The World According to Gap

By Jost Zetzsche

Knowing that John Irving himself delights in obscure word plays, I am sure he will not mind this echo of his most famous title. Besides, it just seems too appropriate when talking about machine translation (MT) and the rest of the (translation) world. For this is a story of gaps, some of which are shockingly wide, and just like the story of Garp, Irving's protagonist, it is a rather bizarre tale.

Though it is not (always) their own fault, the MT community is faced with a number of seemingly unbridgeable gaps with pretty much every group out there:

- The translators who feel threatened by MT and love to ridicule it;
- The multi-language vendors who use it to drive traffic to their websites, but have not really seen much of a positive return on this strategy;
- Translation memory (TM) developers and users who view their systems as vastly superior, but forget that they are a mere sibling of the same base technology; and
- Members of the general public who start off with unrealistic expectations; when they are inevitably disappointed, they continue to use MT, but disparage it passionately.

Though it is not (always) their own fault, the MT community is faced with a number of seemingly unbridgeable gaps with pretty much every group out there.

And, of course, there are the governmental agencies that constantly expect much more from MT technology for their research grants, the mid-sized businesses that often misuse MT technology because they do not truly understand it, and so on.

In this article, I want to look at two of these gaps—between the MT community and translators and between the MT and TM communities—and see what can be done to bridge those.

The Translation Community

I am a translator myself, and while I am not sure that I can speak for the translation community as a whole, let me try to work through some of the conceptions that we may have. ➡

First of all, the perception of MT as a threat to job security seems only too natural. This is not helped by some of the marketing promises of the MT community. And there is no need to look far. The most well-known provider of MT systems, Systran, promises this for several of its products: "Dependable and proven, it translates documents, e-mails, web content, chat, and more—at a substantial savings over traditional translation services" (see www.systransoft.com). This is not really a marketing strategy that endears it to the heart of the translator.

Still, I can overcome that fear once I become more familiar with the limitations of MT technology. I can then see that MT produces quality that is usually not publication-ready when it works in an uncontrolled language environment such as one of the many web-based machine translation engines (this, of course, is where I get all the great examples that I can ridicule).

I can see that the level of success of MT in controlled environments is much higher, but I may still find it objectionable because it is stylistically inferior. After all, I may have the same dilemma that many of my fellow translators have: I value my work—translation—so much that no matter what I translate, be it a marketing text, legal disclaimers, news releases, or user manuals, I try to apply the same kind of excellence. In fact, I may even frown at e-mails from clients that tell me to "really spend every effort" to make a certain translation impeccable because it is part of a bid or some other high-level job. I do not like to be told that because it obviously implies the assumption that I am not always working on that level.

As honorable as this may be, it creates a problem when I forget to distinguish the purposes of the different texts, what audience they are intended for, and what the respective quality requirements are.

Marketing content or literature lose their very purpose and meaning if they are not translated in a way that impacts the user (the reader) far beyond the actual information. In fact, the language in these kinds of text has to be so powerful that it manipulates the user beyond that which he can control (be it through emotions, value propositions, or shopping behavior). Compare that to a legal text. In this case, information in all its detailed nuances is of the utmost importance. Readability is of secondary concern, but ambiguities have to be avoided.

For user guides, information is also very important, but readability or stylistic concerns differ, depending on the user type. If it is for engineers or developers, there is less concern about style than there would be if it were intended for an end-user. After all, any communication with end-users also carries some marketing message that could be thwarted by terrible writing.

And if there are different kinds of expectations by human users, there are also computers. For instance, most of the vast amounts of translated intelligence material is being processed by computers. Who wants to be a translator in that kind of scenario? Apply high quality standards for the translation of something that no one but a computer will ever "read?"

If I, as a translator, have come this far in my thought process, I will probably conclude that it makes no sense to have materials translated by highly qualified human translators when it can be done by computers as well.

But Can It Be Done by Computers?

The answer is that often it cannot, but sometimes it can. In a unique project, Microsoft created MT versions of tens of thousands of knowledgebase articles into several languages. For an example, go to http://support.microsoft.com/kb/281925/en-us and then click on one of the translation links on the right-hand side. You will see an MT version of the article in the respective language, preceded by a disclaimer informing the user of possible pitfalls of the translation. The translation is not pretty, but it communicates (most of the time) what otherwise would not have been communicated at all.

So now I, as a translator, may realize that what we need is to develop usage criteria for translation. For the majority of usage criteria, human translation is of the utmost importance. For others it may be computerized translation with human post-editing, and for still others it may be MT only.

And would I really want this as a translator? Sure. Who wants to waste talent on stuff that a computer can do? I also know that computers will not take away my job security. They may at some point take away certain kinds of jobs, but there is plenty of inter-

We need to develop usage criteria for translation.

esting material that currently is not being translated because it would be too expensive. That is what I would like to do.

Hey, that was not so hard after all! Maybe there can be peace between the translators and the MT community!

(And maybe it would help if MT providers tweaked their marketing message just a bit….)

The Translation Memory Community

It would be futile to restate what Jaap van der Meer said when he eloquently summarized the state of the gap between the different branches of computer-aided translation technology in *MultiLingual Computing*'s issue 71 (Volume 16, Issue 3) from 2005.

