Using Machine Translation for fast, inexpensive, and accurate health information assimilation and dissemination: Experiences at the Pan American Health Organization

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

Abstract

The Pan American Health Organization (PAHO) has developed its own proprietary Machine Translation System (PAHOMTS), which has been used daily since 1980. The software translates between English, Spanish, and Portuguese, and each dictionary contains over 100,000 entries, with specialization in health terminology.

PAHOMTS allows users to gather health information in the desired target language by providing a fast draft that is easily understandable, fairly accurate, and preserves all formatting. This version can be used for gisting without the need for postediting.

When the translation is to be disseminated, the "raw" translation generated by PAHOMTS needs to be edited by a professional in order to obtain a final distributable copy. The final translation is obtained at a fraction of the cost and time needed for human translation and ensures terminological consistency.

Keywords: Machine Translation, health information

A bit of history

The Spanish-English module of PAHOMTS has been operational at PAHO since 1980. Its English-Spanish counterpart was added in 1985, along with the syntactic analyzer (parser). The software was ported from the mainframe to the PC in 1992 and then to the Windows environment in 2000 (León, 2000). Two new modules, English-Portuguese and Portuguese-English, were added in 2003; the remaining two modules, Spanish-Portuguese and Portuguese-Spanish, were completed in 2004 and 2005, respectively.

The basic architecture of the program hasn't changed since 1985: it includes a lookup component which handles morphology, a grammatical component which identifies the internal structure of each sentence, a transfer component which transforms the syntactic structure into the corresponding target language structure, and a synthesis component which creates the target sentence (Vasconcellos and León, 1988). Dictionary entries are rich in morphological, syntactic, and semantic information. While the basic architecture has remained unchanged, the grammar and dictionaries have grown considerably in the past 20 years, and the software has been adapted over the years for new operating systems and word processing packages.

PAHOMTS is a serious Machine Translation (MT) system, highly respected in the MT community, especially in the health-related disciplines and in its language combinations. In several MT evaluation experiments sponsored by the Advanced Research Projects Agency (ARPA/SISTO) between 1992 and 1994, where 3 research systems and 5 commercial systems were evaluated for adequacy, comprehension, and fluency, PAHOMTS consistently received the highest scores for all three criteria (White and O'Connell, 1994).

