Automatic Extraction of Entries for a Machine Translation Dictionary Using Bitexts

Julia Aymerich and Hermes Camelo
Pan American Health Organization
525 23rd Street, N.W.
Washington, DC 20037 (USA)
{aymericj, camelohe}@paho.org

Abstract

It is a well-known fact that Machine Translation (MT) systems greatly benefit from user feedback and constant enhancement of the system's dictionaries. However, getting translators to provide feedback in a consistent and timely fashion is difficult given the time constraints under which translators work. Without feedback from translators, the effort spent on postediting is lost for future translations. This paper presents a method recently incorporated in the PAHOMTS® systems to extract truly useful dictionary suggestions using bitexts of past translations in combination with the PAHOMTS® engine, and without any translator involvement. The suggestions are restricted to a few types of common entries and are ranked in order of priority depending on their frequency of occurrence in PAHO's bilingual corpus. While dictionary entries are suggested automatically by the extractor, a human operator must validate the entries. The process of extraction and incorporation of dictionary entries is described in detail.

The setting

PAHOMTS® is a well-established rule-based transfer MT system that has been operational at the Pan American Health Organization (PAHO) for over 25 years (Vasconcellos and León, 1988). It currently translates in six language directions (English-Spanish, Spanish-English, English-Portuguese, Portuguese-English, Spanish-Portuguese, and Portuguese-Spanish) and is used to process over 95% of translation jobs received at PAHO Translation Services Unit. One of the keys of the successful use of MT at PAHO is the fact that users and developers work closely together: translators and other users provide feedback to the computational linguists, who incorporate the suggestions into the MT systems.

This MT feedback has always been provided in context at PAHO. We understand that contextless feedback may be counterproductive for several reasons. First, the translation error may be due to bad input (format codes, typos, incorrectly split sentences), in which case no action is necessary, or to a bad parse, in which case the action may involve manipulations to the parser or fixing certain codes in existing dictionary entries. Second, the meaning of not-found words may be better decided if the context is taken into account. Third, and most importantly, the dictionary suggestion may be contingent on some elements appearing in the immediate context (a string of words, a certain syntactic construction) or in the general context (certain key words in recent paragraphs, certain elements in the document header, a specific subject matter).

The cycle of manual feedback

For each translation job, PAHOMTS® creates a side-by-side (SBS) file which contains the source segment, the aligned MT output, and flags from the parser indicating whether the segment was completely or partially parsed. All feedback is provided using the SBS file, and until very recently all feedback was manually provided by translators. The process has evolved over the years:

- 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.
- The SBS file later became a Microsoft Word document and a feedback column was added to each segment. Translators typed their feedback in this column and e-mailed the feedback file back to the computational linguists.
- The SBS file underwent a third change whereby the MT *raw* output now appears twice in the second column. The second instance is linked to the corresponding segment in the file with the unedited translation. This allows for synchronized postediting and ultimately creates a SBS file with three columns: the source segment, the MT output, and the final translation (see figure 2 below).

Manually provided feedback is first approved by in-house revisers and later analyzed by the computational linguists, who decide which suggestions are doable and whether the change needs to take place at the dictionary level or at the level of the algorithm (format handlers, parser, synthesizer, etc.). The most common types of feedback provided by our translators are: not-found words (NFWs), multi-word expressions (names of programs, institutions, etc), incorrect translations for noun-adjective pairs, noun-noun (NN) compounds in English, noun + preposition + noun (NPN) expressions in Spanish and Portuguese, selection of alternate translations for words depending on the context (nouns, prepositions, verbs, in that order), translation of acronyms, and syntactic or morphological errors.

Manual vs. Automatic feedback

PAHO translators are expected to provide feedback for the MT dictionaries for all translation jobs; while every effort has been made to create a postediting environment that encourages translators to provide as much feedback as possible, the reality is that time constraints often prevent translators from providing in-depth feedback for each document. Indeed, our translators understand the importance of feeding the MT dictionaries but the sheer volume of work and the tight deadlines make it very difficult to devote enough time to provide useful suggestions on a consistent basis. As a result, our computational linguists usually receive SBS files where, on average, 5% or less segments contain some sort of translator feedback. With a translation volume of 4.5 million words per year, computational linguists cannot scan every SBS file in search of good dictionary entries.

