Column

Enterprise Innovators

Lori Thicke

Building low cost MT

EMC is a global leader in enabling businesses and service providers to transform their operations and deliver information technology through innovative products and services, such as cloud computing and storage. EMC has over 53,500 employees with sales offices and partners in more than 85 countries around the world. Valarie Gilbert joined EMC nearly two years ago as a senior director in EMC's Services and Support team, building tools and systems for online self-service problem resolution. She is a graduate of Carnegie Mellon University and the University of Pittsburgh, holding BS and MS degrees in metallurgical and materials engineering. Valarie also earned a certificate in capability maturity model integration from Carnegie Mellon.

Valarie Gilbert, EMC.

Thicke: At 16, you were one of the youngest women ever accepted at Carnegie Mellon. What does this say about you and how did that prepare you for your career?

Gilbert: It prepared me to confront scary things face on. At the end of the day, I don't like a problem to be ahead of me. I want to face it, figure it out, solve it, move forward. This sometimes means not doing the popular project or taking the easy road. It's not recognition that motivates me, but the pleasure of getting things done and getting them behind me.

Thicke: How did you get involved in localization given your core background in engineering?

Gilbert: Many years ago my career migrated from aerospace materials research to systems engineering — a transition that many with my background were compelled to make. In those days, I had no exposure to localization or machine translation (MT). Eventually, when I began leading several portal prototyping and development teams, my manager asked me to take on localization of the support website, the knowledge base and support forums. The charter was to develop localized support sites and forums that would run parallel and be in sync with the original English versions. The organization wanted a time-to-market process and solution that we could replicate across a number of languages. Content in all three support areas changed frequently and required a workflow process that was easy for authors and contributors. Thus, MT seemed the right answer. True, I had no experience in localization, but career transitions and technology change had never stopped me before. Eventually, with time and investment we built a robust, multilingual MT workflow process.

Thicke: What makes your current implementation of MT so unique?

Gilbert: Well, I don't mean to imply that our current structure is fully robust. Nor are we the only EMC team interested in building MT, but we are experimenting with what we're calling Low Cost MT. It is a small-scale system with limited language capabilities and without the risk of high capital investment. What I think is so unique is that we've built a useful set of tools with limited resources in a short time. Normally, when teams embark on building MT capabilities, there is a variety of commercial solutions and vendors waiting at your doorstep offering installation and consulting services. Typically, you gather a team of translation professionals and build a staff of linguists and engineers with lots of automated translation experience. In our case, our initial team consisted of myself and Pablo Vazquez. As you know, Pablo has a great deal of MT experience, having been in the industry for most of his career. But obviously he and I could not do it alone. We needed additional help building our use cases and establishing an operational process. Requesting further resources would require a formalized business case for investment. Again, all these things take time, and Pablo and I wanted to move quickly. We decided to enlist our existing staff of expert analysts and storage engineers who were eager to embark on this challenge.

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

Thicke: How did you begin?

Gilbert: Sensibly, we wanted to first start with training and support for our engineers. There's a limit to what they could learn from handbooks and white papers. They obviously had to get their hands dirty. We talked to a number of technology providers and to our surprise; many did not want to help. They simply weren't in the business of putting together training programs or providing support to get our engineers up to speed. But after some digging and calling on industry colleagues, we were able to locate a few companies that had several options. We needed a technology-agnostic approach that would teach our engineers the benefits of both knowledge-based and corpus-based translation automation. They needed to know the innovations of a hybrid approach and where each of these techniques would be most useful. We settled on what I'll call "Company C," which provided training for several commercial and open source solutions and guided our team through a self-build process of MT implementation.

We also picked a group of engineers who had a specific interest in this area. They were located in Europe, the Middle East and Africa as well as Asia-Pacific and Japan. English was, at most, their second language. Each of them had already experienced our need for more multilingual content and were looking for ways to fill the gaps between limited local documentation and core English content. In their daily jobs, they used online generic translators to better understand existing English material. They knew the difference between a perfect translation and understandability. Our efforts were of personal interest to them because it could make a difference in their daily jobs.

From the very start, these team members were enthusiastic about training and quality. They were eager to learn how to teach the engines what was in their head; to impart the small nuances of their native language. Bilingualness is a great advantage when trying to jump-start MT.