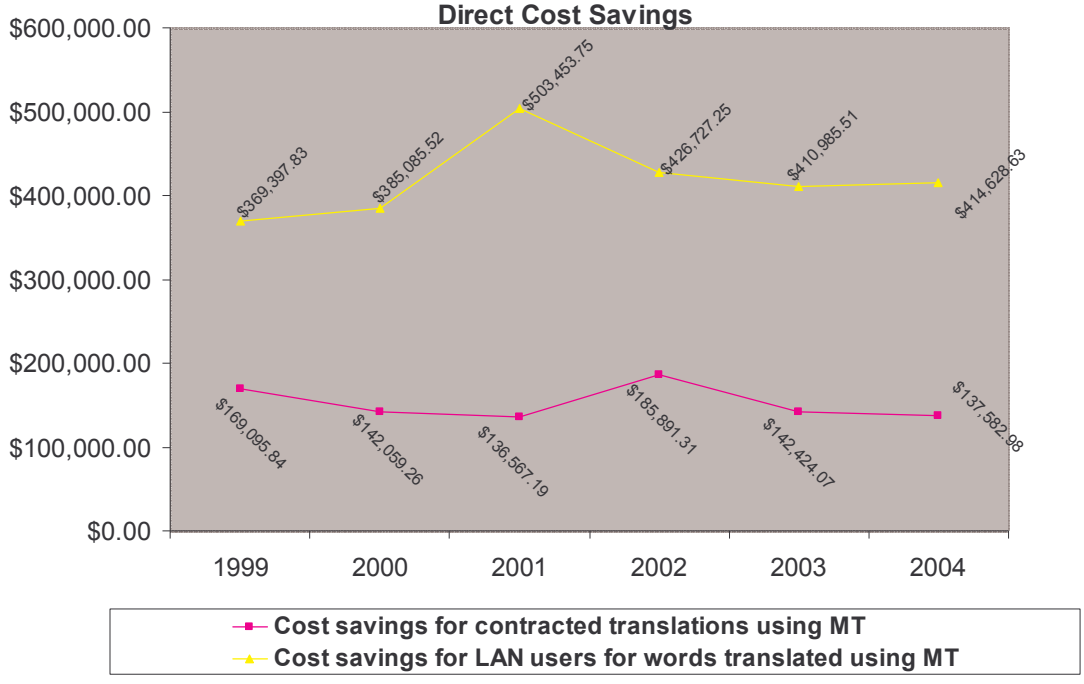
Institutional Setting

One of the reasons for the success of MT at PAHO is the institutional commitment to the MT technology. Early on, PAHO management conducted an experiment (Vasconcellos, 1989) which demonstrated the cost-effectiveness of MT and provided a thumbs up for the daily use of MT in the translation services unit. More than 15 years later, the solid PAHOMTS engine is used in over 90% of all translation jobs for the six language combinations available. Productivity gains are currently estimated at 50%-100%, with a cost reduction of approximately 33%. Translation rates for MT postediting are approximately 70% of the corresponding human translation rates. The translation services unit translates an average of 300,000 words per month. On average, over $150,000 are saved every year by using MT as the primary mode of translation (see figure 1).

PAHO is in a unique position as both a developer and user of MT. This unusual environment allows computational linguists and translators, both staff and free-lancers, to have direct communication and constant feedback, which helps correct mistakes in the output and fosters an ongoing process of software enhancements.

The software is also installed on the PAHO local area network (LAN) and used by HQ staff on a daily basis, translating an average of 200,000 words per month, with an estimated savings of over $400,000 per year (see figure 1). An intranet copy is installed on the PAHO Intranet and can be accessed remotely by any staff member. Full LAN copies are also installed in 24 PAHO/WHO Representative offices in the Americas and in 10 PAHO Centers, including BIREME in Brazil.

Figure 1. PAHOMTS Usage at Washington HQ



Translator productivity is enhanced by the use of PAHOMTS. Thus, whereas a professional translator in an international organization is expected to produce some 2,000 words of translation in an 8-hour work day, our translators have higher productivity rates because they don't need to produce the first draft of the document and don't need to worry about preserving the format. One of our translators was recently able to translate a well-written 17,000-word technical document from Spanish into English in just two working days, plus one extra day for revision. This type of productivity is not typical because the quality of the source document is usually not so high, but it can give an idea of the invaluable assistance PAHOMTS can provide.

Users outside of PAHO

As word of PAHOMTS success spread, PAHO started receiving requests from external institutions and individuals for their own copy of the software. The first licenses of the Windows version were granted in 1993. Since then, 84 licenses have been granted: 11 complimentary copies (WHO, Ministries of Health, Public Health agencies), 15 companies, 9 educational institutions, 16 government agencies, 16 free-lance translators, 8 international organizations, and 9 NGOs.

PAHOMTS is used differently by this diverse body of users. Whereas translators and translation agencies use the software primarily as a tool in their daily translation jobs, other institutions use it primarily for assimilation purposes, for a major translation project, or for daily communication. Some examples include:

- The US Northern Command/SG uses PAHOMTS embedded in an intranet-based Translingual Instant Messaging tool in order to enable communication between English- and Spanish-speaking military health personnel in the USA and Latin America (Jones and Parton, 2004).
- An example of a major translation project is the one carried out by the United States Pharmacopeial Convention (USP). The USP translated its National Formulary (NF) into Spanish in 1994-5 using PAHOMTS. A linguist was hired to customize the dictionaries before undertaking the large translation project. Dictionary updating was performed for a 4-month period, and the translation of the NF, approximately 6 million words, was completed in 10 months. The USP estimated that it would have taken them well over 3 years to translate the NF without MT (Fefer et. al, 1995).
- Several universities use the software with students of translation and computational linguistics.
- Epidemiologists at the Puerto Rico School of Public Health use PAHOMTS to translate technical documents and PowerPoint presentations.