Without feedback from translators, the effort spent on postediting is lost for future translations. To overcome this deficiency, we decided to implement a strategy that would compare the MT and final outputs from bitexts created from past translations, select certain constructions using information from the parser and transfer components, and make suggestions for the MT dictionaries. A similar approach in the 1990s (Nishida and Takamatsu, 1990) also made use of linguistic information and reverse MT engines, but it assumed heavy involvement by posteditors. Their system, while valid for a development environment, cannot be expected to become operational in a production environment, where posteditors simply have time to translate and revise, but not to hand-mark and classify all their changes in the postediting process.

Our strategy cannot provide all possible dictionary suggestions, but it can indeed assist the MT developers in our search for large amounts of truly useful dictionary entries, along with their corresponding translations. The expressions in the list of suggestions are checked against an existing bilingual corpus and ranked according to their frequency of appearance in the corpus.

The dictionary entry extraction module currently focuses on the following:

- Not-found words
- Noun + Prepositional Phrase chains in Spanish and Portuguese, and the corresponding NN compound or noun + Prepositional Phrase in English
- NN compounds in English
- Adjectives modifying nouns
- Adverbs modifying verbs
- Verb + direct object pairs
- Subject + verb pairs
- Translation of bound prepositions with verb, noun, or adjective heads
- Insertion or deletion of definite articles

This list is open-ended and we plan to add new constructions later. We chose these particular constructions because they are the most common suggestions provided by our translators and because these entries can be easily and quickly added to the dictionaries. We also decided that it was important for the feedback extractor to suggest the type of dictionary entry: 1) Substitution Unit (entry where the single words are subsumed under the larger entry, with a single translation); 2) Analysis Unit (or AU, entry where the words are parsed separately and each have their own translations; the AU assists the parser with POS resolution and lexical selection); 3) noun-noun AU (a special type of AU used for NN constructions), or 4) Transfer Unit (TU, a lexical selection rule which specifies the conditions that must be met in order to select alternate translations for any of the words participating in the TU). All of these are stored in the PAHOMTS® dictionaries (León and Schwartz, 1986).

Creating the bitexts

The input text for the feedback extractor is a bitext, a document with aligned source and target segments. This bitext is a by-product of the MT process, which undergoes the steps summarized in figure 1 below.

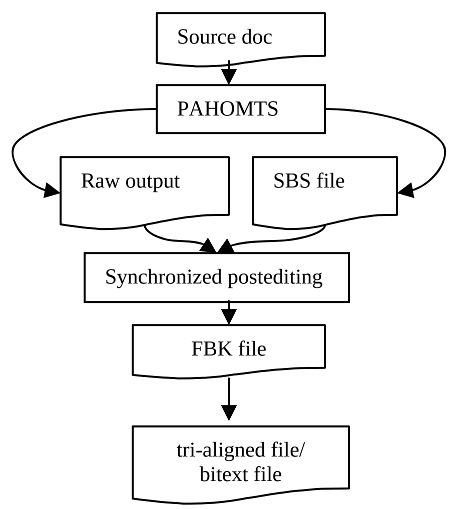


Fig. 1. Creating the bitexts

First, the source document is prepared for MT processing (spell-check, some format checks) and is then run through PAHOMTS®, which generates two documents: the unedited translation (*raw* file, in the same format as the original file) and the SBS file. Both are given to the translator, who then proceeds to postedit the *raw* translation on the screen using the PAHOMTS® postediting macros. In fact, the translator has the option to work on the *raw* or SBS files, which are open in synchronized mode. For each segment, the translator may view the original source segment and may also provide feedback in the feedback column of the SBS file, as seen in Figure 2 below.

