January, 2009

Dispelling the myths of machine translation

Uwe Muegge

Chapter 20

Dispelling the Myths of Machine Translation

by Uwe Muegge

Uwe Muegge
Corporate Terminologist,
Medtronic

Uwe Muegge is the corporate terminologist at Medtronic, a manufacturer of medical technology. He serves in ISO Technical Committee 37 SC3 Computer Applications in Terminology, and teaches Terminology Management and Computer-Assisted Translation at the Monterey Institute of International Studies in Monterey, California.

info@muegge.cc
www.muegge.cc

It is not surprising that myths, half-truths, and misunderstandings abound regarding machine translation: It seems as if the experience most players in the translation field have with this technology does not go beyond toying a little with one of the free online translation tools. Almost every week, I come across an article informing its readers either that machine translation is and always will be a complete waste of time or that machine translation, while being a waste of time today, might actually be useful some time in the distant future. In the hope of setting the record straight, here is a closer look at some of the most common myths about machine translation.

Myth: Machine translation simply does not work

With free online translation services available all over the Web, anyone can run a text through a machine translation (MT) engine and then share the results with the public as proof of the fact that machine translation is capable of little more than the most rudimentary rough translations (gisting), and, of course, providing nearly endless entertainment.

The main problem with these 'tests' is that using any of the free online translation environments gives only a glimpse of the true power of a full-fledged professional machine translation system. For example, the typical online translation service does not allow users to select a subject field or provide user terminology, let alone set stylistic preferences. In fact, many—if not most—of the free text translation tools support no translation parameters other than the specification of the language pair and the source text. No wonder that the translations these machine translation websites produce can be so ridiculously off target.

Fact: Machine translation improves the productivity and consistency of human translators

Whenever new source text for a project is created, that text will have to be translated at some point. Even when you work in what is considered a state-of-the-art globalization environment, i.e. an integrated content management/translation workflow system, you will end up with a certain percentage of low match/no match sentences.

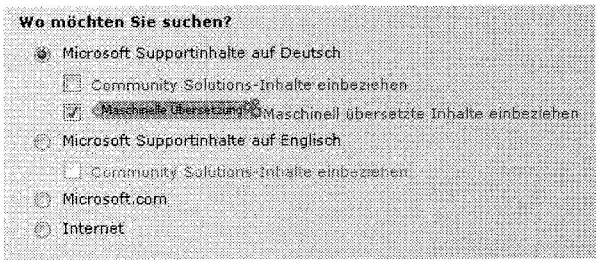
In a well-planned and well-managed globalization project where writers, as well as software developers, use a comprehensive project glossary and a style guide aimed at easy readability/comprehensibility, the low/no match sentences can be pre-translated in a machine translation system before being edited by human translators.

The benefits of machine-generated pre-translation include:

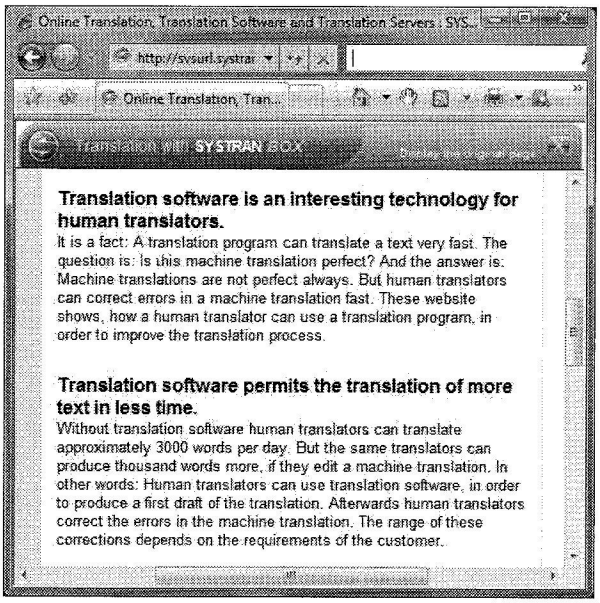
- Translators always have a proposal to work with instead of starting each new translation from scratch. A representative case study recently conducted at Symantec indicates that the productivity of human translators can double when unknown sentences are pre-translated in a machine translation system.[1]
- While most translations will require some editing and many even rewriting, it is fair to expect that a considerable percentage of machine-generated translations turn out to be perfect (this is especially true for short instructions, headings, legends, and the like).
- At a minimum, key terms will be translated correctly and consistently. And not only that, in most cases these terms will also be inflected correctly and appear in the correct singular or plural form. (Try to do that with your translation memory!)