Features of PAHOMTS

PAHOMTS may be installed in three different platforms: standalone PC, Local Area Network, and Intranet server. The software runs under Windows 95 through XP, has a trilingual user interface, and online context-sensitive help. It can translate segments,

using an ActiveX object, or full documents, for which all formatting is preserved. PAHOMTS can translate MS Word, XML, HTML, SGML, PowerPoint, WordPerfect, and text files. The ActiveX object may be embedded in other applications. PAHOMTS may be invoked from the desktop, from an MS Word toolbar, or from a WordPerfect (8.0+) toolbar.

Additionally, PAHOMTS incorporates a batch translation utility to translate a group of files in batch mode and several dictionary utilities:

- dictionary **browse** utility: used to look up existing words, phrases, and rules, it can be accessed from MS Word and incorporates a copy/paste capability.
- dictionary **update** utility: used to add, modify or delete words, expressions, and rules
- dictionary **merge** utility: used by clients to combine their entries with PAHO master entries, and used at PAHO HQ to acquire user entries
- **import** utility: used by clients or PAHO developers to import bilingual glossaries of terms into the PAHOMTS dictionaries. This utility has been used to import the World Bank glossaries (15,000+ entries), International Classification of Diseases, or INN terms (official non-proprietary or generic names of pharmaceutical substances).
- **export** utility: used to create lists of words added by users or small dictionaries that can later be merged

The PAHOMTS parsers generate a syntactic analysis of the source sentence, modify the sentence structure to match the target language grammar, and create a target language sentence. Documents with shorter sentences usually render a higher percentage of complete parses. On average, at least 60% of the sentences are completely parsed, but all sentences are translated, regardless of whether or not they were completely or partially analyzed. Translation accuracy varies depending mainly on the quality of the input document, vocabulary, and sentence length. For narrow subject matters and/or documents with short sentences, translation accuracy can reach 90%.

Translation speed depends on the type of installation, processor, and network traffic but the minimum speed reported is 8,000 words/minute. In batch mode, typical translation speed is 30,000 words/minute.

Several options may be selected at run-time, including grammar type (abstract, letter, manual, report, resolution, survey, speech, post description, news article, summary record), specialized vocabulary, use of accents in uppercase letters, use of separators in numeric expressions, translation of footnotes, etc.

The software also includes an MS Word toolbar to facilitate pre-processing and postediting tasks.

Dictionaries

As of June 2005, the PAHOMTS dictionaries contain the following number of source entries, including single words in uninflected form, multi-word entries, analysis units and translation rules for lexical selection:

English-Spanish	117,951
Spanish-English	113,488
English-Portuguese	102,373
Portuguese-English	95,598
Spanish-Portuguese	94,294
Portuguese-Spanish	85,066

On average, each dictionary contains 101,462 source entries although, evidently, the *older* dictionaries contain more entries than the newer ones. Each source entry has at least one corresponding target entry.

Many of these entries are from the medical domain. In addition, PAHOMTS integrates 13 specialized dictionaries: statistics (epidemiology), radiation, medical research, patient education, environmental health, pharmaceuticals, law, finance, equipment, agriculture, computer science, United Nations, European variety. Two additional microglossaries, initial uppercase and all uppercase, contain special translations related to capitalization. Several microglossaries may be activated for each translation job, in order of priority. Figure 2 below shows some examples of English-Spanish microglossary translations.

Figure 2 – Sample English-Spanish microglossary translations

English	Default	Translation
Patient Education		
headache	cefalea	dolor de cabeza
heartburn	pirosis	acidez
stroke	accidente cerebrovascular	derrame cerebral
itch	prurito	picazón
body fluid	humor orgánico	líquido corporal
Medical Research		
recipient	adjudicatario	receptor
smear	manchar	preparar un frotis
murmur	murmullo	soplo
pupil	alumno	pupila
culture	cultura	cultivo
fault	culpa	defecto
napkin	servilleta	toalla sanitaria
Radiation		
housing	vivienda	cubierta
board	junta	tabla
Equipment		
nail	uña	clavo
board	junta	tabla
stool	heces	taburete
Environmental health		
lead	delantera	plomo
heading	título	espigueo

In addition to microglossary translations, each entry may have up to 99 alternate translations for each part of speech (Verb, Noun, etc.). These alternate translations are triggered by context-sensitive rules, which are stored as dictionary entries. Some examples from English-Portuguese include:

Trigger word	Context word	Features	Default	Alternate
acquire	–	Condition	adquirir	contrair
play	role	–	jogar	desempenhar
play	music	–	jogar	interpretar
play	–	Device	jogar	tocar
delivery	service	–	entrega	prestação
sense	–	Material	sentir	detectar
collection	waste	–	coleção	compilação

These rules test for the direct object. For example, the first rule indicates that, when the object of the Verb *acquire* is a condition or disease, the Portuguese translation for the Verb should be *contrair*, and not *adquirir*. Similarly, when the direct object of the Noun *delivery* is the word *service*, the Portuguese rendering of *delivery* should be *prestação*, not *entrega*. PAHO master dictionaries contain thousands of these rules.

