



MT at Symantec

Enterprise Innovators is a series of interviews with trailblazers whose innovations in language technologies are helping establish new best practices within the localization industry. In this first column of the series, Lori Thicke speaks with Fred Hollowood, research director of Symantec Corporation, a global leader in providing security, storage and systems management solutions. With more than 17,500 employees worldwide, Symantec is a Fortune 500 company headquartered in Mountain View, California. Hollowood is based in Dublin, Ireland.

Thicke: Symantec is known for being on the cutting edge of machine translation (MT) in our industry. What is your background in MT? Just how long have you been involved?

Hollowood: I got involved in 2003 in order to investigate MT as a possible technology for the translation of rapidly perishable security content. We were faced with a situation where there were few MT practitioners in the industry and limited consultancy services at our disposal. We did take some third-party advice, but we decided to work through the issues by taking on Johann Roturier, a Ph.D. student from Dublin City University. We recognized immediately that although this approach was a slower route, it was more sustainable as we could build our expertise and extend the culture in-house. There was a lot to learn in MT, source profiling and natural language processing. We were fortunate that Johann had a background in translation studies and we also had professional translators in-house, so we had a supportive, hands-on environment in which to test our initial forays into MT.



Fred Hollowood,
Symantec

Thicke: 2003 was pretty early for MT. Have you always been interested in pushing the frontiers?

Hollowood: I have always had a focus on technology. In the early days, delivering localized products required a lot of long hours and manual effort. Over the years, we've automated these processes, thereby releasing engineers and linguists to engage in higher value-added activities. I have enjoyed building the infrastructures and redesigning the processes to make this happen. We all produce so much more now than we did a decade ago. This is down to technology handling repetitive tasks while humans do what they are best at. MT is a fine example of this. By generating a candidate sentence that is preformed and contains the correct terminology, we allow the translator to work on a sentence that requires minimum editing – a simple goal, but not without complex implications.

Thicke: Since Symantec is one of the world leaders in software, we have come to expect something exciting from them. What's new in MT at Symantec?

Hollowood: We have been using SYSTRAN, a rule-based machine translation (RBMT) engine, for our product documentation, and the service has settled into our standard processes. MT is now part of everyday life in project preparation, and our vendors are accustomed to post-editing the output. More recently, we have involved ourselves in the statistical machine translation (SMT) world, largely with the Moses open-source system, and are considering its use in areas where content is not deeply tagged and the source is not controlled. Given that we are a global company with significant sites outside the United States, there are several content repositories inside the company that would

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

benefit from translation.

Thicke: Tell me about controlled source content and its implications for MT.

Hollowood: *Controlled* is a term we use when we talk about source text that has been written so as to avoid certain grammatical and stylistic structures. For example, many company style guides encourage the use of short sentences for comprehension and clarity. Another case would be the use of passive voice, often explicitly mentioned in style guides as it is frequently problematic in translation. Both these issues were important in maximizing the efficiency of our RBMT process.

Thicke: So, source content that is well written and contains metadata in tags is handled well by a rule-based approach. What about SMT?

Hollowood: As I mentioned earlier, we have been an RBMT house, using SYSTRAN, and it serves us well with our highly tagged and controlled source. It allows the post-editors to have well-defined expectations on both the type and predictability of the edits required. The SYSTRAN hybrid allows increased fluency of output in a number of languages. Our investigations into the Moses SMT engine have allowed us to evaluate its performance. We have not yet deployed it but can see possible uses in various content types, particularly with what I call lightly structured narrow domain content.

Thicke: What exactly do you mean by

narrow domain content?

Hollowood: When we look at our technical content at Symantec, we see several domains. Security and availability are immediately obvious. Yes, these two domains do share some terminology, but they also have terminology particular only to themselves. I would call these two domains narrow. The content one finds in each is particular to that domain and not the other. A broad domain is a domain that includes a wide selection of subject areas. A broad domain SMT engine covers general inquiries quite well but can never give specific detailed translation that a specialist would demand. Narrow domains have set terminology, requiring a specific target translation for a specific source. A rule-based engine generally gives you good control of terminology and is particularly effective when the content is highly tagged. I think it is difficult to get an RBMT engine to do a good job on a broad domain, however, because the dictionary conflicts become too complex.

Thicke: Could you give an example of content types that are suitable to SMT?

Hollowood: Any well-formed content is suitable for SMT as long as you have enough of it to train your engine in the range of languages you require. It can be a "chicken and egg" type of problem. You have to have translations before you can train the engine to translate.

Thicke: By publishing the results you are achieving with MT, Symantec has not only shown its leadership in this area, but it has also stimulated a great deal of interest in MT. What kinds of gains are you seeing?

Hollowood: In our product documentation we are experiencing throughput improvements in the region of 50% to 100% in various languages. That is to say that a translator is able to post-edit in excess of 5,000 to 6,000 words a day in some languages on a well-formed source. This allows us to deliver products in shorter timeframes at more advantageous pricing. Of course, some languages are more efficient than others. Using English as source, French, Spanish and Portuguese provide better results than German or Japanese. Even in the cases of Japanese and German, significant efficiencies are now possible.