Fact: Machine translation enables the translation of material that would otherwise not be translated

Very few organizations, if any, currently translate all materials into all the languages spoken by all of their current or future customers. The primary reason for this is that for many types of documents, especially in the after-sales domain, the budget is simply not available for large-scale human translation.



German search page for the Microsoft Knowledge Base with machine translation option enabled



Example of a German>English machine translation from the author's website, www.muegge.cc

A number of organizations are using machine translation solutions for making large volumes of text available to their global customers in their local language without involving any human translators in the process. The Microsoft Knowledge Base, which contains more than 200,000 documents in English, is a well-known example of a text repository where the number of machine-translated documents by far exceeds the number of those translated by humans.

Myth: Machine translation systems can only handle word-for-word translation

The belief that machine translation is basically limited to the sequential substitution of words in the source language with words in the target language is as widely-held as it is wrong. All popular machine translation systems, including the free online translation services such as systransoft.com, translate.google.com, and windowslivetranslator.com employ highly sophisticated algorithms that are the result of years of research and development.

Fact: There is not one but many very different machine translation technologies that are all capable of producing excellent translation results in the right environment

Machine translation has been around for more than 50 years, and during this half century a wide range of MT technologies have evolved (e.g. dictionary-based, rules-based, example-based, statistical) plus countless hybrid forms. Here is a brief discussion of the three machine translation technologies that are most relevant for commercial applications today.

Rules-based Machine Translation

Rules-based machine translation, also known as transfer machine translation, is the dominant MT paradigm today. SYSTRAN, Babel Fish, and @promt, to name just a few, are all rules-based systems. Rules-based MT systems use a three-stage translation process:

1. Analysis: Parses the source sentence to create a tree of the syntactic structure of that sentence.
2. Transfer: Converts the syntactic tree for the source language into the corresponding tree for the target language.
3. Generation: Populates the target tree with corresponding words to create a sentence in the target language.

The benefits of rules-based machine translation include:

- Mature, proven technology that can be implemented quickly and at relatively low cost.
- Many commercial systems available, covering many language combinations.
- Highly customizable through dictionary and style settings (some systems also support the customization of the rules base).

Rules-based machine translation systems have been in use in commercial settings for many years, e.g. at Autodesk, Daimler, and the European Commission's Translation Service. The two primary challenges for rules-based MT are first, that the rules base of any system is by necessity limited, meaning that for best results, authors need to adjust their writing style; and second, while commercial rules-based machine translation packages are available for dozens of language combinations, many languages are still not covered.

Statistical Machine Translation

Statistical machine translation (SMT) is getting a lot of media attention these days, especially after Microsoft announced that it is using a proprietary SMT system to translate its huge Knowledge Base document repository,[2] Google won a large-scale machine translation evaluation contest with its statistical machine translation engine.[3] Statistical machine translation systems typically consist of two major components:

- Translation Model: Generates translation proposals based on corresponding word sequences in aligned source and target training data.
- Language Model: Selects the best translation proposal based on training data in the target language only.

The good news about statistical machine translation is that once an SMT system has been trained on customer-specific data, this is the MT technology that typically produces the highest translation quality. On the flip side, that training effort requires a substantial body of existing translations: Language Weaver, the leading vendor of statistical machine translation systems, recommends a bilingual corpus of two million words or more per language pair. Because of the demanding training requirements,

combined with the fact that statistical machine translation systems tend to have a higher sticker price than some of the rules-based systems, this MT technology is primarily used by government agencies—the intelligence community in particular—and large corporations.

Direct Machine Translation

In its most primitive form, the only thing a direct machine translation system does is to replace the words in the source language with words in the target language in the same sequence and without any linguistic analysis or processing. The only resource direct machine translation uses is a bilingual dictionary, which is why this MT technology is also known as dictionary-driven machine translation. Due to this rather unsophisticated technology, direct machine translation has been considered obsolete for many years, and there are hardly any commercial products available that use direct MT.

Despite its limited capabilities, I strongly believe that direct machine translation still has a place in today's arsenal of automated translation tools. For a number of common real-world applications, word-for-word or phrase-for-phrase substitution is all that is required for successful translation. Think of domains where both vocabulary and syntax are standardized, as is the case with weather reports, financial profiles, and many e-commerce applications.

