CHALLENGES AT THE WORLD TRADE ORGANIZATION

Evaluation and implementation of a Statistical Machine Translation System

Olivier Pasteur, World Trade Organization, Switzerland

This paper addresses the ongoing challenges of integrating statistical machine translation (SMT) into the translation workflow of an organization such as the WTO, in which significant requirements for technical correctness must be balanced against those for political sensitivity, while dealing with budget constraints, complex workflows, tight deadlines, and high quality requirements. The integration of machine translation necessarily entails fundamental changes in the work of translators. Strategies to help translators accept these changes as well as training for those who are not computer-oriented have to be planned and provided for.

Background

The World Trade Organization is a member-driven organization, where decisions are taken by consensus by its 153 Members. In short, it is a forum to negotiate agreements regulating world trade and a legal framework to implement and monitor existing agreements. It handles trade disputes, monitors national trade policies and provides technical assistance and training for developing countries. About 650 staff members from 79 countries work in the Organization's single headquarters in Geneva. About 6,000 meetings are organised per year, for which documents are prepared and distributed in the three official languages (English, French and Spanish).

The WTO secretariat is organized into 22 divisions, dealing with all technical aspects of world trade. Most of these divisions generate work for the translators, some more than others.

The Languages, Documentation and Information Management Division manages the entire document production chain. It employs 160 staff members.

The translation services translate about 150,000 pages a year, 110,000 internally and 40,000 externally. About 10,000 documents are produced every year in the three official languages. As of now, more than 400,000 documents are available for downloading in English, French and Spanish.

The English Translation Section has 4 permanent staff members, with 1 to 3 in-house freelance translators at any given time. The French and Spanish sections have 20 permanent staff members each, with 5 or more in-house freelance translators at any given time and 20 translators working from home on special service contracts. Approximately 25 % of work is outsourced to external translators.

Working methods vary from the traditional dictation to the most sophisticated computer-assisted translation tools, depending on the type of document or the computer background of the translator.

Translation support services are provided at all stages of the translation cycle. They aim to save translators and revisers time by finding sources, terminology and passages that have already been translated. Tasks range from referencing and pre-translating documents to developing computer-assisted translation (CAT) tools and providing relevant training and support.

The Translation Support Section acts as the interface between the translators and computer experts and cooperates closely with the Informatics Division. Three teams share the responsibility for translation support.

The CAT team, composed of a small group of experts (two computer-oriented translators and one expert in computer tools), provides high value-added services such as pre-translation of documents, management of translation memories, administration of terminology databases and development of computer applications that help translators in their work.

The Support team (one and a half persons) acts as a hotline for translators (in-house and external), installs and updates the translation applications and provides training for users.

The Referencing team, whose 5 members have a thorough knowledge of WTO documentation, serves about 80 translators by providing them all the necessary background documentation. For each document handed out for translation, the team typically provides a list of relevant references and, when appropriate, a comparison between the successive versions of the document to be translated. Using computer-assisted referencing tools, the team provides, for documents with occasional quotes from other documents, an electronic copy of the original with all

passages occurring in other documents highlighted and hyperlinked to the sources in all languages. For documents that have been revised or altered, or that are largely lifted from other documents (new version, etc.), all deleted, added and displaced portions of text are marked in different colours and styles. Translators can thus focus only on new translation and use the recycled translation provided for them.

Translation memories are systematically used for certain types of documents and for documents in which more than 50% is lifted from previously translated documents.

Other tools are also used by translators and support staff. These include off-the-shelf terminology database management and term extraction systems, an in-house developed trilingual concordancer (i.e. a full-text search engine which displays the hit documents in the requested source and target languages and automatically aligns them), a legacy documentation search and retrieval system, a translation resources portal, etc. Most of these tools are available to the general public and may thus be used by external translators.

Because of the wide range of subjects it covers and taking into account the sensitivity of certain documents and negotiations and the tight deadlines, translators consider the WTO a challenging organization to work for. But the work is generally considered rewarding because of the variety and the relevance of subjects covered, the challenging nature of texts that are often in the spotlight, the good working environment and the excellent tools made available to them.

So what would be the added-value of an automatic translation system in this already highly-computerised translation environment? And what happens when an international organization such as the World Trade Organization decides to go ahead with machine translation?

Why MT? Why SMT?

