in chat rooms to sifting through huge amounts of data to determine which texts need to be translated manually to the almost instant translation of web sites, machine translation has become, once again, respectable. To those of us, however, who have been working with MT in specialized applications for some time, this new interest has not had much effect.

In my discussion of MT, I hope to avoid too much duplication with the other presentations and will limit my remarks to my own experiences as a user of translation technology. I will give a brief overview of how my use of MT has changed over the years and I will conclude my observations with some estimations as to the future of MT in my areas of expertise.

About 15 years ago, I ran a small translation company in Rochester, NY. One day we were approached by the translation department of Xerox Corp., also a local company. They were looking for translators with experience in machine translation and/or post-editing. Although I had neither, I was intrigued and expressed my interest in becoming involved in these aspects of translation technology. Xerox provided my company with some training and support, and that was the beginning of a long and enduring professional relationship that continues to this day.

In these early days of my MT involvement, Xerox still relied upon a rarified environment to provide machine translation output of high quality:

- Xerox publications were authored by Xerox employees who were technically versed professional writers, well-immersed in their subjects.
- In addition to good writing, Xerox had, as far back as the early to mid-1970s, instituted controlled language (MCE = Multinational Customized English) with a seamless system of writing rules, authoring tools and compliance checking.[1]
- The domain in which MT was used was very clearly defined; analog copiers. The terminology, though highly specialized, was relatively limited and unambiguous.
- In a stroke of genius, the team implementing the first MT system at Xerox had decided to limit the general dictionary to the bare minimum, focussing its efforts instead on developing a highly specialized multi-target dictionary for the technical terminology.[2]

Together, these elements resulted in high-quality raw MT output. And, to be honest, the output I remember from these “good old days" remains unmatched in many cases today.

For a while, post-editing the Xerox output was not a frustrating task, many of the translators I introduced to the task took to it very quickly and easily. One could say that we had a very fortunate first experience with MT.

Of course this did not last forever, the ideal environment was marred by the following factors:

- New products were introduced with terminology that introduced lexical ambiguities;
- Author training budgets were slashed;
- Untrained authors (outsourced) began to contribute to the documentation;
- Increasing lack of terminology coding, and
- A stagnant MT system.

As a result of these factors, MT output quality suffered tremendously, especially since the actual translation software did not experience any significant improvements over the time.[3]

While witnessing and learning from the effects of reduced quality in source writing and terminology management on Xerox MT, we became involved in post-editing for other clients' MT outputs. We were then drawn into coding terminology and developing terminology lists for clients. We experienced first-hand how the quality of the source text and that of the specialized lexicons directly affects the output of any MT system.

Finally, when MT became available on personal computers we were free to develop our own MT dictionaries and processes. Using off-the-shelf software, we began to map out and implement MT workflows for clients' translation needs. These clients came from industry with high documentation demand and tight deadlines such as automotive or process control clients.

In general, we were faced with a situation much different from those in which we had seen MT used. Instead of the "luxury" of maintaining and refining just one lexicon for one client in one domain, we found ourselves working for customers from various domains with different terminology needs. Naturally, our processes would differ from those used at corporations with in-house translation department and/or MT support. In order to be useful in this environment, MT must be applicable, with decent output, to a multitude of domains and publication styles.[4] It must be able to process input from a multitude of file formats, and its output must neatly fit into the same formats.

For several clients, we managed to employ MT in an efficient way: Clients with large volumes of similar source files, such as user guides and/or service manuals for related products, e.g. different car lines or software of similar nature. Our focus then turned to pre-editing, as we did not have any influence on the source text quality, and, clearly, most publications we encountered were not written with translatability in mind. At the same time, we began an integration of translation memory into the production process.

We managed to fashion a process with the following steps:

- taking the source materials from the client,
- put them through various pre-editing and de-formatting filters,

- running them through translation memory and machine translation tools and
- filtering them back into their original formats.