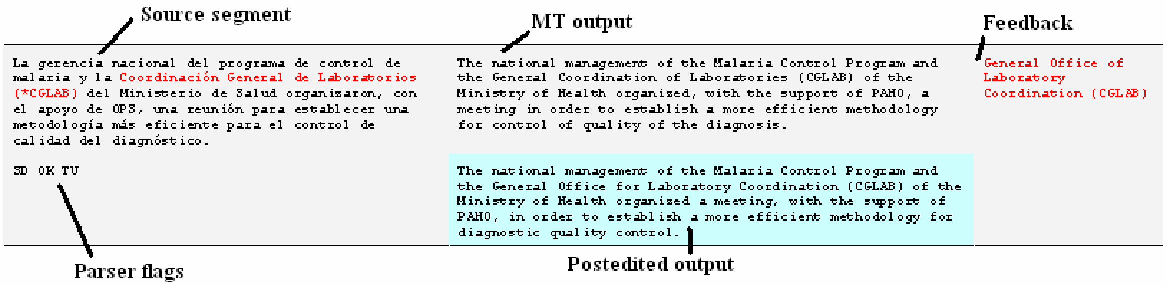
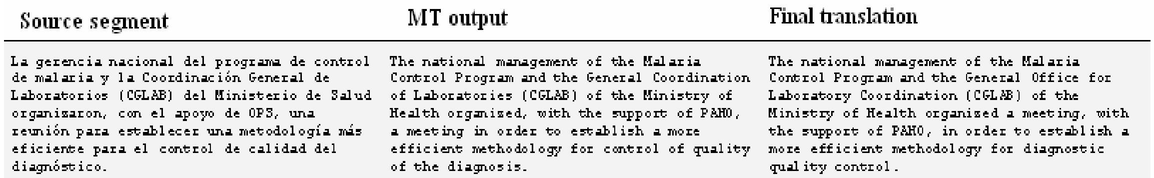


Figure 2 - A row in the feedback file



Source segment	MT output	Final translation
La gerencia nacional del programa de control de malaria y la Coordinación General de Laboratorios (CGLAB) del Ministerio de Salud organizaron, con el apoyo de OPS, una reunión para establecer una metodología más eficiente para el control de calidad del diagnóstico.	The national management of the Malaria Control Program and the General Coordination of Laboratories (CGLAB) of the Ministry of Health organized, with the support of PAHO, a meeting in order to establish a more efficient methodology for control of quality of the diagnosis.	The national management of the Malaria Control Program and the General Office for Laboratory Coordination (CGLAB) of the Ministry of Health organized a meeting, with the support of PAHO, in order to establish a more efficient methodology for diagnostic quality control.

Figure 3 – A row in a tri-column aligned file

The original SBS file thus becomes the feedback (fbk) file. The *fbk* file contains the translator's suggestions on the third column, the MT output in the top part of the second column, and the final output in the blue shaded row of the second column. If the translator chooses to work on the SBS file, he or she will synchronize the final translation in the blue cell by copying it to the corresponding segment in the *raw* file. Conversely, if the translation is postedited using the *raw* file, the final segments will be copied from the *raw* to the *fbk* file. All synchronization tasks are carried out using proprietary PAHOMTS® macros. Once the translation has been completed and the feedback has been approved by in-house revisers, the feedback file is sent to the computational linguists and dictionary coders, who implement the suggestions.

At this point, the assistants run a macro that "cleans up" the *fbk* file by removing all format, parser flags, and feedback, and moving the segments with the final translation into the third column, as shown in Figure 3.

The obvious final step to create the bitext is to simply remove the middle column. Thus, for each translation job processed at PAHO Translation Services, we obtain a document with quasi perfect aligned segments that is input to a translation memory, bilingual corpus, terminology extractor, and dictionary entry extractor.

Extracting dictionary entries

A common approach to extracting dictionary feedback is to use word n-grams and purely statistical methods (Knight and Chander, 1994; Kauchal. and Elkan, 2003; Elming, 2005). In our case, that would involve comparing, using the tri-column aligned file, the MT output and the final translation by performing string matching operations. However, we chose to use the full MT systems at our disposal and to have them guide the process in a more linguistically-oriented fashion, one that makes use of POS resolution, Noun Phrase structure, and parse trees in order to extract linguistically significant strings. This method allows us to suggest expressions that can become dictionary entries, as opposed to random strings that could never exist in a dictionary. Another reason for preferring linguistic techniques is the inability of statistical techniques to handle long-distance dependencies.

The process takes place in four steps:

1. Extraction of <u>source segment</u>, complete parsing, and generation of MT output. Only segments with complete parses (about 65%) are used and partial parses are discarded. As a result of this process, the following data items are created:
   a. Array of source words and associated dictionary entries, including unique identifiers (ID#) for each word
   b. *Vertical path* (string of parts-of-speech) for all words in the segment
   c. Lexical selection rules that have been triggered for all words in the segment
   d. Parse tree
   e. Array of target entries
   f. MT-generated target segment
2. Extraction of <u>translator's final segment</u> and complete parsing. Partial parses are discarded. Data items a-d above are created for the target segment.
3. <u>Identification of constructions</u> in the source segment that are likely dictionary entries, along with identification of the corresponding target string, using the source and target parse trees, the source and target arrays, and the unique ID#s. If the target string cannot be identified with precision, the entire target segment is saved as a suggested translation.
4. For each construction in the list of suggestions, deciding whether the source string is a good candidate for an MT entry. The first step is to check whether the MT output and the final output match.
   a. If they do match, a dictionary suggestion is proposed only in the case of NN compounds for