Whereas computer-assisted translation tools have long been used at the WTO, machine translation had until recently never moved beyond testing. However, at the WTO as in other organizations, efforts are being made to streamline work processes and make the most of translation tools available on the market.

The challenge was, and still is, can machine translation help a translator work faster while maintaining the same level of quality? In other words, can an automated translation system produce an output of usable quality, that is, a draft translation worth editing?

With this in mind, several tests were carried out in 2009 by the Translation Support Section on rule-based machine translation systems. But these did not produced the expected results and proved to be time– and resource–consuming. However, the latest developments in the field of statistical machine translation showed that this approach to automated translation in the context of the WTO translation services was more realistic and less costly. Unlike the rule-based approaches to MT, the statistical approach could take advantage of the millions of sentences already translated by WTO translators without having to build huge specialized dictionaries and specific rules, thus reducing overhead costs.

Therefore, in April 2010, the WTO invited potential partners to submit tenders outlining the approach and mechanisms by which they would support the development and implementation of its business requirements.

Scope of the project and business requirements

The project was to be carried out in two phases: a pilot phase, based on a limited scope of the WTO corpus and a limited number of language pairs, to assess the viability of the SMT approach and the ability of the implemented system to produce MT outputs of usable quality; and a production phase to provide a full-fledged statistical machine translation system, based on the entire WTO corpus, delivering MT outputs in the six language pairs in use at the WTO.

It was understood that the production phase would be launched under the condition that the system, as implemented and fine-tuned during the pilot phase, would be able to produce MT outputs of usable quality and to handle growing amounts of data and requests.

The company Simple Shift was selected among the potential suppliers to whom the Request for Proposal was sent. Simple Shift is a Geneva-based IT consulting and service company with a strong expertise in language engineering and multilingual information retrieval.

Project Management and Implementation Methodology

It was agreed with Simple Shift to use the open source statistical machine translation software Moses as the central platform for the WTO's SMT system (SMTS). Moreover, it was agreed that additional in-house software developed by Simple Shift (in particular the phrase segmenter and aligner) would be provided to the WTO, including the source code so as to guarantee the organization's autonomy.

The segmenter in the SMT system installed at the WTO is purely statistical. Thus it needs to be trained on bilingual corpuses. The Canadian Hansard corpus was used for the English-French pair and the Europarl parallel corpus for the English-Spanish pair. As with all training-based systems, the difficulty was essentially to provide the computer with enough examples of all the possible sentence endings, as well as false sentence endings such as abbreviations and other symbols which include, for example, a full-stop mark.

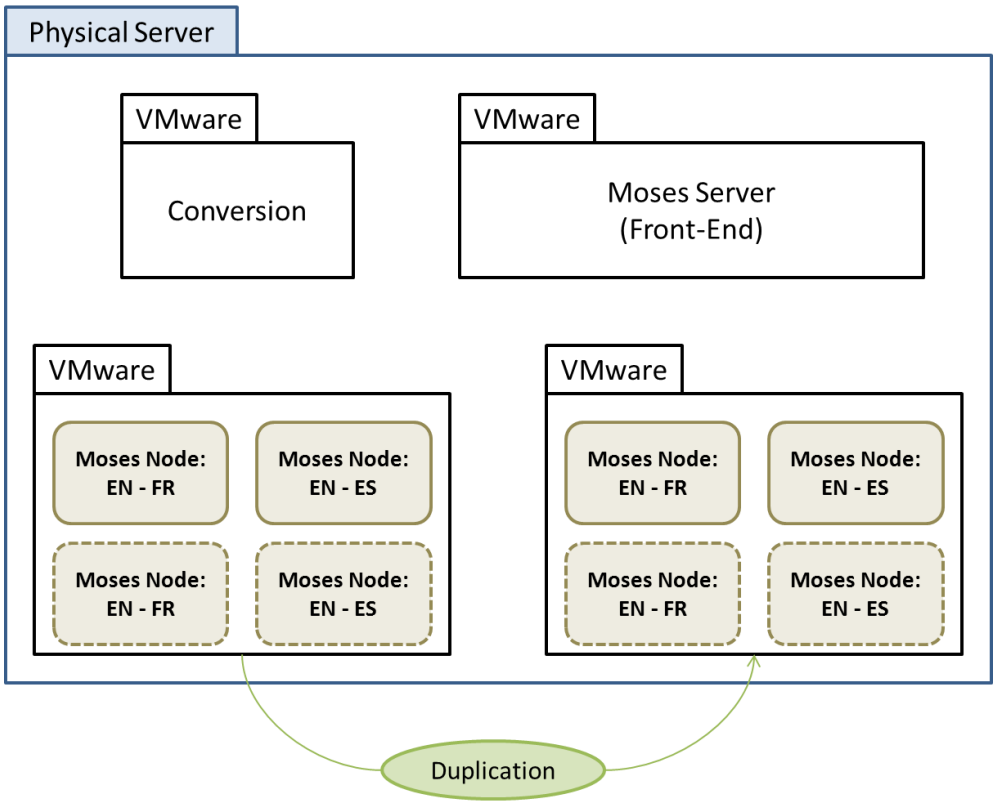
Just like the segmenter, the aligner is trained on a bilingual corpus. The aligner installed at the WTO is also trained on the Canadian Hansard and the Europarl corpus. The aligner chooses, for each source sentence, the target sentence which is the most probable translation. The system is provided with an ordered list of source sentences and a non-ordered list of target sentences. The most frequent words (articles, prepositions, etc.), which are defined in a so-called "stopword list", are removed in order to limit the index size, and thus reduce the number of possible translation options and increase the system's performance. When the system analyzes the first word of the first source sentence, it looks up in its training material to decide which target word is the best possible translation More precisely, it looks for the source word in its training material, and then it looks at all the words in all the corresponding target sentences and calculates the probability, for each target word, that it could be the correct translation of the source word. The target word with the highest probability is chosen. Additional processes are also performed: in case of an alignment "hole" (i.e. the space separating two blocks of well-aligned sentences), the program stops and removes the first and last sentences of the well-aligned block in order to keep only the sentence pairs for which the alignment is certain. The aligner also performs a geometrical calculation of how realistic the alignment is, on the basis of the length of the source and target sentences.

