The Machine Translation Maturity Model at PAHO

Julia Aymerich
Pan American Health Organization
525 23rd Street N.W.
Washington, DC 20037
aymericj@paho.org

Hermes Camelo
Pan American Health Organization
525 23rd Street N.W.
Washington, DC 20037
camelohe@paho.org

Abstract

This paper describes the work environment, IT tools and interactions that have come about after almost 30 years of MT use at the Pan American Health Organization (PAHO). We present the key ingredients that have contributed to the creation of a mature and stable environment for our translation service and provide some practical suggestions and solutions that we have worked out over the years.

1 Introduction

PAHOMTS® is a well-established rule-based transfer MT system, developed and maintained by computational linguists and translators in the PAHO Translation Services unit (TR). It has been operational at PAHO since 1980 (Vasconcellos and León, 1988). It currently translates in six language combinations (all directions between English, Spanish, and Portuguese, three of PAHO's official languages) and is used to process over 90% of translation jobs received at PAHO TR. Our average workload is 4.5 million words per year and our clients are administrative and technical units at PAHO HQ in Washington DC and other PAHO offices throughout the Americas.

The PAHO TR unit, then, has been using MT successfully for almost 30 years. This technology is fully integrated in our daily work and is the default mode of translation at PAHO. The system is also used by other administrative and technical units throughout the Organization and is licensed to external institutions. We have reached a highly stable and mature environment that we believe could be emulated by other institutions despite one obvious advantage at PAHO: we are both developers and users of the technology.

2 Keys for Maturity

The PAHO TR unit functions at a high-maturity level of continuous improvement, the highest level of the "translation readiness" maturity model proposed by Iverson and Kuehn (1998) and the highest level of software engineering maturity model proposed by Florac et al. (1997). This means, among other things, that there is a strong sense of teamwork, processes are continuously and systematically improved, common sources of problems are understood and eliminated, translated text is tracked and reused, glossaries are in place, translation costs and processes are reviewed periodically, a plan is in place for continually updating reference materials, new technologies are proactively pursued and deployed, and there is a commitment to translation quality and cost-effectiveness.

These characteristic attitudes were not achieved overnight, though. We gradually invest in projects in a controlled manner, involve all actors in the process, and continually evaluate our work. This kind of gradually evolving roadmap has been effective for us.

Being part of a multilingual organization, translation is a standard practice at PAHO. Tools developed in-house, together with available technology, are in place for continually updating reference materials such as bilingual corpora, MT dictionaries, and translation memories. The bilingual corpora are shared with others in the organization and with our freelance translators. The translation process, in conjunction with the MT software, is periodically evaluated.

It was understood very early on that using a static MT system in isolation would not work in the long run, that it would be essential to receive feedback from translators who were postediting the output, and that this feedback should be incorporated in the engine right away. This way, transla-

tors did not have to correct the same mistake twice if they pointed it out the first time. Additionally, revisers send the final version of a document back to translators so they can learn new terminology and stylistic practices and improve the quality of their work.

This constant interaction between translators, computational linguists and ad hoc terminologists has been the key ingredient in our success story, but not the only one. In order to reach our current level of maturity, the following ingredients have proved critical, each of which will be described in detail in the sections below:

1. Excellent and reliable software
2. A strong network where all players can interact
3. Integrated tools
4. Human component

3 Excellent and reliable software

PAHOMTS® was developed in-house over the years by a small team of computational linguists, who always worked side-by-side with the translators who were postediting the output. The software is designed to handle standard syntax, is used in a broad domain (not just in medical or public health texts), and the linguistically rich dictionaries are fully customizable. After the initial development efforts, maintenance and development have always been done using real documents submitted for translation rather than artificial preconceived grammar rules. For example, even though Spanish syntax does not allow for a comma between the subject and the verb phrase, it is a common mistake in Spanish to use this comma. PAHOMTS® simply accepts it and deletes it when rendering the sentence in the target language.

Beyond the initial development, PAHO managers understood that an in-house system like ours would need to be maintained and adapted to new terminology and software environments if it was going to be around as a long-term solution. Without in-house computational linguists, the project would not be sustainable. We currently have two staff computational linguists and we hire contractors to assist us with specific projects. Even in times of crisis, when posts were being cut right and left, our computational linguist posts were always respected. Under the guidance of our staff, PAHOMTS® has undergone some major adaptations: from the mainframe to DOS and then to Windows; from the Wang processor to WordPerfect and then to MS Office. A key aspect of the software is that it is able to maintain the format of the original document through customized format handlers, thus allowing the translators to concentrate on the words of the document and not worry about preserving the format.

