The Emerging Role of Machine Translation
Alex Yanishevsky, PROMT

Introduction

In *De Revolutionibus Orbium Coelestium,* Copernicus posited a heliocentric vision of the universe. Of course, people did not immediately cease to believe that the sun revolved around the Earth, but the pronouncement certainly signaled a paradigm shift. The necessity for a paradigm shift also became clear in the localization industry with the advent and popularization of computer aided translation (CAT) tools. Currently, machine translation (commonly referred to as MT), will be the harbinger for a similar change. Leading industry analysts concur with this view: machine translation has been identified as one of the top three valuable technologies for localization and arguably even for global economies.[i] This article will look at the forces that have driven the re-emergence of machine translation and will explore how the localization industry can best address this new paradigm.

History

Machine translation has a long, complicated and somewhat checkered history. To appreciate how far machine translation has come and to what degree it continues to succeed, it is worthwhile to step back briefly and understand its origins. After all, humans have always tended to wonder about the phenomena they observe. In the same way that somebody first asked why our bodies do not move in the same way in all directions (what makes falling different from going up?), we humans have always wondered why we speak different languages rather than one, particularly since living in a society requires the ability to communicate with one another. The idea of a universal language can be traced back to Biblical times with the building of the Tower of Babel. Much later in the 17th century, René Descartes proposed a universal language, where equivalent ideas in different tongues would share one symbol. Then, in the late 1830s, Charles Babbage, considered the father of modern computing, put forth the idea for an analytical machine, which among other functions, would be capable of storing dictionaries. This notion was to machine translation what Leonardo Da Vinci's sketches were to modern aeronautics.

During the 20th century, machine translation experienced a meteoric rise, an equally meteoric fall, and resurgence at the end of the century continuing to the present time. Warren Weaver, one of the pioneers of machine translation, mentioned the possibility of using computers to translate between languages in a memorandum in 1947. A mere seven years later, theory was put into practice during the celebrated Georgetown-IBM experiment in 1954. The "electronic brain," as it was called in the IBM press release, managed to produce an intelligible translation of 60 sentences from Russian into English using 6 grammar rules and 250 dictionary entries. Professor Leon Dostert, a Georgetown language scholar and of the project leaders, proclaimed, "[in] five, perhaps three years hence, interlingual meaning conversion by electronic process in important functional areas of several languages may well be an accomplished fact."[ii] Unfortunately, this triumph was eclipsed by a scathing report from ALPAC (Automatic Language Processing Advisory Committee) in 1966 which noted, "…we do not have useful machine translation [and] there is no immediate or predictable prospect of useful machine

translation."[iii] After this pronouncement, funding for development of machine translation systems became drastically curtailed.

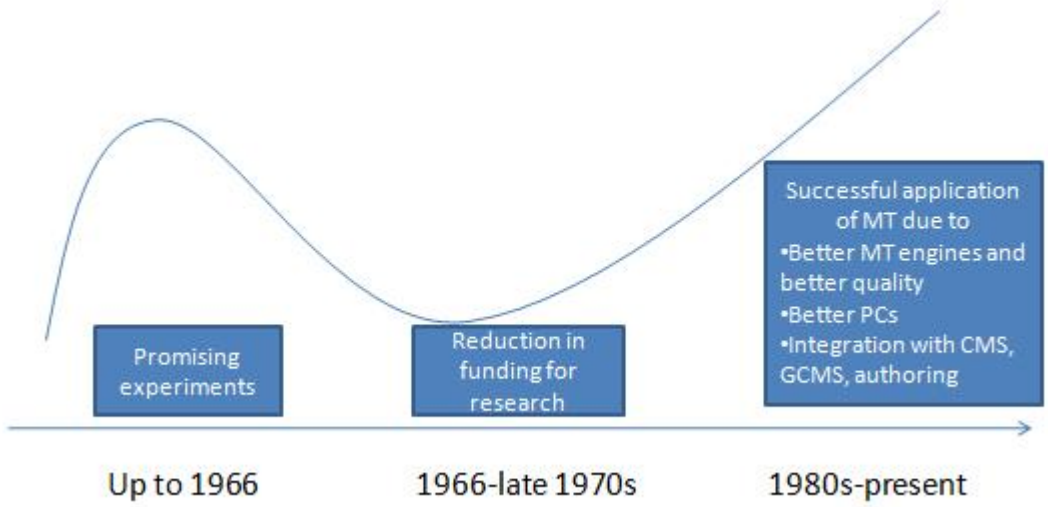

Figure 1. A brief history of machine translation in the United States from the 20th century to present.

After languishing in relative obscurity for about a decade and a half, the machine translation industry re-emerged with renewed vigor due to several factors.