The result of the segmentation and alignment processes is the production of translation memories in TMX format. The content of these translation memories includes the source word or phrase, the target word or phrase and the corresponding translation probability (*e.g.* car ||| voiture ||| 0.8). On the basis of these TMX files, the program randomly extracts a number of source and target sentences (for example 2,000 sentences) and produces 3 files: a training file composed of all the translation memory except the 2,000 sentences; a tuning file containing 1,000 source and target sentences; a test file containing the other 1,000 source and target sentences. The result of this process is the creation of six files: the three files described above, respectively for the source and for the target languages. In this process, all the sentences which have no translation or which are entirely in capital letters are removed, as well as all sentences containing more than 40 words.

Lowercasing and tokenization (*i.e.* separating each word from possible noise such as punctuation signs and apostrophes), language and translation models building, compression and tuning are then performed with the relevant modules included in the Moses application.

The Moses environment

The SMTS architecture relies on a single physical server hosting three to four virtual machines: one for the converter and node (translation system) builder; one to allow users to access and use the application; one or if possible two (depending on the performances of the physical server) to host the Moses nodes. A specific SMTS should be built and installed for each language pair and direction. Each node is encapsulated in a virtual machine (under VMware); each virtual machine includes two to four nodes, depending on the server's performances. Each node is dedicated to a language pair and direction and has its own IP address and port number. The SMTS is a full web application, requiring only a browser from the client computer.



Input and output formats

Two input and output formats will be available in the production version of the SMTS:

- TXT: Input and output files will be in plain text format. This will in particular be the case when text is directly entered or copied/pasted in the web-based graphic user interface of the SMT system;
- XLIFF: The input file will include the source segments, which will contain the sentences to be translated, and the target segments, which will be empty at the initial stage. Tags relating to the text layout (bold, italics, etc.) will not be kept in the translated segments. The output file will then be an XLIFF file containing the source segments of the original documents as well as the target segments as translated by the SMTS. This target file can then be opened with an XLIFF-supported bilingual editor (such as Trados Studio) for post-editing purposes.

Methodology used at the WTO for output evaluation

One must keep in mind that the central issue in evaluating any type of translation is that several outputs can be considered acceptable. Moreover, evaluators often do not agree on what a correct translation is. In order to avoid "reinventing the wheel", an extensive review of the literature on this topic was made. Then a list of evaluation criteria which the human translators could use was proposed and various evaluation metrics which seemed to fit the WTO's working context and project objectives were set: a statistical metric, which is rather objective but does not reflect the direct usability of the SMT output, and two human metrics, which are more subjective but were crucial in determining whether the SMT pilot project could be deployed into production.