Our rule-based system has also evolved with the times and has become more hybrid-like; it is currently used in combination with translation memories (TMs) and terminology tools, and we have plans to add a statistical postediting module.

4 A Strong Network

While the Machine Translation system is at the core of our operations, it would not be as useful if it were not integrated in a network of other IT tools and people.

4.1 A network of people

Four groups of people have direct interaction with the software or its output and the functions of each group are clearly defined:

1. Multilingual support staff receive and process electronic translation requests, perform spell and format checks in the source documents, locate reference materials when possible, select run-time options, run the translations, do some quality checks on the output to make sure it meets the requirements for postediting (check parser statistics, not-found words, preservation of formatting), interact with requesting units, staff, and external translators, prepare the final bitexts, deliver the translations to the requesting units and take care of other administrative tasks such as verification of funds, negotiation of deadlines, and invoicing. Our support staff receive intensive training in all the tasks they perform and have become specialists in the many tricks of word processing. Their presence and expertise free up translators from administrative and procedural tasks so that they can concentrate on translating.
2. In-house translators postedit MT output, occasionally do human translation (when there is no soft copy), revise translations done by freelancers, provide suggestions for the MT dic-

tionaries and grammars, check dictionary changes suggested by external translators, assist in the preparation of multilingual glossaries, maintain the PAHO style databases, prepare and distribute terminology notes, suggest new tools that would be useful in their work, coordinate translation work and interact with external translators to help them solve terminology and style problems, develop training materials for external translators and provide them feedback on their work in the form of documents with tracked changes and other notes, interact with requesting units and staff, help solve terminology and style questions of technical units, and interact with other professional translators in order to keep abreast of new developments in the field.

3. Computational linguists maintain the MT dictionaries and algorithms based on feedback from translators, adapt it to new software environments, maintain the in-house translation tracking system, prepare terminology lists to be imported into the MT dictionaries and glossaries, maintain the databases used by translators for their terminology research, coordinate the creation of bitexts from past and current translation jobs, develop new tools requested by translators, and run monthly and yearly statistics.
4. External translators postedit MT output, do research using the tools provided by PAHO, and provide suggestions for the MT dictionaries and grammars.

All in all, support staff and computational linguists support the work of all translators; staff translators support the work of freelancers. The entire system works well because the functions are clearly defined and the translators and revisers can concentrate on the actual task of translating and revising, and coordinating the work. It also works well because PAHO management has understood the importance of maintaining the tools developed in-house and has provided us with enough resources to do so.

4.2 A network of tools

MT is supported by an array of IT tools, most of which are home-grown while a few are commercial.

Editing tools: PAHO postediting macros for MS Office, synchronized postediting tools, PAHO style checkers.

Research tools: PAHO web page for translators and editors, which includes thousands of bitexts, glossaries, official documents, indexed web pages with PAHO and WHO conference documents, etc. A commercial tool is used for indexing and searching, but PAHO computational linguists and translators maintain the databases.

Translation memory tools: Translators use commercial TM software and the computational linguists use commercial alignment and terminology management tools.

Feedback mechanism: Each translation job comes with a "raw" translation file and a "side-by-side" (SBS) file which contains a table with three columns: source segment, target segment, and feedback. Translators enter their suggestions on this file, which is later passed to the in-house revisers for clearance and the computational linguists, who are able to use the context to determine the best way to implement the suggestion (Aymerich and Camelo, 2007). Originally, the feedback was handwritten by translators on the SBS file, which was a text file that could not be easily manipulated on the screen.

Tracking tools: PAHO Translation Tracking System (TTS) is used to store all the information about translation jobs processed. It is also used as a vehicle for communication between in-house and external translators, support staff, computational linguists, and requesting units. The TTS is also used to create translation statistics and to locate reference documents for new jobs.

4.3 Empowering the translator

At PAHO we have the ideal situation, because the people in charge of maintaining the MT software and related tools are in constant contact with the users and are sometimes users themselves. Instead of developing in a vacuum, we always check with the translators first and work on projects that will be useful for them. At the same time, translators are provided with guidelines for postediting, are usually provided with a revised version of their translation, and are informed when their feedback is incorporated in the engine.