Users may add new words, multi-word expressions, or translation rules and thus customize their dictionaries to their specific needs. They can also create up to 6 microglossaries. Terminologists and linguists at PAHO HQ are constantly adding new dictionary entries to the master dictionaries. When outside users install a software upgrade, they may merge their customized dictionaries with the PAHO master dictionaries.

Information assimilation

Using the translation interface or the Active X objects, PAHOMTS can be used to obtain a quick and inexpensive translation of a document or segment. Since the *raw* translation does not require postediting for assimilation purposes, there is no cost involved beyond the original license fee. Speed can also be a crucial factor in many situations and, with the current translation throughput, speed is not an issue.

An example of a user who employs PAHOMTS for information assimilation is a Hospital in Panama that receives daily health news headlines in English from the InteliHealth Professional Network. The headlines are translated into Spanish with PAHOMTS and reviewed by doctors, with no postediting involved. They can quickly decide which articles are of interest to them and then have them translated with PAHOMTS and postedited by a native speaker, if they decide to distribute the article.

Another potential use of PAHOMTS is in the translation of the Supercourse lectures. This course is "designed to provide an overview on epidemiology and the Internet for medical and health related students around the world" (http://www.pitt.edu/~super1/, La Porte et. al, 1994). The developers are seriously considering the possibility of using MT for the translation of the more than 2,000 lectures already available. This would be an ideal

application of MT because the target audience is already knowledgeable about the subject matter, the PowerPoint presentations are written in simple and direct sentences, and the vocabulary is restricted to the medical domain.

Yet another example is the use of PAHOMTS at the National Library of Medicine (NLM) of the United States. As reported in Rosemblat et. al. (2003), the NLM uses PAHOMTS to translate Spanish queries into English in order to perform cross-language information retrieval on the clinical trials database.

Information dissemination

A document that is to be distributed must be postedited by a professional translator. If the document will be distributed internally or in an unofficial setting, the translation should be revised by, at least, a native speaker of the target language who is familiar with the subject matter, if a professional translator is not available.

MT is not meant as a substitute for translators but rather, as a tool to aid translators in obtaining faster and less expensive translations, as well as to reduce the time spent performing terminological searches. For example, since the PAHOMTS dictionaries contain hundreds of names of organizations, and these official names appear with reliability marks in the *raw* translation, posteditors know that they can trust the translations and don't need to spend time researching the terms.

The MS Word postediting Macros facilitate the posteditor's task by saving keystrokes and automating frequent change operations. Some examples include:

- General macros: changing display, lower/uppercase, switch words to the right or to the left, delete words, look up in PAHOMTS dictionaries
- English macros: delete definite article, create possessive expression, create Noun-Noun compound, undo Noun-Noun compound, add a serial comma
- Spanish/Portuguese macros: singular, plural, masculine, feminine, feminine plural, masculine plural, delete definite article, -mente adverb, add diacritic

PAHOMTS produces a *raw* translation that should be considered a draft document that needs to be postedited. The final translation can be completed much faster than with traditional human translation, at a fraction of the cost. Additionally, terminological consistency is greatly enhanced by using MT.

A clear example of a major project for health information dissemination using PAHOMTS is the translation of the USP-NF mentioned above. Another sample is its use by the Pan American Center for Sanitary Engineering and Environmental Sciences, CEPIS, which used PAHOTMS to translate all abstracts of the full articles in their database. By using the built-in grammar option for abstracts, which changes Spanish active sentences into passive sentences, they achieved very high accuracy in the *raw* translations and were able to translate the abstracts in record time with limited staff.

