MT's 'perfect storm,' Russian and beyond

Lori Thicke

I If 2010 has a localization mantra, it's sure to be "Do more for less." Product managers want to extend localized products to their customers in international markets, purchasing managers are looking for overall budget reductions, and project managers, as always, are required to hit seemingly incompatible targets for cost, speed and quality.

Meanwhile, content to be localized has now outstripped human capacity to translate it. According to a 2009 *BusinessWeek* article, "Even with today's . . . cutting-edge technology, there are more words to be translated than most companies or governments could ever afford to handle. This shortfall limits opportunities for companies to market and support their products across languages, and to conduct business on a global scale."

Incredibly enough, this explosion of corporate content is dwarfed by the proliferation of user-generated content (Web 2.0), which smart companies can use to tap into their communities of international users — if they have the capacity to navigate colossal volumes of multilingual content.

More content . . . more languages . . . more speed. With the global financial contraction now added to the mix, the imperative to reduce costs is at the center of virtually every localization decision. If this "perfect storm" of factors hadn't already put machine translation (MT) on the agenda for most enterprise-class companies, the September 2009 White House paper, "A Strategy for American Innovation: Driving Towards Sustainable Growth and Quality Jobs," issued a strong call for

"automatic, highly accurate and real-time translation between the major languages of the world."

MT is poised to hit the mainstream.

Russian MT in the real world

PROMT, a technology company founded in St. Petersburg, Russia, is well placed to respond to this challenge from the Obama administration and from the market for ever-better MT technology. One of the leading rule-based MT (RBMT) engines in the marketplace, PROMT engages 82 of its 200 employees worldwide in resource and development.

Founded in 1991, PROMT was a spin-off of several research projects in the former USSR. The founders, Svetlana Sokolova and Alexander Serebryakov, were both graduates of the Department of Applied Mathematics and Mechanics of St. Petersburg State University. Sokolova, the technical lead of the computational linguistics lab at Saint Petersburg Pedagogical Academy, famous for pioneering Russian RBMT research, went on to found PROMT. With an approach that was revolutionary at the time, the first software release was a success on the market. That first MT engine was developed for the Russian language, and with offices in St. Petersburg and Moscow, PROMT still has its roots firmly in Russia. However, nowadays PROMT also has a strong focus on the Americas and Europe, with sites in Boston, San Francisco and Hamburg. Two decades after its first English-Russian engine, PROMT now covers seven languages in 21 combinations, is also offering Simplified and Traditional Chinese through a partner integration, and is currently working on adding more language pairs.

Adobe Systems localizes over 70 products into upwards of 32 languages, and as a result, localization is a significant portion of the product development budget and the product release timeline. Adobe has recently been introducing MT into its localization process. Ray Flournoy is Adobe's senior program manager of the Machine Translation Initiatives, based at the company's San Jose office. "We began our first experiments with Russian MT about a year ago, with our first large-scale production localization

Lori Thicke is cofounder and general manager of Lexcelera, cofounder of Translators Without Borders, and a member of the MultiLingual *editorial board.*

project completed during June 2009," says Flournoy. "Our experiences with Russian MT have been positive, with all projects so far showing efficiency gains from using MT plus post-editing instead of translation from scratch."

Like many enterprise users of MT, Flournoy's metrics are based on tangible productivity gains. They track post-editing time and compare it against the time required to translate from scratch. Improvements are made directly in the engine so that "as the engine quality improves, the post-edited text quality should remain at the same high level. Only the time required to produce the post-edited text should decline."

For Adobe, Russian MT has produced efficiency gains just slightly below the results they see for languages such as French or Spanish, which are closer in structure to English and thus perform better. Preliminary results indicate that for these languages, the MT post-editing was performed approximately 40% to 45% faster than human translation. "We are seeing efficiency gains with Russian that range from 18.5% at the low end to 40% on the high end," says Flournoy.

Adobe's future plans are to expand its use of MT beyond document localization. According to Flournoy, "future uses include publishing raw MTs of our online help documents and using raw MT output to catch UI problems from text expansion earlier in the development cycle."

Challenges of Russian MT

PROMT's fearless approach to building language models may spring from its experience creating an engine for an inherently thorny language pair, English <> Russian, with more "amenable" languages such as the FIGS+P – particularly French and Spanish – posing less of a challenge. According to Olga Beregovaya, CEO of PROMT Americas, Russian doesn't lend itself willingly to MT because "Russian is very complex, and in many cases the behavior of, say, noun phrases is much less predictable than with Germanic or Romance languages." Alex Yanishevsky, senior solutions architect, PROMT Americas, would agree with this analysis: "Certainly, achieving better results in Romance and Germanic languages is a quicker endeavor since Russian, in contrast, is a highly morphological language."

One issue in Russian is that a person's name may have a different ending in a passive construction. In the example below, the new ending to *Alex* is underlined:

Alex did this.
Алик сделал это
versus
This was done by Alex.
Это было сделано Алик<u>ом</u>

Impersonal constructions also pose a problem, as Russian has an implied pronoun. For example,

Светает

(*The dawn*) is breaking.

With only the verb in use, *the dawn* is implied.

PROMT has developed some 800 paradigms of word inflections for Russian, compared with only about 250 for English. Although it is not every language that has a corresponding "best fit" in terms of MT approach, this complexity of the Russian language is one reason why Russian, along with German, is better suited to RBMT. Future research for language pairs will necessarily involve getting away from English as either a source or target language. As Flournoy sees it, "I've been expecting that the market would start demanding more MT engines for language pairs that don't include English. Russian already has a head-start on that because PROMT started with Russian as its pivot language; however, I would expect that we will see even more of this. In particular, there are very few engines that cross from the Western languages to the East Asian languages. I would expect that a Russian <> Chinese engine or a Russian <> Japanese engine would fill a growing, undiscovered market demand in the coming decades."