which no dictionary rule exists. PAHOMTS® generates NN compounds in complete parses and under certain circumstances (Aymerich, 2001), so it might be advisable to add a NN Compound AU to make sure the NN compound is generated correctly regardless of the parse.

b. If the MT and final outputs don't match, the MT dictionary is checked to make sure a lexical selection rule doesn't already exist. If there is no rule, the feedback extractor identifies the correct translation in the final segment and suggests the type of entry to add to the dictionary.
c. If the entry proposed is a not-found word, Substitution Unit or Analysis Unit (both of which apply to contiguous items), its frequency is checked in the bilingual corpus and the frequency score is added to the feedback entry.
d. After the entire document is processed, the frequency of each suggestion in the text is also recorded.

The end result is a file in which each suggested entry contains five fields, all of which are sortable:

- the type of entry (single word, Substitution Unit, Analysis Unit, NN compound AU, Transfer Unit)
- the source string (for single words, SUs, and AUs) or the words participating in the lexical selection rule (for TUs)
- the suggested target string (for single words, SUs, and AUs), or the suggested translations for each word in a lexical selection rule (for TUs). The target string can be the entire target segment if no good match is identified.
- the frequency of the expression in the bilingual corpus. This field is used to sort the feedback entries so that entries with higher frequency appear at the top. The user can also set a frequency threshold below which feedback entries are discarded.
- the frequency of the expression in the source document.

The following sections describe in detail four of the more common types of dictionary entries extracted.

Feedback 1: Complex Noun Phrases

The parse tree is used to locate NPs in the source segment. For Spanish and Portuguese, only NPs that contain non-terminal elements are considered, and these can be Prepositional Phrases or other NPs in coordination. For English, NPs that contain more than one noun (noun-noun compounds) are also considered to be complex NPs.

For each valid NP, the feedback extractor creates a list of candidate NPs, all of which must meet the "complex NP" definition outlined above. For example, for the segment *Área de Vigilancia Sanitaria y Atención de las Enfermedades*, the following NPs would be added to the list of candidates: *Área de Vigilancia Sanitaria y Atención de las Enfermedades* (words 1-9), *Vigilancia Sanitaria y Atención de las Enfermedades* (words 3-9), *Vigilancia Sanitaria y Atención* (words 3-6), *Atención de las Enfermedades* (words 6-9), and *Área de Vigilancia Sanitaria* (words 1-4). The following are examples of NPs which would not be considered for feedback extraction because they cannot yield valid entries: *hábitos higiénicos y nutricionales saludables, the four official languages* (all terminal symbols), *otras áreas que requieren ser legisladas, matters which will be taken up* (NP contains a Relative Clause).

The feedback extractor then processes all NPs in the list, one-by-one. For each element in the NP we have information about the actual word, the word number, the word's unique identifier, the POS resolved by the parser, and head-modifier relations. Moreover, we have this same information for all the words in the target array created after the transfer component, and we must find the correspondence between the two sets.

For example, consider the segments in Table 1 below. One of the NPs extracted for possible feedback is *mejoramiento de los estándares de vivienda*. The NP is extracted by following the chain of NPs and PPs in the parse. The first word in the NP that contains *mejoramiento*, after eliminating determiners, is taken as the first word in the source string. We travel down the parse tree by moving from NP to PP, and from PP to NP. The search stops at the last NP that contains only terminal symbols. The last word in the last NP in the chain (*vivienda*) is taken as the last word in the source string.