Thicke: How did they structure the training? This wasn't an off-the-shelf program. They obviously created something customized for your needs.

Gilbert: Exactly. "Company C" created a translation automation boot camp customized for our diverse group. The program had two main components and was both formal and flexible. The formal portion was a step-by-step discussion of technology. They spent four days reviewing MT principles, translation memories (TMs), reviewing the XLIFF file format and the principles of translation management. The flexible training allowed the engineers to dive into specific areas of interest. One team focused specifically on parsers for Chinese, while others were far more interested in statistical parameters and impacts of the corpus. Thus, we built a team with a broad-based knowledge and a range of skill sets.

Thicke: Where are you now?

Gilbert: Well, after the boot camp we created an infrastructure team with five basic areas. The first one was MT training, focusing on the specifics of how to train our engines, and how to improve the training process. Then we defined the pre- and post-translation processes, including filtering, parsing, normalization and even authoring requirements. Next was system availability and administration, evaluating how to get the best out of the MT systems in terms of performance, memory, clustering and uptime. MT quality looked at defining and executing our translation quality program; generating our test scores and calculations and establishing our predictive capabilities. Last of all was architecture and integration — improving ten of the MT engines and connecting them with the rest of the enterprise. And also creating our integrated workflow technology.

As a collective group, the team now has two main goals. Obviously, they want to continue perfecting their specific technologies. They are working hard to improve engine quality, automate our processes and make our systems more robust. But more importantly, we have to integrate into the broader enterprise and establish ourselves as part of the end-to-end operation. Our challenge will be to create an internal team that can support our open source MT. It will likely have to be much broader than just our team to accommodate IT methodologies. We'll need to conform to internal standards of uptime and performance. How do we support enhancements, new requirements and real time use cases like translation on demand? There is still much ahead of us.

Thicke: What do you feel are your successes?

Gilbert: In a period of just five months, we were able to take a group of technologists with no MT experience and later witness them having in-depth conversations about duplication of data, parsers and segmenting rules. They are passionate about improving our corpus and heavily investigating methods to automate quality assurance. I consider our "home grown" effort a great success. Our MT systems are working at the same speed as the technology providers and our scores are as good as some commercially trained engines.

We are certainly not in the business of developing MT engines; we simply want to use them. We didn't need to scour the industry for noted additional experts or high-priced vendors. With minimal support we were able to build in-house expertise with our own talented engineers. Going forward, we are working on developing a service-based infrastructure and training more engines, more languages and expanding the reach of the open source MT.

I think we're on the road to accomplishing our main goal, which was to introduce MT concepts to our broader team. Once they are more familiar, we want to run trials proving speed and quality of translations and establish some cost benefits against human translation.

Thicke: Reflecting on the overall experience, how do you think you did?

Gilbert: As I said, I consider our effort a great success. While we're certainly not done, nor have we translated bulk loads of content, we've put together a solid baseline team that we can build upon and grow our practice. The key to success was a combination of engineering skills; language skills; responsibility and ownership of interest areas; broad boot camp training; and a great training partner and a network of like-minded professionals interested in growing the industry.

Most importantly, we had access to an outstanding group of industry colleagues through the many associations and committees to which we belong. The TAUS series of conferences and meetings served as excellent forums for collaboration and troubleshooting through roadblocks. During MT workshops, we could discuss parameters and models to jump-start our testing. Access to the TAUS Data Association's TMs also gave us a robust corpus that accelerated our training.

While our next phase may include commercial engines, we are now certainly better buyers. We have profoundly increased our basic understanding of MT and can review commercial offerings much more critically. We grew and expanded the capabilities of our engineers and increased our internal IP. This is indeed our value proposition.

Thicke: What advice would you give to others?

Gilbert: My advice to others is threefold. If you are embarking on MT for the first time, don't hesitate to jump in. Building something internal is a great way to provide learning and experience for your team. There is nothing better than hands-on knowledge. Also, the right team members can bring enthusiasm and momentum. They can help make the technology concepts commonplace for your organization and remove barriers for those who think MT is untested. Automated translation is certainly not the right answer for all scenarios, but it is definitely a great addition to a localization tool set. Lastly, find a great partner who understands your strategy and listens to your concerns. It was invaluable to have a partner that was interested more in our goals than trying to sell their services. **M**