Thicke: If there are few companies today deeply involved in MT, I'd say even fewer dare tackle "difficult" languages such as Japanese and Chinese. Tell me about the work you've done in this area.

Hollowood: On the research side, our focus has been to improve the reception of MT in Japan and China by improving quality. Yanli Sun, one of our Ph.D. students, was working on the translation of prepositions from English into Chinese within technical documents in an industrial localization context. The aim of the study was to reveal the salient errors in the translation of prepositions and to



Business Management

Project Management

Terminology Management

Translation Memory

Web Solutions, Workflow Automation, Web Services Interface Layer, Advanced Leveraging, External Collaboration Solutions, Machine Translation, CAT Tools, Authoring

MultiTRANS

Translation Management System

For Governments • Enterprises • LSPs

www.multicorpora.com
USA / Canada: 877.725.7070
Europe: +32 (0) 2.213.00.20

explore methods to remedy these errors.

This study first examined which prepositions were handled unsatisfactorily by our MT system. Based on this information, three novel approaches were proposed to improve the translation of prepositions. The approaches included building an automatic preposition dictionary for the RBMT system; exploring and modifying the process of statistical post-editing; and pre-processing the source texts to better suit the RBMT system. Overall evaluation results – either human evaluation or automatic evaluation or both – show the potential of our new approaches in improving the translation of prepositions.

Thicke: And in Japanese?

Hollowood: Midori-san has addressed a different but related area focusing on Japanese. She has worked on understanding human post-editing behavior, as this is crucial for reducing post-editing effort. There is a lack of large-scale studies on post-editing in industrial contexts that focus on the activity in real-life settings. This study observed professional Japanese post-editors at work. A mixed method approach was employed to both quantitatively and qualitatively analyze the data and to gain detailed insights into the post-editing activity from various viewpoints. The results indicate that a number of factors – such as sentence structure, document component types, use of product specific terms and post-editing behavior – all affect the amount of post-editing effort in an intertwined manner. The findings will contribute to a better utilization of some MT systems in the industry as well as the development of the skills and strategies of post-editors. The team has also published several papers in the last few months on MT, post-editing and automatic evaluation, so all in all it's been a good year in sorting through some of the fundamental issues in the area.

Thicke: In October 2010 at the TAUS conference in Portland, you made an announcement about sharing a tool you've developed.

Hollowood: Yes, another departure this year has been to contribute to open source. We released SymEval, our MT evaluation technology via SymForge. This is a useful tool, which measures the

differences in a test and reference document on a segment-by-segment (with tokenization) basis. This tool generates a general text matching score, which is useful for production coordinators, and highlights differences, which is useful for linguistic assessment. We are hoping the open-source community out there will use our tool and expand its functionality.

Thicke: What can you share with other companies wanting to follow in your footsteps as to a reasonable return on investment (ROI) timeline, considering the cost to train up the engines for each language pair?

Hollowood: The ROI is totally dependent on the volume pushed through the system and, of course, a word-based discount from your vendor. Training your engines or populating your dictionaries is paid back when you commit sizable content streams for localization. Rapid turnaround and consistency are benefits traditionally hard to quantify in monetary terms.

Thicke: What would be your advice for a company looking to deploy MT internally?

Hollowood: Look beyond what you are currently translating to consider the possible opportunity of providing quality gisting services on other company content, not traditionally associated with product. Companies

have a reservoir of content developed internally, usually in English, that is not translated into all of the languages of interest because of the huge costs involved. I have heard of estimates as high as 9x volumes more than is currently translated. This does not include the fast-developing area of customer forums, which are many times larger again. Applying automated translation services to these content silos is surely one of the next goals in the industry.

Thicke: In retrospect, what was one smart thing you did and one thing you wouldn't do again?

Hollowood: The one smart thing we did was to take the time to understand MT. There are many pitfalls and false promises along the way, and knowing what was happening and avoiding knee-jerk answers kept the project real. In the early days, I relied heavily on the quality assessments of translators. They were not always favorable. It was some time before I learned to temper these evaluations with automatic metrics and user evaluation of MT output. Giving the evaluations of end users a higher priority is key. For example, these days we are becoming accustomed to "less than perfect" grammatical rendering on mobile devices. If customers need immediate solutions, perfect grammar and layout are not their priority. **M**

audio localizations
in country, on time

London, Paris, Frankfurt, Milan, Rome, Madrid, Lisbon,
Prague, Budapest, Warsaw, Ankara, Athens, Amsterdam, Stockholm,
Oslo, Copenhagen, Moscow, Cairo, Tel Aviv, Mexico City,
Buenos Aires, San Paulo, Beijing, Taipei, Seoul, Sapporo.

BINARISONORI

www.binarisonori.com