The following is some of the information available after the parse tree is converted into a flat table representation:

#	word	ID#	pos	Role	Mod
14	mejoramiento	**015584**	N	Head	
15	de	000100	P		
16	los	000013	D		17
17	estándares	**067837**	N	Head	
18	de	000100	P		
19	vivienda	**018238**	N	Head	

El incremento de la alfabetización y el desarrollo de destrezas vocacionales, el **mejoramiento de los estándares de vivienda**, el desarrollo agrícola con un objetivo básico nutricional, y el crecimiento económico con una distribución equitativa de los beneficios son requisitos fundamentales para el mejoramiento de la salud en los países y en la Región.	The increase in literacy and the development of vocational skills, the **improvement of the housing standards**, the agricultural development with a nutritional basic objective, and the economic growth with an equitable distribution of the benefits are essential requirements for the improvement of health in the countries and in the Region.	For example, greater literacy and the development of vocational skills, **improved housing standards**, agricultural development with nutrition as the primary objective, and economic growth with equitable distribution of its benefits are essential for improving health in both the countries and the Region.

Table 1: Sample source, MT, and final segments

The target array does not contain a translation for the second preposition *de* (which has been deleted by the transfer component), and two of the nouns have been reversed:

#	word	ID#	pos	Role	Mod
14	improvement	**015584**	N	Head	
15	of	000100	P		
16	the	000013	D		18
17	housing	**018238**	N	Mod	18
18	standards	**067837**	N	Head	

By matching the ID#s of all Spanish head nouns (015584, 018238, 067837), we can conclude that words {14-19} in the source string correspond to words {14-18} in the MT output.

The feedback extractor then searches for the MT output segment {*improvement of the housing standards*} in the final translation produced by the translator. In this case, because no match is found, we must parse the translator's version of the sentence, using the converse translation engine (English-Spanish in this case), in order to locate the corresponding English phrase. The English parser produces a parse tree and the POS resolution for the words in the segment. We cannot simply use the unique ID#s because these numbers are not the reverse image of one another in the English-Spanish and Spanish-English translation modules. The parse tree must be used instead.

The feedback extractor traverses the English parse tree, looking for NPs whose head noun is the same as the target for the Spanish head noun (*improvement*) or the last head noun in the Spanish NP (*standards*). We must look for both nouns because we don't know whether the English NP contains a NN compound (*housing standard improvement*), a Prepositional Phrase (*improvement of housing standards*), or some other construction. After traversing the parse tree, we cannot locate a NP whose head is *improvement*, but we can locate one whose head is *standards*. Using the parse tree, we extract the word numbers for the words in this NP (words 12-14), construct the target string (*improved housing standards*), and add it to the feedback file. Because the Spanish string contains a preposition and the English string does not, we determine that the suggested entry should be a NN compound AU (NN1 = AU that deletes prepositions and reverses nouns).

The search for the English phrase is restricted by the number of head nouns: the English construction should have the same or less head nouns than the Spanish. We now have a source string, a target string, and an entry type. All we have to do is check the frequency of the Spanish expression *mejoramiento de los estándares de vivienda* in the Spanish corpus, and add this number to the feedback entry, which looks as follows:

Freq	Type	Spanish	English
5	NN1	mejoramiento de los estándares de vivienda	improved housing standards

Other complex NPs extracted for this sentence are:

Freq	Type	Spanish	English
114	NN1	mejoramiento de la salud	improving health
7	NN1	estándares de vivienda	housing standards
2	NN1	incremento de la alfabetización	greater literacy
1	NN2	distribución equitativa de los beneficios	equitable distribution of its benefits

If some of the words in the source NP are in uppercase, the feedback extractor suggests an SU in addition to an AU. Some examples of suggested SUs are: *Comité de Ética para la Investigación de la OPS* → *PAHO Ethical Review Committee*; *Registro de Proyectos de investigación de la OPS* → *PAHO Research Registry*; *Iniciativa América Libre de Humo de Tabaco* → *Smoke-free Americas Initiative*.

Feedback 2: Noun-Adjective combinations (AUs and TUs)

The search for adjectives modifying nouns is somewhat different. After the sentence is parsed and the flat table representation is created indicating the syntactic relations between words, the feedback extractor looks for adjectives that modify a noun. The noun-adjective pairs do not need to be contiguous because the parser records modifier information for long-distance dependencies. Once the adjective-noun pair is located in the source string, the corresponding translations are located in the MT output by using head-modifier information, POS, and unique identifiers. Next, we parse the translator's final segment, and locate adjective-noun pairs. For each pair, we try to match the translation of the adjective and its head noun. If both match, the entry is discarded. Conversely, if only one matches, the feedback extractor suggests a TU that checks the modifier-head relation between the adjective and the noun. An AU is also suggested if the words are contiguous. If we cannot match the translation of both adjective and noun, the feedback extractor locates all adjective-noun pairs in the final segment; if we have subject and object information for the NPs, it is used in order to filter out irrelevant NPs.