- First, machine translation performance inevitably benefited from the staggering advances in hardware and software witnessed by the computer industry during the last two decades of the 20th century.
- Second, the consolidation of Web 2.0 increased the amount of data to be translated, and even caused new translation needs to arise, following the dramatic proliferation of user generated content. The potential market for industry and user-generated content combined has been estimated at close to 100 billion dollars. User-generated content, in particular, has several features which make it suitable for machine translation. On the one hand, it is produced in great quantities and is updated rapidly, which means that it requires on-the-fly translation. On the other hand, it is also less likely to be translated by human translators (due to its relatively ephemeral character). In other words, there is an enormous body of information constantly generated at unprecedented speed which is also being increasingly lost behind language barriers.
- Third, the market has been a driving factor in industry development. More than ever, there is growing pressure for getting to global markets faster and cheaper; both of these factors favor automated translation solutions. This is easy to understand: whereas having a human translator translate an inherently linguistic product such as a novel is nothing but necessary, having a human translator use human capabilities and time to translate product catalogues or accounting

> information is extremely expensive and can reasonably be equated to having a neurosurgeon put first-aid bandages on thousands of people a day.

Machine translation provides a solution to the challenges of both market cost-efficiency and fast-paced information generation, allowing a higher throughput of translation for a fraction of the price of human translation. The readiness of machine translation for prime time, as it were, is echoed by industry analysts as well, "Although available for decades, machine translation is now a **viable, game-changing opportunity for competitive advantage** by allowing organizations to provide dynamic translation processes."[iv]

In fact, dramatically increasing volumes of data available today have actually made new translation approaches possible, which were unthinkable of only a few years ago. Besides the genuine rule-based philosophy, in which the text is processed according to hand-crafted linguistic generalizations that are based on statistical patterns observed in human languages, quantitative methods have appeared which actively take advantage of large databases: whenever these contain the same information in two (or more) languages, a strict sentence-to-sentence correspondence and in some cases on a more granular level, phrase to phrase and even word to word correspondence between the texts can be easily established. By merely determining which word sequences in one text systematically correspond to words in the other, a machine can either match (statistical machine translation) or generalize (example-based machine translation) equivalences in order to generate a translation. Since the availability of this type of multi-lingual resources (known as *parallel corpora*) has grown by orders of magnitude during the last years, this method has become a feasible approach to providing adaptable, robust and even self-learned translation software.

Each approach has its own merits, but also its own shortcomings. For example, parallel corpora must be reliable and must contain voluminous units in a specific domain to propose a useful translation. Hence, the latest technological development consists of a hybrid approach that integrates the advantages of the above three, resulting in an even more accurate and relevant translation. As both sheer power and processing speed improved with the help of more robust translation engines, ideas of yesteryear about being able to translate languages on the fly finally became a reality. Since then, online translation engines and translation engines embedded into MS Office and Open Office applications have become so commonplace that experienced users cannot imagine their lives before translation software was available. The conceptual change in going from a monolingual scenario to a multilingual scenario can be equated to an overnight transition from a monopoly to a free market economy.

What content is a candidate for machine translation?

The section above provided a compelling argument for the use of machine translation. Clearly, this second time around, machine translation is here to stay. However, to be fair, we must acknowledge that there is a particular type of content that is better suited for machine translation. Although anecdotal evidence (translation of *les enfants et les femmes enceintes* as *pregnant children and women*) and apocryphal evidence (*The spirit is willing, but the flesh is weak* translated into Russian and then translated back as *The*

whisky is strong, but the meat is rotten) of mistranslations from machine translation engines is amusing for cocktail parties, critical considerations are not being taken into account: either translation engines are provided with unsuitable text, relatively speaking, or people expect the engine to act like a human being. If anything, objectively it would be more just to fault human writers for writing poorly in the first place and fault translators for howlers like "degenerativnaya geometriya-дегенеративная геометрия" (translation of degenerative geometry into Russian) since they, unlike a machine translation engine, can actually do research. So what content is a candidate for machine translation?

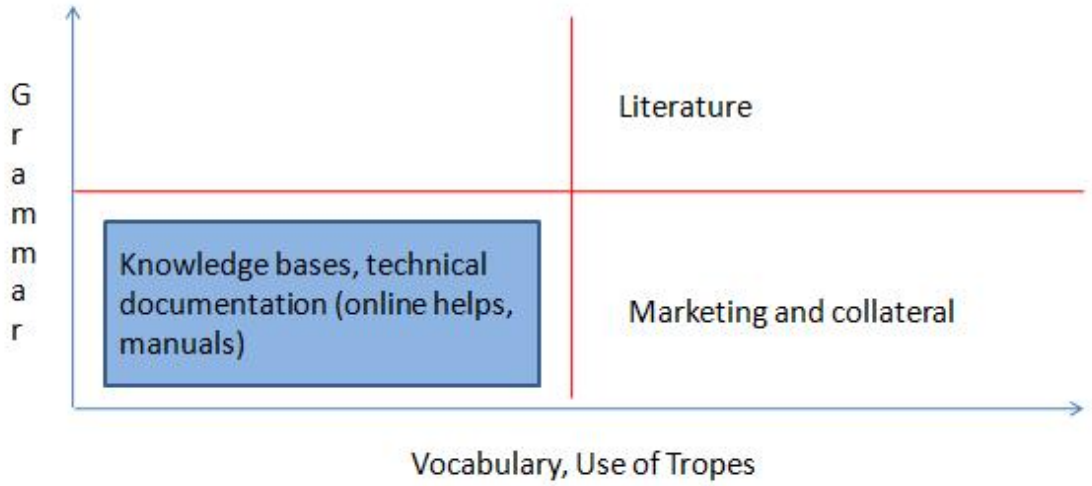

Figure 2. Machine translation is best suited for language with a restricted or controlled vocabulary and grammar.