Computational linguists and staff revisers also comment on the individual translators' feedback in

order to teach them which suggestions are most useful and which are not relevant (for instance, typos in the source), not clear or not implementable. Some of these comments are summarized in periodic newsletters, which report on number of terms incorporated and some examples, samples of useful and not useful feedback, percentage of translators providing suggestions, etc. The annual contracts for external translators specify that they are required to provide feedback for the MT program, but translators are not paid extra for the feedback provided. In general, our translators cooperate because they feel they are in the loop and they know their suggestions will be taken into account. They also feel that they are part of the development team and they have contributed to improving a product which will eventually benefit them as well.

5 Integration

One key aspect of a successful work environment is the integration of tools available to translators.

5.1 The work environment

We have found over the years that translators do not like to change their work environment. If they are postediting an MS Word document, they want to work in MS Word. This is why we have incorporated all tools into MS Office so that the translators do not have to move around too many applications while postediting a document.

Our postediting macros are integrated in an MS Word toolbar and an MS PowerPoint add-in. The macros allow users to perform editing operations in English, Spanish, and Portuguese, as well as look up words or expressions in the MT dictionaries, clean up MT marks, do accurate word counts, do intelligent uppercasing, etc.

The synchronized postediting toolbar allows translators to link the "raw" and "side-by-side" files so that they can see the source segment for any target segment they are working on and can provide feedback in a straightforward manner. Translators are given the choice to work on the raw file, which contains all the format of the original document, or on the SBS file, which ignores format and presents the text in three columns, as mentioned above. Either way, once the postediting is complete, the final version is synchronized so that raw and SBS versions match. The SBS file is later converted by another macro into a bitext, and the bitext is fed into the databases that will be accessible to all translators as a reference. For a full description of the PAHO macros, see the PAHOMTS® User Guide (2008).

Another tool that is integrated in MS Word is the PAHO style checker. This macro runs translations against a database of frequent errors made by external translators (maintained by the in-house translators), provides the explanation for each error and, in most cases, suggests a better alternative or preferred translation. The macro was developed by the computational linguists and it is run before the translation is revised as a first pass to detect and correct the most common mistakes. Some of the errors are terminological in nature and others are simply stylistic preferences at PAHO.

5.2 Combining TMs and MT

Another form of integration is the combination of Translation Memories and MT. We use the interactive version of Wordfast, which allows for easy integration with PAHOMTS®. Even though the documents we process are not at all repetitive in nature, TMs come in handy for a few types of documents. One example is the yearly financial report, which has some variation but the same core structure and general content. Another example of the powerful combination of MT and TM was evident during the recent influenza A (H1N1) crisis. For the initial five weeks of the outbreak, we were receiving daily "talking points" for the media from WHO. The countries in the Region were anxiously awaiting these daily updates early in the morning so they could be prepared to inform the public as soon as possible. The information was produced in English at WHO and was needed in Spanish. Similarly, we received a daily report from PAHO Emergency Operations Center at the end of the day, which also had to be translated into Spanish for immediate distribution. Additionally, we received some requests to translate large technical documents about influenza preparedness and related topics. For the first few days, we used MT exclusively and the computational linguists started preparing bitexts that were fed into the search databases. Once we had enough bitexts, we performed terminology extraction to create an initial glossary, and prepared a TM. After the TM was in place, we used it in combination with MT for the

incoming daily reports. This helped our two in-house Spanish translators produce documents with consistent terminology and style, and increased their productivity to an average of close to 5,000 words a day.

5.3 Terminology research tools

Yet another form of integration is the use of databases that can be accessed from the PAHO LAN by staff translators and other PAHO staff, and from the PAHO intranet by external translators. These databases, accessed through dtSearch and maintained by the computational linguists, are updated daily and indexed weekly (or more often, if necessary) and can be accessed by the translators from MS Word. In addition to public glossaries from PAHO and other international organizations, the databases include bitexts of translations that date back as far as 1991. The bitexts for the earlier years (1991-mid 2006) were created using a commercial tool, MultiTrans, and were checked by our support staff; the recent bitexts are automatically created using the synchronized postediting module described above. All bitexts are aligned at the sentence level and our translators report finding this tool extremely useful both for terminology research and to find how translation problems have been solved in the past. Full text searches are often more useful than mere glossary-style searches as they provide examples in context and make it possible to search for specific collocations. The computational linguists also use the databases to verify terminology feedback from translators before adding it to the MT dictionaries and to create bilingual glossaries that can be imported to the MT dictionaries.