Quantitative metric: using the BLEU score

To sum up, BLEU (Bilingual Evaluation Understudy) is a statistical metric of the quality of a machine translation which is calculated by comparing three elements: a series of source sentences, their translation performed by human translators (called "reference translation"), and their translation performed by a computer. The calculation is based on the "precision" metric.

BLEU has several significant advantages: it correlates well with human judgment, it calculates quickly, it can deal with several reference translations for a given source sentence and there is an extensive benchmark of BLEU scores on the Internet.

However, the BLEU scores are criticized on several points: the intelligibility or grammatical correctness is not taken into account; all words in a sentence have similar weight, whatever their frequency and semantic importance; it cannot deal with languages where words have no boundaries (although this is not an issue with the WTO's official languages).

Moses comes with a built-in BLEU score calculation module. This BLEU module can be used to calculate the BLEU score of any other MT system. It was considered useful to build a customized benchmark on WTO documents. Therefore, the BLEU score was chosen as the quantitative metric of the SMTS project to facilitate both the calculation processes and the comparison with the performance of Moses on other corpuses, or with the performance of another SMT system (typically Google Translate) on WTO corpuses.

Qualitative approach: proposed methodologies for an evaluation by human translators

The stated goal of the SMTS pilot project at the WTO is to assess whether the quality of the SMTS output is high enough for a human translator, or reviser, to post-edit instead of performing a full translation from scratch.

From this point of view, the central quality criterion should probably be the time needed to post-edit a given sentence from the SMTS output. This element is more critical than the gravity of the errors, because all errors (minor and major) will have to be corrected for the document to meet the WTO's quality standards, and it often takes about the same time to correct major and minor errors.

However, an evaluation process may only lead to three possible types of conclusion: the SMTS output is good enough to be post-edited; it is so bad that it will be impossible to improve it until it reaches the minimum quality level required; it could be used if improved to some extent.

A quality evaluation which would take into account only the time spent to post-edit SMTS sentences would make the second or third types of conclusion impossible, since there would be no information as to which problems are the most acute. This information, however, is necessary to conduct experiments to decide whether the system could be improved to some extent.

Therefore a two-fold approach was suggested. First, a simple methodology was proposed, based on the time spent to correct each sentence. If this first evaluation results in the conclusion that the SMTS output is worth editing, the evaluation process can be terminated. However, if the result indicates that the SMTS output requires too much correction time, or that almost none of the sentences could be kept for post-editing, a second round of analysis would be required to determine in more detail the types of errors met in the output. This in turn would help to establish priorities in the improvement effort. This second round of output evaluation should thus be based on a more detailed methodology.

<u>The quantitative evaluation</u>

The first evaluation sheet is based on the correction effort (and thus the time) required for each SMTS output sentence. The example below is built on a short series of 10 sentences extracted from a WTO document, as translated by the SMTS during the pilot project.

	Number of Sentences	In %	Score per Sentence	Total Score	Comment
Number of sentences which did not require any correction	1	10.0%	5	5	
Number of sentences which could quickly be corrected (1 or 2 errors)	4	40.0%	4	16	
Number of sentences with more correction work (3 or more errors)	3	30.0%	2	6	
Number of sentences which were not worth correcting	2	20.0%	1	2	
TOTAL	**10**			**29**	

Table 1 – Simple Metric Scoring Sheet

On the basis of the four analysis criteria above (left column), which address individual sentences, the following score may be calculated for the whole document:

Total Score	29
Maximum Possible Score (= Total Number of Sentences * Maximum Score)	50
Overall Document Score (= Total Score / Total Maximum Possible Score)	58%

Table 2 – Simple Score for the Overall Document

The document time score is a percentage of how useful the document is as compared to a document which would have required no correction at all. Translators should then define the percentage threshold above which an SMTS output is deemed useful.

The qualitative evaluation

The detailed evaluation sheet used for this evaluation was adapted from the score sheet of the SAE J2450 metric[1]. Here again, the table below was built on the same series of 10 sentences extracted from the same output.