> Disdain on the side of the professional translators for the hilarious and stupid MT mistakes gave birth to a new variant of MT called translation memory (TM). TM started off as a lower-level feature of commercial MT systems such as ALPS AutoTerm. But the success of TM came with dedicated products such as IBM TM/2 and TRADOS. The marketing message was tuned in to what the professional translation industry wanted to hear: 'Forget about MT; it doesn't work. Instead, use our TM product because it leaves you in full control of the process.'
>
> The message worked well. Within a period of 10 to 15 years, TM products have found their way to the workstations of more than 50,000 translators in the world. But the message has also caused a sort of 'cognitive disorder' in the translation industry, namely that TM is good and MT is evil, foregoing the fact that TM is just a new variant of MT, closely related to the school of thought around EBMT [example-based machine translation]. The damage is done, however, and it will take years to convince the community of business translators that post-editing fuzzy matches from TM databases is, in fact, not different from post-editing fuzzy matches from any other MT system.[1]

Van der Meer very pointedly describes the gap that was artificially created between the two siblings as a result of the TM side's marketing message. (He does not mention that the MT side also played a part in creating the schism by either looking down on or, at best, ignoring its TM relation.) Regardless of fault, however, the key question comes down to this: Is there a way to reconcile these two groups? Or, put differently, is a reconciliation even desirable?

I would argue that there are two ways where it is desirable and, in fact, inevitable.

Rather than repeating what van der Meer has said about the history and development of MT technology, I would much rather refer you to his article again. He describes the two main schools of thought in the MT community as the linguistic rules-based and data-driven approach. In short, the rules-based approach relies on a syntactical analysis of the source language as well as on dictionaries. It then attempts to transfer that data to the target language to synthesize it in target language-specific rules and language. The other approach is that of data-driven MT. Here there are no linguistic rules, just large amounts of raw, bilingual data that are processed so that the translation system produces translation by sheer statistical means. While there used to be gaps between these two camps, there is now an increasing realization that a hybrid model would be most desirable. As a result, even Language Weaver (see www.languageweaver.com), the poster child of the data-driven approach, is experimenting with incorporating syntactic information that is learned automatically from data. And the above-mentioned system by Microsoft also uses a combination of both technologies.

Maybe there can be peace between the translators and the MT community!

For the pure rules-based approach, bilingual data was not necessary (aside from dictionaries), but it is at the very core of the data-driven systems. And here we come to a somewhat awkward phenomenon: the much-shunned community of TM users has assembled huge repositories of bilingual data, arguably the largest collections of readily available bilingual data anywhere. And rather than investing massive amounts of resources to produce that data, it was produced nearly accidentally as 0a by-product of translation with the help of the many available TM tools. Owners of that data have previously treated this data as just that—a by-product. However, awareness of translation processes and technologies have changed, and the data is typically now centrally controlled by the translation buyer

rather than the service providers.

In this area of almost ironic convergence, it is the rebellious and shunned prodigal son who comes home with a large treasure chest, just waiting to be employed by the slightly pompous parent. (And you say there is no drama in the world of translation!)

A second area of gap-bridging lies in the way that the translation memory market has developed since the two siblings separated. Tools that started off as low-level TM tools have developed into full-fledged translation environment tools. These tools not only use TM technology, but typically have very sophisticated terminology management facilities, file conversion utilities, project management components, and quality assurance processes—many features that MT tools have neglected to develop.

At the same time, the translation environment tool market is still rather crowded (and becoming increasingly so with new tools appearing left and right), providing a glut of competitors to chase after the market-leading spot now occupied by SDL TRADOS.

To me this seems like a phenomenal opportunity for a merger between one of the many TM contenders and an MT vendor. Though this was attempted with limited success when SDL purchased ALPNET's and Transparent Language's technologies, the stakes are different today. Data is not only the key to the MT engine, but also to the data-based MT engine. A new and very attractive tool proposition could be offered to the translation industry with all the bells and whistles that existing translation environment tools offer.

But why a merger of these technologies? Here's why. The bare-bones operation of translation memory technology with database-lookup procedures for fuzzy and perfect matches is really quite simple, but the results are strikingly good. A perfect match in a well controlled environment is as good as it gets, and even the very best MT technology will not make this any better. The only problem with TM technology is that there is only a finite number of matches, and this is where an MT engine that is continuously trained by the same database the TM relies on can perform what it can do best: translate for post-editing.

The publisher of Irving's *The World According to Garp* introduced the book with a now-famous dust jacket description: "This is the story of T.S. Garp, the bastard son of Jenny Fields…." While it is tempting to play on our topic's similar paternity issues, I will gladly refrain. After all, Garp suffers a rather violent death in the novel, and none of the stakeholders discussed in this article are likely to suffer the same fate. Translators, TM, and MT are here to stay. What will change is that the gaps between them will dissolve.

Notes

1. http://fm.multilingual.com/FMPro?-db=archives&-format=ourpublication%2ffeaturedarticlesdetail.htm&-lay=cgi&-sortfield=Magazine%20Number&-sortorder=descend&-op=cn&Author=van%20der%20meer&intro=yes&-recid=33690&-find=.