It should also be noted that many dictionary entries in PAHOMTS® are coded for high reliability. These include official names of programs or institutions and terms that have a preferred translation at PAHO or that translators have a tendency to modify. Whenever a translation with high reliability is selected, the program adds special marks that are clearly visible for the posteditor. When he or she runs into these terms in the "raw" translation file, they know that the research has already been done and they do not need to look further into the correctness of the term. A postediting macro removes all such marks when the translation is complete.

5.4 The Translation Tracking System (TTS)

Finally, our in-house intranet-based TTS is the element that holds everything else together. It is essentially a tracking and workflow management tool for translation requests and is used by three groups of people:

1. Requesting units (the clients) upload their translation requests, along with instructions and relevant references. Clients can also track the progress of the job and download the final translation.
2. TR staff verify workload distribution and assign jobs by availability and specialization of translators; locate past translations that can be relevant for incoming jobs; create invoices and verify availability of funds; store source, target, and bitext files; and communicate with the clients and translators.
3. Translators, both in-house and external, receive their job assignments, download raw and SBS files, download references, communicate with in-house revisers and support staff, and upload their final translations and feedback files.

The TTS sends e-mail alerts that contain updated information about new or existing translation jobs and also contain links to the relevant page on the PAHO intranet for ready access. This way, no files are ever e-mailed as attachments; instead, they are all securely stored on the intranet server and uploaded and downloaded by the interested parties. Also, all relevant staff are copied on e-mail alerts so that the entire team is informed and can act on a request even if the person in charge is absent or busy. Similarly, team members from a requesting unit can download final translations even if they did not personally originate the request. Because all interactions between external translators and the TR team regarding a translation job are also stored in the TTS, the entire history of a job is available for future reference.

The database entry for each external translator contains contact information and areas of specialization; staff translators use a 1-5 star ranking system based on the experience and performance of external translators. Freelancers can also indicate their availability on their individual calendars, specifying both when they are not available for translation or the periods of time when their availability is high. Thus, when a new job needs to be as-

signed, the TTS displays a list of translators for the language combination in the following order: translations with high availability, translators who are currently working on a PAHO job, and translators who are currently unavailable. Within each group, translators with a higher ranking are displayed at the top. This system helps TR staff quickly locate the best person for each job.

Because all information about translation jobs is recorded in the TTS, we are able to produce accurate statistics on demand. The statistics include MT/HT distribution, in-house/freelance distribution and cost, number of words by target language, translator, event, or time frame, costs, cost savings with MT (an average of 33% savings), percentage of jobs delivered on time, workload by translator, etc. We keep monthly statistics to help us verify that no information is missing from the system. We also produce annual statistics for managers.

6 The human component

Despite the high level of automation of our processes, we do not dream of full automation and we acknowledge that the human component is critical. We will always need highly qualified individuals to postedit the MT output, to provide feedback, to perform a manual check of suggested dictionary and glossary entries and, most importantly, to encourage and assist one another. Freelance translators typically work in isolation and are in geographically diverse locations. By bringing them into the loop and making them feel that their contribution is welcome and they are part of a team, they tend to cooperate in the development process.

7 Translation Workflow Overview

The lifecycle of a translation job starts whenever a PAHO HQ Unit submits a translation request using the Translation Tracking System. Staff members are expected to log in on TTS and complete and submit an electronic Translation Request Form (TRF). All documents for translation are uploaded to the TRF in a suitable format. Background materials can also be uploaded, which should be in the target language (in particular, previous versions of the same document) for conceptual and terminological reference. Clients are also requested to indicate the name and contact information of the technical person who is responsible for the content (so that in-house translators know where to go to ask questions) and the purpose of the translation (so that the translator can adjust the register and style accordingly). Clients must also allocate funds to cover the cost of contractual translation services when necessary.

All translation requests are handled according to PAHO TR's priorities, date of request, and availability of resources. Requests are also checked against the bilingual corpora in order to gather background material, if available.

Requesting units can also ask for an estimate before processing a translation. After the estimated cost and delivery time are provided by TR, units can either cancel the request or ask TR to process the job. If contractual cost is too high, a requesting unit can negotiate or drop the deadline for the job to be done in-house.

Once a request has been approved, the work is assigned to an in-house or freelance translator. For very long jobs with tight deadlines, it is sometimes necessary to divide jobs between two or more translators. Jobs are assigned to in-house translators whenever possible. However, because our in-house capabilities are so small, 80% of the work is outsourced and later revised in-house, time permitting. Next, it is decided whether an appropriate TM is available. If so, the document is translated in interactive mode using a combination of TM and MT. Otherwise, an assistant processes the document in batch mode with PAHOMTS®. Human translation is used only if the source or target language is French, the source document cannot be converted to a format that is readable by the MT software (for example, a PDF file with graphics only), only a hard copy was provided by the requesting unit and its quality is not good enough for text extraction using OCR software, or the document is too idiomatic.