For an outsourced activity, our translation process was indeed an efficient means to fast turn-around at reasonable cost to the clients. On average, we managed to reduce turn-around time by about 35% and production cost by approximately 45-50% for translation from English into French and German.

This environment of high-volume translation with short turn-around requirements is the one for which MT is ideal: MT programs have a fast throughput, can work 24 hours a day and are consistent. With the right lexicon and some tweaking of the rules where possible (the L&H Barcelona engine, e.g., permits such tweaking), the results of MT can be quite respectable, allowing for a speedy post-editing. Especially when the publisher's expectations towards stylistic finesse are low, MT is in its element.[5] But this is also the area where, to me, MT does not seem to have much of a future, as I will explain below.

Over the last several years, I have become aware of an important development with ramifications, I believe, for the future of MT in the areas of my expertise. Certain companies began to build publication databases in which authored texts are stored in the smallest-feasible segments (much like in translation memory). The General Motors method presented before at AMTA meetings is fashioned in this way.[6] Following the principle that all source text should have to be translated only once, these databases will, one day, ensure that all source text should have to be authored only once.

And, if this approach is successful and adopted in many companies, one day there may be no new text at all, because in many fields there is, in my view, a finite number of sentences that are needed to describe all procedures and actions to take when working, e.g. on a car, a turbine, a locomotive etc. The relational databases in which they are stored will ensure that for each segment of the source language there is an equivalent in the target language(s). And in due course, they will become self -sufficient, multilingual document production systems without need for MT or, in fact, translation at all.

One good example for this type of development is the automotive industry. Year in and year out, they have to publish large volumes of service manuals for each and every model they manufacture. For example, one North American automaker sells 17 different models in Canada. Since the service manuals must be made available in Quebec in French, this means that 17 service manuals of about 2,500 pages each need to be translated, from one model year to the next. However, unless new technology is introduced, these manuals contain little new material from one issue to the next making them, after the initial translation, good candidates for translation memory.

TM is often fooled, however, by variations in writing, even when the actual content does not change. Once these variations are eliminated by using the source authoring database, and once the segments are linked to their respective translation segments, there will be little use for traditional TM products, let alone MT on a continuous basis. In

fact, a manager of a large translation company predicted recently that within a few years, European automotive translation into the dominant languages will no longer exist in any meaningful volume.[7] I tend to agree with this outlook.

As support for this reasoning, let me cite the following example:

For one automotive client, we translated a service manual for one specific passenger car from English into Canadian French. This was done completely using our MT process, as there was no existing legacy data for this language pair. Once translation was complete, source and target texts were aligned and integrated into a translation memory.

When the manual for a different model needed to be translated, we analyzed the new source text against that translation memory. Over 40% of the new text completely matched text within our TM database. Thus, in effect, the translation effort was reduced by 40% for a cross-platform project.

Experience has shown that there are few changes from one model year to the next, unless a model is redesigned completely. Thus, when the manuals for the next model year arrived, there were perfect matches of between 65% and 85%, resulting in further decline of translation necessity. Imagine this trend to continue, especially with the use of the multilingual legacy database mentioned above. It seems likely that after a few iterations, there will be little new text to translate.

New technologies to be developed, however, will account for temporary bursts in MT use until they are fully documented. But the amount of completely new developments in the foreseeable future e.g. in the automotive industry will likely remain limited to innovations such as hybrid drive technology (electric motor/combustion engine used in the Toyota Prius, for example) and fuel cell technology which is expected to be in full production within the next ten years.[8] Once these technologies are integrated into the publication process and databases, MT as it is presently used will certainly be phased out.

While I have not had much direct insight into the use of MT in other industrial applications, it is my expectation that the automotive industry's path towards less translation and MT will be followed in other industries as well.