Some examples of adjective-noun feedback are: *contexto favorable* → *favorable climate* instead of *favorable context*; *actores claves* → *key participants* instead of *key actors*; *acatamiento masivo*→ *widespread compliance* instead of *massive compliance*; *tabaco ajeno* → *second-hand smoke* instead of *alien tobacco*; *dudoso honor* → *dubious honor* instead of *doubtful honor*; *censura previa* → *prior censorship* instead of *previous censorship*; *comercialización agropecuaria* → *agricultural production* instead of *livestock marketing*.

Feedback 3: Lexical Selection Rules (TUs)

Transfer Units (TUs), which are stored in the dictionary, are lexical selection rules that indicate the context that must be present in order to select alternate translations for the triggering word or for any of the context words. The context can refer to the lexical entries that appear in the vicinity of the trigger word, or can refer to specific syntactic constructions in the parse tree. The words in the TU do not have to be contiguous because the relations between them are recorded in the parse tree. TUs can refer to syntactic functions (Subject, Object, Complement), Head-Modifier relations (adjectives modifying nouns, adverbs modifying verbs), subcategorization (presence of dependent clause, indirect object, bound preposition, infinitive clause, etc.), coordination, position in the sentence, role in the phrase, etc. The most common types of TUs are Object TUs and TUs that check for the presence of a bound preposition.

Let us look at the extraction of one common type of TU, the Object TU. In order to extract possible Object TUs from a bitext, the feedback extractor needs to use the parse trees for the source and target languages. After the source sentence is parsed, a list is compiled with verb-object pairs. For each pair, the program records the translation selected by PAHOMTS®. Then, the translator's version is parsed, all verb-object pairs are listed, and matches are attempted for each verb-object pair. Those pairs that do not have a match are checked against the list of TUs that applied in the source sentence. Furthermore, if there is Subject information for the verb, we use it in order to filter out irrelevant verb-object pairs. The remaining pairs for which no TU has applied are added to the feedback list. In the case of TUs, the frequency cannot be recorded because the words are not necessarily contiguous. Some sample TU feedback entries are: *permitir acceso* → *allow access* instead of *permit access*; *asegurar calidad* → en*sure quality* instead of *assure quality*; *suplir las funciones* → *assume functions* instead of *replace functions*; *cambiar patrones* → *change habits* instead of *change patterns*; *favorecer mejoramiento* → *improve* instead of *favor improvement*.

Feedback 4: Not-found words

The logic to identify possible translations for not-found words is of course very different from other feedback entries explained above, but it also makes use of the source and target parse trees, capitalization, and POS resolution information.

Whenever a not-found word is encountered in the source segment, the feedback extractor compares the MT output with the final translation and filters out all the words that exist in both versions, unless they are capitalized. Next, the feedback extractor checks whether the parser has resolved the part of speech of the not-found word and eliminates the target words which have been parsed with a different part of speech. Function words and auxiliaries are also removed. The remaining words in the final translation are looked up in the target dictionary for the original language direction. If there is only one word that is not found, it is suggested as the translation for the not-found word in the source segment. If there are two or more, their corresponding source entries are looked up in the source string, and those not found are saved as possible translations for the not-found word.

Incorporating the feedback in the MT dictionaries

Once the feedback file is created, the interface displays all feedback entries in 6 columns: segment in which the string occurs, frequency in the corpus, frequency in the document, suggested source string, suggested target string, and suggested dictionary entry type. All of these fields are sortable, so the user can choose to work first on all not-found words, all modifier TUs, all NN compounds, etc. The user can also eliminate useless entries or entries with very low frequency in the corpus or in the document at hand.

When the user clicks on any given suggestion, the interface displays the complete source sentence, the complete MT output, and the complete translator's version of the sentence, so that the user always has access to the context. The segments suggested by the feedback extractor are clearly identified with color codes within the complete sentences. If the source or target segments suggested by the feedback extractor are not correct, the user can edit the text before adding the entry to the dictionaries.

Once the user is satisfied with the source and target strings and the type of entry, he or she simply clicks on the *Add entry* button in the interface. This opens the PAHOMTS® dictionary update interface, which will have started the creation of the new entry, following the instructions from the feedback extractor (single word, SU, modifier TU, noun-noun AU, etc.) The user can modify any fields as necessary before saving the entry.