Combining MT with Translation Memories

Many institutions nowadays are making use of Translation Memory tools for their translation needs. Translation memories are databases that contain aligned source and target sentences (or *segments*). When a new translation job is processed, the system searches in the database of past translations and suggests the previous translation for a segment if a match is found. Matches may apply to complete segments or to subsegments within a sentence. Translation memories work best with repetitive texts and require many hours of work in order to align a significant number of translations.

At PAHO translation services, translation memories are used in conjunction with machine translation for certain types of documents (León, 2002). Staff translators work primarily in interactive mode, as follows: for each segment, when a match is found in the memory, the suggested translation is displayed. The translator may edit the translation as needed and proceed to the next segment. If no match is found in the memory, the segment is processed with PAHOMTS. Again, the posteditor will edit the *raw* translation, which is appended to the translation memory.

Translation memories are also fed with past translations from a 60+ million corpus. The large aligned corpus will be made available to freelance translators on the web so that they may search for terms. The corpus is also being used for automatic terminology extraction in order to improve the PAHOMTS dictionaries.

Software enhancements

PAHOMTS is nowadays a solid MT program because it has been enhanced daily for the past 20 years, and the improvements are based mainly on actual translations produced by the software. Thus, the system's dictionaries are improved daily as follows:

- when a translation is run, PAHOMTS creates a list of not-found words. A dictionary updater adds the words to the dictionaries and, depending on the number and frequency, the translation job is re-run.
- while posteditors are working on a translation, they record suggestions for grammar and dictionary enhancement in an electronic file (the side-by-side file) which contains, for each source sentence, the corresponding *raw* translation and a feedback column. The feedback is later implemented in the system's grammars and dictionaries by the computational linguists and dictionary updaters.
- the computational linguists also analyze side-by-side output files and look for sentences that have not been fully parsed, in an effort to locate and correct gaps in the syntactic parser. The percentage of full parses has been steadily increasing over the years.
- an automated terminology extractor is used to locate expressions that are used consistently in past translations, along with their translations. The computational linguists later review the lists and import them into the PAHOMTS dictionaries using the bilingual glossary import utility.

References

Aymerich, J. (2004) *Machine Translation in Practice at PAHO*. Tutorial presented at AMTA 2004, Washington, DC, September 28, 2004

Jones, S. and Parton, G. (2004) *Collaboration Across the Multinational Battlespace in Support of High-stakes Decision Making - Instant Messaging with Automated Language Translation*. MITRE Technical Paper. http://www.mitre.org/work/tech_papers/tech_papers_04/04_0823/04_0823.pdf

LaPorte RE., Akazawa S., Hellmonds P., Boostrom E., Gamboa C., Gooch T., et al. (1994) *Global public health and the information superhighway*. *BMJ* 1994;308:1651-2. (25 June)

León, M. (2002) *La Traducción Automática y Memoria de Traducción*, in Proceedings of the Primer Congreso Internacional: El español, lengua de traducción, Almagro, Spain. http://europa.eu.int/comm/translation/events/almagro/html/ponencias_es.htm

León, M. (2000) *A New Look for the PAHO MT System*, in J.S. White (Ed.) AMTA 2000, LNAI 1934, pp. 219-222

León, M. and L. Schwartz (1986) *Integrated Development of English-Spanish Machine Translation: from Pilot to Full Operational Capability: Technical Report of Grant DPE-5543-G-SS-3048-00 from the U.S. Agency for International Development*. Washington, DC: Pan American Health Organization

Rosemblat G., Gemoets D., Browne A., Tse T. (2003) *Machine Translation-Supported Cross-Language Information Retrieval for a Consumer Health Resource*. American Medical InformaticsAssociation Annual Symposium (AMIA03)

Fefer, E., Stavchansky, S., Pezoa R., Martianera, J. and Vernengo, M. (1995) *Introduccción a la versión en español de la USP23-NF18*. Internal USP report.

Vasconcellos, M. (1989) *Long-Term Data for an MT Policy*, in Literary and Linguistic Computing, Vol. 4, No.3, pp. 203-213. Oxford University Press

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, pp. 187-236. Cambridge University Press

White, J., and O'Connell, T. (1994) The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. Proceedings of the 1994 Conference, Association for Machine Translation in the Americas.