If we use English as an example, although English has almost 1 million words, most technical communication requires only 1% or about 10,000 words and examples such as controlled English used at Boeing or Caterpillar can use a mere 1,500 words! Thus, content can run the gamut from individually created worlds with subjective meanings as in literature to tighter controlled writing as in the aviation industry or in auto manufacturing. Content that employs freer, unfettered language and uses figures of speech (similes, metaphors, etc.) is less suitable candidate for machine translation as it is difficult for engines to disambiguate such tropes. In enterprises, marketing materials typically fall into this more literature-like category since brochures frequently use figures of speech to tout their products. Conversely, content that is repetitive more restricted or fettered or even at its extreme, strictly controlled, is a good candidate for machine translation. In enterprises such content may include knowledge bases, manuals and online helps. Naturally, the expectations of the end user should be fulfilled and balanced by the localization budget. Frequently, knowledge bases are machine translated without subsequent post-editing, but with a disclaimer alerting the user to the fact, whereas manuals and online helps require human intervention. Success stories at companies like Caterpillar, Boeing, Microsoft, Océ and Symantec, to name but a few, prove that a restricted source language (English, in this case) coupled with machine translation can produce significant savings. Océ reduced costs by 60% by implementing an all-

encompassing program that included machine translation (MT), translation memory (TM), Controlled English (CE) and Extensible Markup Language (XML).

Addressing the new paradigm

As with any new paradigm, implementation of machine translation in a localization workflow or program requires re-evaluation of the status quo and the willingness to understand and effect change.

Four factors are crucial to successful implementation:

1) **Education** of all involved parties.
    a) For companies, education means becoming aware of the full range of activities encompassed by the content supply chain, beginning with terminology management, spanning controlled English and machine translation, and concluding with human post-editing. For localization service providers, education means becoming aware that machine translation takes translation technology, already successfully in use, one step further.
    b) For translators, the last link in the content chain, education means realizing that machine translation is to them what microwave ovens were to modern families. Cost-efficiency gains can in fact be quantified: research indicates that machine translation and post-editing achieve translation speeds as high as three times that of translation performed by unaided human translators.[v]
2) **New cost schemas**, namely, the question of which charges are appropriate for the different categories. Potential solutions involve either charging differently depending on the category or charging per word or hour for post-editing translation output, or some combination of both.
3) Development of a new position, **the post-editor**. Machine translation will drastically change the language industry landscape in terms of cost, throughput and workflow. Once more, the comparison with the early days of translation memory tools seems apt, since not only did translators not lose jobs because of computer-assisted tools, if anything, computerized tools enabled them to satisfy the growing demand stemming from the increasing bulk of content to be translated. The translators who benefited the most were those who included translation memory in their toolbox more promptly. Likewise, the translators who promptly become familiar with machine translation technologies will also be those who will be able to meet the ever growing market demand and will yield significant increases in throughput and cost reduction, leading to a higher return on investment and, therefore, obtaining a competitive advantage.
4) Working with the **machine translation engine**. It is widely accepted that there is no such thing as a “perfect” translation. There can be as many translations as there are translators. Consequently, an effective machine translation post-editor must leave behind subjective opinions and specific stylistic considerations and must focus rather on addressing systematic phenomena and improving lexicographical entries. Rectifying such errors is tantamount to an economic investment, as it will lead to further engine fine-tuning and exponentially increase cost-efficiency.

Conclusion

The ever burgeoning interest in machine translation is evidenced by more frank discussion of the subject by clients themselves, as well as increased inquiries and passionate discussion at conferences by localization service providers and translators. Consequently, much like computer-aided translation tools in the late 1980s and 1990s, machine translation solutions will become mainstays by dint of necessity. The role of a human translator will never cease to exist. On the contrary, it will expand to include a post-editing role. Exciting new horizons will open up and bring their own unique challenges and rewards. Our task is to recognize the role machine translation will continue to play in the coming years and prepare for the new challenges that we, as clients, localization service providers, technology companies and translators will be facing. Language barriers are merely a reflection of increasing human diversity, and translating from one language to another will always be a human need like the need for understanding. After all, until Copernicus' work was translated, the earth revolved around the sun only in Latin!

---

[i] Gilbane Group. Gilbane.MCBI.Report.July.2008.pdf

[ii] http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html

[iii] http://www.hutchinsweb.me.uk/MTNI-11-1995.pdf.

[iv] Gilbane Group. Gilbane.MCBI.Report.July.2008.pdf

[v] Allen, J. (2004). Case Study: Implementing MT for the Translation of Pre-sales Marketing and Post-sales Software Deployment Documentation at Mycom International. In *Machine Translation: From Real Users to Research*. Springer Berlin/Heidelberg.

NOTE: The author would like to thank his colleagues, Anthony Alfonso, Olga Beregovaya and Jordi Carrera for their valuable input and corrections.