Running on empty?

Despite some dire predictions when SDL bought Trados a couple of years ago, the landscape of translation environment tools (aka CAT tools) is alive and well. Not only are there a good number of new contenders gaining ground in a surprisingly dynamic market, but they are also introducing interesting new technologies, both in the form of workflow management and in increasingly fine-tuned terminology and translation memory processing engines.

Photo: James Steidl

By Jost Zetzsche

But it's this last point that makes me wonder. We have a fleet of beautifully tuned engines, but just like mechanical engines, these don't run without fuel – in this case, the data. When translation memory technology entered the scene about 15 or 20 years ago, it just seemed natural for everybody to "dig for their own fuel" by performing human translation that would be retained in the translation memory databases and later leveraged for future translations. While everyone saw the benefits of this technology right away – things like increased consistency and leverage of content within the same document – the true benefit only showed up as the size of the databases increased. A 2004 study on Translation Memory demonstrated the dramatic increase in the usability of translation memories when they have been used for five years or more.[1] So, to stay with our engine/fuel analogy, we dug in our own backyards to find some fuel, experienced modest success in the beginning, and now with the passing of time our engines are being fed more consistently.

Translation data sharing – vision or reality?

Are they running at 100 percent yet? Definitely not. In fact, I would guess that they are barely in double digits!

Can you imagine what kind of output these translation memory engines would produce if they were not only fueled by their own home-grown data but by the data of entire vertical industries? As a consumer, can you imagine what it would mean to finally have common and well-formed terminology across competing products? Or as a translation buyer, can you imagine what it would mean for your product to finally compete on the merits of its functions and features rather than translation and terminology? Or to be able to translate not only in the most common European and Asian languages, but also afford to do the "smaller" languages?

Ultimately, this is what the vision of translation data sharing is about. Sound a little too visionary? Here are some very recent and real initiatives that are working toward exactly that vision.

Creating a cooperative platform

TAUS, the Translation Automation User Society, organized a summit in March of this year with representatives of 26 organizations – among them multinationals such as Adobe, Cisco, eBay, European Patent Office, International Monetary Fund, Microsoft, and Oracle – "to explore how a cooperative platform for sharing language data can potentially increase levels of translation automation, through for instance advanced leveraging and training of machine translation systems"[2]. While the participants of the summit recognized that there are definite hurdles to overcome (such as legal questions, classification of data, and infrastructure), the next meeting to tackle some of the hands-on questions is already scheduled for fall.

What's the value of your data?

Another initiative is the licensing scheme that TM Marketplace has been offering for a couple of years. Rather than giving data assets away to competitors, this concept looks at the value of the translation memory data that has been assembled over many years, puts a price tag on it, and sells licenses for its use. For instance, General Motors is offering TM Marketplace licenses for more than four million of its translation segments (app. 37 million translated words) in six language combinations to other vendors in the automotive industry, language vendors who are active in the automotive industry, or machine translation developers.

Why would a large corporation like GM do that? First, because they have an interest in aligning their suppliers and competitors to their terminology (since they are the first automotive market leader to offer their data, their terminology will likely become the standard). And second, GM has realized that there is no reason to keep the value of their data asset locked up in their corporate vault when it could just as well be realized in new and innovative ways.

Tailored terminology for freelancers

And there is a third initiative, the "made-for-order model," done without any direct participation of the original data owner. Huge amounts of bilingual data in the form of PDFs, web pages, or various other formats are available for download on the internet. While these data sources are not prepared to be used as translation memories, the emergence of industrial-strength alignment tools and expertise makes it possible to turn these documents into bilingual translation memory data. These translation memories can be "custom-ordered" and tailored for specific industries or products.[3] This last model is obviously geared toward a very different target audience, typically freelance translators. And while it may not directly support a "higher level" of data sharing, the least it might do would be to encourage data owners to find more profitable ways to part with their data.

Free fuel from one common source

To come to a point of "free-flowing (data) fuel" across vertical industries, a number of things need to happen, including the development of legal structures, classification schemes, quality control, an infrastructure that would support the flow of data including the necessary tools and standards, strategies to merge existing and differing terminologies and much more. But really these are very minor considerations compared to the one primary development that makes it all possible to start with: Data owners need to come to the realization that their data holds redeemable value to others, and that it is in their interest for others to use their data. Once this cornerstone falls into place, the others will follow.

Not only is it encouraging to see some very prominent companies such as GM already participating in data sharing, and others like the TAUS members actively pursuing it, it guarantees that the vision of commonly produced and shared data fuel will come to pass.

[1] LISA Translation Memory survey 2004, www.lisa.org/products/surveys/tm04survey.html

[2] TAUS press release, www.translationautomation.com/downloads/TAUSSummitNewsRelease.pdf

[3] You can find a White Paper on the legal ramifications at www.tmmarketplace.com/whitepapers/align

contact

Jost Zetzsche is the president of the translation memory broker TM Marketplace.

jost@tmmarketplace.com
www.tmmarketplace.com