It is important to stress that, while the process of extracting possible dictionary entries is automatic, the suggestions are always validated and added to the dictionaries by a human. When the extractor suggests several possible translations for a source entry, for example, the human will decide which is valid and will simply discard the rest. Similarly, the user will use his or her judgement as to whether a context rule suggested by the extractor is valid in all contexts or is an ad hoc translation in a particular document and should not be added as a dictionary entry. By manually validating suggested entries, we make sure that no superfluous or even harming entries are being added to the dictionaries.

Some preliminary results

The feedback extractor described in this paper was developed at PAHO in the first three months of 2007 and has been tested on the Spanish-English bitexts created during this period, comprising a total of 237,470 words (124,440 Spanish source words and 113,030 English target words).

Dictionary feedback provided by translators for these texts yielded 175 dictionary entries. The dictionary suggestions extracted by the new application yielded 4,409 dictionary entries. The process of checking and adding the entries was quite fast, because all the context was provided and the extractor assists the dictionary update program by suggesting the type of entry.

In order to get more fine-grained evaluation results, we selected a small subset of documents (16 short documents, 219 segments), ran them through the extractor, added all valid entries to the dictionaries, and manually identified the entries that were not suggested by the program. On average, the extractor was able to identify 65% of all valid suggested entries (recall). The missing 35% was due to incomplete or incorrect parses. Of all the entries suggested by the extractor, only 26% actually became dictionary entries. This low precision is mainly due to two factors: the extractor makes several suggestions for an entry when unable to identify the target with accuracy and some suggestions are irrelevant because of their low frequency.

Rerunning the 16 documents through PAHOMTS after doing the dictionary work yielded the following improvements in BLUE, NIST, and METEOR scores:

	First run	Second run
BLEU	0.3760	0.4329
NIST	7.2855	7.8315
METEOR	0.6469	0.6813

Table 2: Automatic metric scores

Conclusions and future work

Although it became obvious during the testing phase that the feedback extractor needs some fine-tuning (to weed out some irrelevant suggestions and to suggest smarter translations in some cases), the process has proven extremely useful for fast acquisition of truly useful dictionary entries. We are confident that, after further testing, this will become an invaluable tool in our work environment.

It is worth noting that the evaluation results, positive as they are, are not as significant in a production environment as the fact that we are now able to get large numbers of useful dictionary entries with minimal effort and with no translator involvement.

Several areas where we will concentrate our efforts in the coming months are: better handling of one-to-many and many-to-one alignments, addition of more constructions, possible use of existing semantic features to improve the accuracy of suggested translations, better handling of constructions that occur in parentheses, incorporation of smarter morphology for string matching, and better handling of constructions involving conjoined elements. Once all of these enhancements have been incorporated and tested, we will use the feedback extractor tool on all bitexts created for all language combinations and expect to see a major improvement in translation quality based on the large lexical acquisition process.

References

Aymerich, J. (2001) Generation of Noun-Noun Compounds in the Spanish-English Machine Translation System SPANAM. In Proceedings of VIII MT Summit, Santiago de Compostela, Spain, September 2001, pp. 33-37.

Charoenpornsawat, P., Sornlertlamvanich, V., and Charoenporn, T. (2002). Improving translation quality of rule-based machine translation. In Coling-02 on Machine Translation in Asia - Volume 16 International Conference On Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, 1-6.

Elming, J. (2005) Using Machine Learning to improve Rule-Based Machine Translation. NGSLT Machine Learning: Term paper, Dpt. of Computational Linguistics, Copenhagen Business School. December 2005.

Kauchal. D. and Elkan, C. (2003) Learning Rules to Improve a Machine Translation System. In Lavrac, N., Gamberger, D., Todorovski, L. and Blockeel, H. (Eds.), Machine Learning: ECML 2003, 14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia, September 22-26, 2003, Proceedings, Lecture Notes in Computer Science, Springer-Verlag (pp. 205-216)

Knight, K. and Chander, I. (1994). Automated Postediting of Documents. In Proceedings of National Conference on Artificial Intelligence (AAAI).

León, M. and Schwartz, L. (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

Nishida, F. and Takamatsu, S. (1990) Automated Procedures for the Improvement of a Machine Translation System by Feedback from Postediting. Machine Translation 5 (3), pages 223-246.

Vasconcellos, M. and León, M. (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