RBMT-SMT debate continues

The technology underpinning the PROMT engine relies on linguistic rules, an approach PROMT shares with other engines such as SYSTRAN and Lucy, while Asia Online, Language Weaver and Google represent another camp: statistical approaches relying more heavily on algorithms.

In the commercial localization space, rule-based engines such as PROMT and SYSTRAN still dominate the relative newcomer, statistical MT (SMT). According to a recent report by the Language Technology Centre (LTC) on the language industry in the European Union, "the number one in machine translation continues to be SYSTRAN, followed by Google in second place."

The two companies represent approaches that have both strengths and weaknesses. RBMT is based on rules as well as on dictionaries that establish specific terminology (by domain, company or product line), making it a candidate for increasing terminological consistency in localization projects. SMT models are trained on the massive quantities of text they have been exposed to, generating greater sentence fluency.

To illustrate the difference between the rules approach of RBMT and the statistical learning of SMT, take the classic example of *the black cat*. Translating it into French, RBMT has the hard-coded rule that the adjective follows the noun: *le chat noir*. SMT will also translate this correctly because chances are that it will have seen this particular combination of words in its training.

If you change *the black cat* to *the blue cat*, RBMT will still get it right because the rule remains the same: adjective after the noun. But unless SMT has seen other examples of *the blue cat* in its training material, it is likely to deliver up, incorrectly, *le bleu chat*.

Despite this unpredictability of SMT, at the end of the day, both systems, if correctly trained, are capable of producing the type of results commonly reported by companies:

■ 18% to 50% cost savings on post-edited (fully human) quality, for a variety of content types,

■ up to 95% cost savings on raw (that is, non-post-edited) MT, such as might be used for search analytics and customer support,

■ at least a twofold increase in productivity.

However, there are important differences in the two approaches. The main strengths and weaknesses of the two approaches can be summarized as follows:

Set-up and customization: RBMT is quicker to set up if it is based on a system that already supports the language "out of the box." In this case, the training effort is concentrated on adding terms that may be missing or incorrectly translated in a given context, with a granularity that can go as fine as the individual product level. Dictionaries are arranged in priorities so that the user interface for a given product, for example, takes precedence over that of another product, even within the same family. SMT, on the other hand, enforces the terminology that was in the training material, whether or not it is the desired terminology in a particular context.

Resources: SMT requires engineering and processing power for its training, while RBMT requires specialized linguists.

Training materials: A good RBMT engine can be trained on glossaries, translation memories and/or a relatively modest amount of bilingual text. SMT requires millions of segments of clean bilingual and monolingual data — from, say, a million segments for a FIGS language to four to five million for Japanese. Since not every company is able to access the amount of material needed to train an SMT engine, its best choice becomes a rule-based engine like PROMT. As Yanishevsky says, "when all is said and done, SMT is really a creation *ex nihilo* and relies on voluminous, reliable and clean data, and all three conditions are often difficult to meet."

Terminology: RBMT's strength is term consistency, with dictionaries to enforce terminology. SMT, which looks for the most likely term, is generally unpredictable in this sense although Asia Online also integrates dictionaries for greater control.

Sentence fluency: If RBMT offers more terminological consistency, SMT generally offers more fluent sentence structures.

Adding new language pairs: SMT's great advantage is that it is infinitely easier to add a whole new language combination — say, French > Swahili — because there are no grammatical rules to code. Adding a new language pair is simply a matter of feeding in sufficient content in that language pair.

Updating and continuous improvement: RBMT can be retrained on the fly, with continuous updates possible, even on a daily basis; SMT retraining typically has longer cycles, so improvements are slower to integrate.

Integration into workflows and translation management systems: Both systems make APIs available to assist in integration.

A multi-engine future

While the rule-based-versus-statistical debate rages, enterprise users are taking a pragmatic approach, according to Common Sense Advisory's report, "The Business Case for Machine Translation." The report states that users should "expect to buy — and integrate — multiple MT products. One engine often won't address all MT needs; organizations requiring a variety of languages from different linguistic families will likely find themselves with more than a single machine translation solution."

Enterprise users have discovered through trial and error that the choice of MT may well be language dependent. While some languages, such as French and Spanish, can achieve impressive results through either RBMT or SMT, others such as Russian and German are best managed today within a rule-based process. Although it may be slightly more complex to integrate two different systems into their workflows, many corporate users are finding the better quality of the best-matched language models outweighs any implementation issues.

Hybrid advances may radically change this picture. Many MT engine suppliers are now moving toward a hybrid approach, blending the best features of both into a single engine. SYSTRAN released its hybrid in 2009, and PROMT will be following suit with its hybrid version in 2010.

Says Yanishevsky, "clearly, hybridization will be the development of the future for both SMT and RBMT engines. However, fundamentally, we believe that it is faster and more efficient to hybridize with rule-based underpinnings than with SMT underpinnings since it is easier to graft statistics onto rules rather than vice versa. The hybridization of our MT engine will linguistically smooth an already robust and quality output."

One of the conclusions of the LTC report "The Language Industry in the EU" states that "it seems very likely that the use of machine translation will grow to cater for exponentially rising translation needs in increasingly globalized contexts." Add this to the call of the Obama administration for more and better MT, and it seems clear that 2010 will indeed be the year that MT enters the mainstream. **M**