Work and reference files are exchanged by subsequent log-in and download in the TTS. Work notifications are sent to translators by e-mail. All translators have authenticated access to the TTS website from anywhere in the world. Each translator is assigned his/her own page to download content and source documents and to upload translated files and feedback. This page is accessible only to him/her and TR staff. It is also possible for TR staff to know the status of a translation job at any time.

At this stage, the translator actually translates or postedits the work received using the tools pro-

vided by TR (like the PAHOMTS® toolbars and the dtSearch page) or his/her own tools. This is the most important step as the main cost of translation is largely determined by how efficient an environment is provided to the translator.

When finished, the translator uploads the translated files, translator notes, and feedback for the MT software. The translation job is then routed for review (proofreading, terminology check, etc). Quality control tasks are performed in-house. The work is checked for translation accuracy and for overall document correctness. MT feedback is also checked for accuracy and cleared by the reviser. If changes were made to the translated file, the reviser uploads to the translator's page a version of the document with track changes activated.

When the final work is ready, the head of TR clears the job for delivery and an office assistant uploads the final documents to the requesting unit page in the TTS.

The TTS also calculates costs both for in-house and freelancers based on information entered for each job order. Word counts are done on the target file for Spanish, Portuguese, and French. For English, the word count of the raw file is used. For revision jobs, the number of hours spent on revision is recorded. Costs are calculated based on the rates recorded for each task in the TTS and invoices are automatically generated from the TTS.

If feedback was provided for a translation job, the TTS also sends an e-mail alert to the computational linguists, who update the MT dictionaries and translation engines accordingly.

Finally, the source and final translations are converted into a bitext file and uploaded to the corresponding bilingual corpus so that translators can access them as a reference for incoming jobs. Jobs processed using MT are automatically aligned using PAHOMTS® toolbars. Other jobs are aligned using commercial tools. Periodically, the computational linguists use the bitexts to perform terminology extraction and feed the MT dictionaries and the PAHO glossaries.

Conclusions

MT has reached a high point of maturity at PAHO because we have worked hard over the years to create a mature and stable work environment, and one that has evolved over time to adapt to new circumstances while still staying faithful to the original vision. One critical element has been the spirit of teamwork fostered among all actors in the process, by bringing translators into the development loop, providing feedback between computational linguists and translators and among translators, and promoting open communication and knowledge sharing. Another key element has been the seamless integration of MT and other tools into the workflow of our translation service. Despite the successful adoption of MT, this technology is still seen as a tool; it is one of the critical components of the way we work, but a tool nonetheless. We should never forget that this whole process is driven by and targeted to human beings with a high level of specialization and the human component is the central piece of the puzzle.

References

Aymerich, J. 2004. Machine Translation in Practice at PAHO. Tutorial Notes, at AMTA 2004: Sixth biennial conference of the Association for Machine Translation in the Americas. September 2004, Georgetown University, Washington, DC, USA. http://www.mt-archive.info/AMTA-2004-Aymerich-tutorial.pdf

Aymerich, J. and H. Camelo. 2007. *Automatic Extraction of Entries for a Machine Translation Dictionary Using Bitexts*, in *MT Summit XI*, 10-14 September 2007, Copenhagen, Denmark. *Proceedings*; pp.21-27 http://www.mt-archive.info/MTS-2007-Aymerich.pdf

Florac, W., R. Park, and A. Carleton. 1997. *Practical Software Measurement: Measuring for Process Management and Improvement*, Software Engineering Institute, April 1997.

Iverson, S., and H. Kuehn. 1998. *Assessing "Translation Readiness: A Maturity Model. Proceedings of the 45th Annual Conference of the Society for Technical Communication*, Arlington, VA: Society for Technical Communication.
http://www.stc.org/ConfProceed/1998/PDFs/00010.pdf

PAHO Machine Translation System User Guide. 2008. http://www.paho.org/english/am/gsp/tr/2008_PAHOMTS_UserGuide.pdf

Vasconcellos, M. and M. León. 1988. *SPANAM and ENGSPAN: Machine translation at the Pan American Health Organization*. Computational Linguistics 11, pp 122-136. Also in J. Slocum, editor (1988), Machine Translation systems, pages 187-236. Cambridge University Press