Score 2 : Error Types	Minor Errors			Serious Errors			Category Weighted Score
	Number of Minor Errors	Minor Error Weight	Total: Minor Error Score	Number of Serious Errors	Serious Error Weight	Total: Serious Error Score	
Wrong Term or Meaning (WT)	0	2	0	7	5	35	35
Wrong Syntax (WS)	4	2	8	7	5	35	43
Omission (OM)	0	2	0	1	4	4	4
Structural Error (SE)	3	2	6	4	4	16	22
Misspelling (SP)	0	1	0	0	3	0	0
Punctuation or Parenthesis Error (PE)	1	1	1	1	2	2	3
Miscellaneous Error (ME)	0	1	0	0	3	0	0
TOTAL	**8**		**15**	**20**		**92**	**107**

Table 3 – Detailed Metric Scoring Sheet

After each sentence has been analyzed based on the criteria in the left column above, the total score is to be divided by the total number of words in the source text in order to calculate the detailed score for the overall document:

Sum of Category Weighted Scores	107
Number of Words in Source Text	180
Overall Document Score	59.4%

Table 4 – Detailed Score for the Overall Document

It should be emphasised that when using the detailed metric scoring sheet, each word should only produce one score, even if it could be affected by several criteria. For example, if one source word is translated by the wrong term, and that wrong term additionally includes a syntactic error (e.g. wrong gender or number), the evaluator should choose which error is the most important, and only score that word under the most important criterion. This is because the total number of errors is then compared to the total number of words, so adding several errors to a single word would introduce a bias in the total document score. Evaluators were also asked to follow two meta-rules to classify errors: 1) When an error is ambiguous, always choose the earliest primary category; 2) When in doubt, always choose "serious" over "minor".

The evaluation process

The pilot phase of the SMTS project in WTO lasted 3 months (July-September 2010). As mentioned earlier, the goal of this pilot phase was to assess the technical viability of the project (implementation of the SMT system, testing on a limited corpus and evaluation of results).

The suggested two-fold approach was followed during the evaluation process. A simple evaluation, carried out by a panel of three translators (one per language), allowed for a rapid calculation of the percentage of reusable sentences in a real-life environment (translation by Moses of a large document in the same field as the translation model built during the pilot phase).

[1] SAE J2450 is a "Quality Metric for Language Translation of Service Information". It was set up by a Task Force "in an effort to establish a translation metric that could be used by automotive companies to compare quality of translation deliverables". It is only one element, albeit important, in a total Quality Assurance process.

Results of the simple evaluation metric

	EN → FR	EN → ES	FR → EN	ES → EN
Document translated	WT/TPR/S/235	WT/TPR/S/235	WT/TPR/S/236	WT/TPR/S/234
Number of sentences evaluated	1204	351	579	99
Percentage of reusable sentences	**47,66 %**	**52,33 %**	**55,00 %**	**58,00 %**

A more detailed evaluation, performed by the translation support staff, pinpointed the most frequent types of errors and helped to establish priorities in the improvement effort. Due to lack of time and resources, this evaluation was limited to two language pairs (EN → FR and EN → ES). The evaluators tried to differentiate the various types of errors and to classify them by broad typology: wrong term or meaning, wrong syntax, omission, structural error, misspelling, punctuation or parenthesis error, miscellaneous errors. However, in the course of the evaluation, translators found that the errors spotted were more varied in nature than expected and that it was difficult to weigh their individual significance.

On one hand, a structural error might be considered a critical problem by a computer expert since it could reveal a problem in the sentence structure training algorithm while a translator might consider it a minor problem if all that is needed is to swap a subject and a complement.

On the other hand, the numerous little syntax errors, which are not considered serious from the computer expert's point of view because they might only be a sign of an inadequate training corpus (which can always be completed and improved), will certainly annoy – if not infuriate – translators, who will have to correct a large number of "small" errors, whether they are serious or not. This will in turn compromise both the acceptability and the efficiency of the tool.

Towards a full-fledged statistical machine translation system at the WTO

Given the results obtained by the "quantitative" evaluation (the percentage of reusable sentences) and the planned improvement efforts to deal with the main errors spotted by the "qualitative" evaluation, it was considered worth starting the production phase and implementing a full-fledged statistical machine translation system at the WTO.

A pre-production version (without most of the user interfaces) will be ready by December 2010. This version will allow us to test all the functionalities of the production version (including the XLIFF support) but without GUIs. The initial development effort can thus focus on translation quality and integration work. The production version with all features and interfaces should be ready by February 2011.

The SMTS will then have become an additional tool in the set of existing CAT tools made available to WTO translators. A tool that could later be fully integrated within a single interface, combining the strength of machine translation, translation memories, multilingual text concordancer and terminology retrieval.