In one recent implementation, Medtronic, a large medical device manufacturer, used direct machine translation to translate a large product database into multiple languages.[4] Human translation was not an option for this project because of cost and, yes, quality concerns (an analysis of previous human translation projects indicated an unacceptably high error rate among numeric values such as product numbers and dimensions). Also, initial tests had shown that both translation memories and rules-based machine translation systems produced poor results with text that has the following characteristics:

- little or no repetition on the sentence level;
- high repetition on the word/phrase level;
- telegraphic/elliptic style, e.g. 'winds from southerly direction, speed reaching 55 km/h', 'American Technology Associates (AMTA) strong buy, Avion (AVIO) market outperform', or 'plate 2456dr15 right-angled, slotted, 15 ea'.

This type of translation project is most definitely among those that any self-respecting human translator could easily do without. And since direct machine translation does not require human post-editing in a best case scenario, using MT in this kind of environment might for once be welcomed by translators (who would hate to do these translations themselves) and translation buyers (who would love the idea of almost instant, almost free translations).

Myth: Machine translation is only for large organizations

Yes, it is true: If you read any success stories about machine translation, they typically come from the Caterpillars, Microsofts, and Symantecs of this world. But that is true for many—if not most—emerging technologies. It is also true that some of the most powerful machine translation systems in use today are the result of the multi-million dollar research and development programs only corporate giants can afford. But that does not mean you have to spend big bucks to deploy a machine translation solution.

Fact: Being both affordable and user-friendly, many machine translation packages are available for even the smallest of businesses, including freelancers

If you do a little research, you will find that many commercial machine translation packages are in the same price range as their translation memory counterparts, and that is mostly true for both workstation solutions for single users and client-server solutions for many users. And the secret is out that while corporate and small-business versions may differ in many ways, the core translation engine is typically the same in both products. In other words: In terms of out-of-the-box translation quality, there is generally little, if any, difference between the $1,000 professional version and the $50,000 corporate version of a given machine translation product.

In addition, the developers of commercial machine translation systems have invested heavily into making their products as intuitive to use as possible. In fact, I would even say that it is easier—and certainly faster—to produce your first translation with a typical MT product than it is with the typical translation memory tool.

A few more facts to consider:

- Many low-priced machine translation products feature a built-in translation memory (TM) module to improve the efficiency of the post-editing process ('never correct the same mistake twice'), and a few MT systems (like @promt Expert) offer seamless integration with the SDL Trados translation memory system.
- A number of translation tools vendors (such as Across) that cater to small and mid-sized companies offer TM-MT system bundles and/or MT integration via API.
- User education and MT system customization (e.g. building dictionaries), which are major factors in achieving the best possible translation results, are often easier to accomplish in smaller organizations than in larger ones.

The Bottom Line

Since its inception, machine translation has been a highly controversial technology, and it will probably continue to be so for some time. Much of this controversy is based on false assumptions about what machine translation can do and who might benefit from using this type of technology. Let me say it loud and clear: In general, the commercial machine translation systems available today cannot replace human translators, especially when those MT systems are operated by users who have no linguistic background. However, when the goal is to improve the efficiency of the human translation process or to create comprehensible translations in environments where human translation is not an option, and when these systems are operated by trained and motivated translation professionals, then machine translation is and has been a very powerful solution.

---

1. **Systran Software Inc. 2007.** Systran Case Study: Symantec. *Systran Software Inc. website.* [Online] 2007. [Cited: June 6, 2008.] www.systransoft.com/download/case-studies/2007.12.Symantec.pdf.

2. **Microsoft Corporation. 2008.** Machine Translation - Home. *Microsoft Corporation website.* [Online] 2008. [Cited: June 6, 2008.] http://research.microsoft.com/nlp/projects/mtproj.aspx.

3. **Institute of Standards and Technology. 2006.** NIST 2006 Machine Translation Evaluation Official Results. *National Instititue of Standards and Technology website.* [Online] November 1, 2006. [Cited: June 6, 2008.] http://www.nist.gov/speech/tests/mt/2006/doc/mt06eval_official_results.html.

4. *Fully Automatic High Quality Machine Translation of Restricted Text: A Case Study.* **Muegge, Uwe. 2006.** London: The Association of Information Management (Aslib), 2006. Proceedings of the Twenty-eighth International Conference on Translating and the Computer. ISBN 978-0-85142-5.