Of course, I am far from sounding the death-knell for MT. There are too many other uses in which MT is firmly entrenched. And there will be the need for MT in environments that are characterized by high volume, dynamic content and rapid turn-around requirements such as we experience now on the World Wide Web. There also is an increasing focus on developing MT applications for languages far off the mainstream.[9] As translation technology makes the cost of translating large volumes of text ever more economical, and as new languages are added to the roster of MT pairs, MT will thrive in other areas, as we can already see on the Internet.

Even in these environments, however, MT will not be able to stand alone much longer in the long term. It is my prediction that users on the Net will not be satisfied with "gist" output much longer but will demand language quality in all they read on the Internet. Also, they will not much longer be satisfied with the strict limitations in terms of input formats that the existing services offer (txt, rtf, html). Half-hearted attempts to sell customers on raw translation with a cursory corrective glance (the so-called "content validation" being offered by some companies) will ultimately fail for various reasons.[10] The solution still lies within a fully integrated translation system in which MT plays but one part, albeit a pivotal one.[11]

In summary, I feel that there is a mixed future for MT, as long as development does not stop and the base output quality of MT continues to be improved. MT as a stand-alone tool will be seen only in the real-time applications such as Internet and Intranet translation of virtual meetings, email and chats. And even there, over time, it will have to integrate such processes as automatic format filtering, automatic pre-editing, etc. In other translation applications, MT *per se* will recede into the background and become one module in an integrated process. This is already visible in Enterprise Translation solutions such as L&H's iTranslator and Transcend's Enterprise Translation Server: The type, maker and other specifics of the MT engines used are hardly even mentioned anymore. Yet, as it fades from the view, Machine Translation will become a ubiquitous tool.

---

[1] A recent review of these activities can be found in Ann H. Adams, et al., "Developing a Resource for Multinational Writing at Xerox Corporation", in *Technical Communication, Journal of the Society for Technical Communication,* Vol. 46, Number 2, May 1999, pp. 249-254.

[2] To this day, and based on the experience with the Xerox dictionaries, I strongly believe that one of the keys to high-quality MT output lies in a very restricted general dictionary rather than in using the overblown and unmanageable dictionaries with which MT companies adorn their products. But this is a tangent which cannot be followed in this paper.

[3] It should be noted here that over the last few years, MT quality has been improving again at Xerox, see the article cited in Footnote 1 above.

[4] Unfortunately, not all translation companies seem to agree with this point of view. I have had access to MT output from at least two companies which seems to have been processed through MT without any preparation or even concern about the output, rendering the raw translation all but useless.

[5] See for example Kurt Godden, "Machine Translation in Context", in Farwell, Gerber and Hovy, eds. *Machine Translation and the Information Soup.* Springer-Verlag, 1998. pp. 158 - 163..
[6] Kurt Godden, "Machine Translation in Context", in Farwell, Gerber and Hovy, eds. *Machine Translation and the Information Soup.* Springer-Verlag, 1998. pp. 158- 163.
[7] During a private discussion with the director of a translation company
[8] See "Mikrowelle serienmäßig", an interview with the Director of Fuel Cell Development at DaimlerChrysler, *Der Spiegel* Nr. 36, Sep. 4 2000, pp. 168-170.
[9] See for example Hemant Darbari, "Computer Assisted Translation System - An Indian Perspective", in *Machine Translation Summit VII, Proceedings,* pp. 80-83.
[10] For one, it will be difficult to find translators/editors who will do this kind of work in the long run. After all, for content validation, the editor must be fluent in both the source and the target language, and it takes almost as much time to read both source and target and to correct factual mistakes as it would take to do a complete post-edit. As a bare-minimum service, its cost would be much too high to be economically feasible. Furthermore, several free-lance translators told me they would never do this type of work, as they would not want the risk of having their name put in connection with a text of raw translation.
[11] I have previously outlined such an integrated system. See Walter Hartmann, "The Next Step: Moving to an integrated MT system for High-Volume Environments", in *Machine Translation Summit VII, Proceedings,* pp. 266-271.