SIMPLE SHIFT
Ruelle du P'tit-Gris 1
1228 Plan-les-Ouates
Switzerland
Tel: +41 (0) 79 419 45 98
Email: karim@simple-shift.com
Website: www.simple-shift.com

ASLIB 2011

Building Blocks to Integrate a Moses MT Engine into the ProductionTranslation Workflow

with reference to the World Trade Organization and other customers

Introduction

The workflow of a translation servicemay be defined as the model (static view) or processes (dynamic view) according to which documents to be translated reach the translation service,are affected to a translator and then possibly a reviser, the translation is produced, and the translated document is pushed out towards the customer, the editing service, the web/paper publishing service, etc.

The benefit of a well-designed and implemented translation workflow can be shown at several levels:

- For translators, it facilitates the reference search operations and the feedback into the reference repositories, and it accelerates the internal flows (from customer to translation service Head to translator to reviser, and back to customer or out to others);
- For translation customers, it simplifies and accelerates the translation request procedure and it gives a better visibility on the translation status and deadline compliance; and
- For content managers, it streamlines the document production processes, shortens the time to publication and improves content security.

If we zoom out, the translation workflow may be integrated into a larger one which can be department- or organization-wide, and is generally based on a content management system (CMS). The interaction between the CMS and translation workflows most frequently occur at the level of documents and their life cycle, and of users and roles (i.e. security rights – this has an impact on how an SMT system can be used by external translators, see below).

If we zoom in, the translation workflow shows sub-processes, especially during the actual translation process, as the translator uses computer-assisted translation (CAT) tools, which in turn are fed with new translations. In organizations where a referencing service is available, that service also uses CAT tools, statistics and other support tools to facilitate the translator's work.

Version	Date	Author(s)	Status
1.0	12 October 2011	Jacques Guyot	Final

In this article we will zoom infurther down to the level where a translator or referencing service uses an SMT tool to support the translation operations. We will see what such a tool can do for them and how it may be integrated into the translation workflow.

We will then describe how such an integration was actually performed at the translation service of the World Trade Organization (WTO), and how it is considered to be done for the translation services of the European Organization for Nuclear Research (CERN) and the Universal Postal Union (UPU).

1. Overview: What can an SMT system like Moses do for an international organization ?

There are two main categories of tasks an SMT system can perform in an international organization: Either it performs a preliminary translation ("pre-translation"), and this can only be done for a translation service, or it produces a final version, and this can only be for internal use of non-critical documents, except if the average output quality is deemed high enough.

a. Pre-translating for the Translation Service, for human post-edition

This is the most reasonable use of an SMT system, since in most cases the output quality is not good enough to produce the final version of an official translation. The SMT is either used to translate an entire document from scratch, or used as a complementary tool to a Translation Memory (TM) system, where it will translate the sentences which could not be found by the TM tool (see item c. below).

This type of use requires integrating the SMT tool in the workflow of the translation service, becausein that case the SMT is strictly considered as a CAT tool and thus is only used by translators.

b. Translating for other departments (no post-editing)

In some situations, documents such as forms, notices, etc. may include essentially controlled-vocabulary. This may allow for a near-final or final version to be produced by the SMT, to be quickly proof-read. This is for instance the case of Vacancy Notices, as illustrated below in the CERN example.

The SMT tool can also beintegrated in a workflow outside the translation service to provide so-called "gisting" services, *i.e.* it provides an instant but rough translation of any document to help the user understand the content. In that case the SMT solution appears as a Google-like translation service, for example on the Intranet home page, and can be called by a mouse click. Needless to say, such a gisting service must be limited to the translation of informal messages, non-critical documents, etc. for internal use.

c. Feeding the Translation Memory system with automatically-segmented and aligned segments

In translation services where a TM tool is available, many translators start by submitting the document to the TM analysis and use a large part of the TM suggestions. However, in this scenario there is no help to translate the sentences which are unknown to the TM tool: either the TM provides a suggestion or the cell remains blank. In this situation the SMT system can be automatically called up by the TM tool to propose a translation of the unknown sentences.

If used in conjunction with an TM tool, the SMT system must be adapted in terms of input and output formats (see the WTO and CERN examples) and an API must be developed to allow the TM system to run the SMT engine (the final results being presented in the TM environment).

d. Aligning full documents for bitext-based CAT tools

Some CAT tools do not use translation memories, but full-text monolingual documents used in source/target pairs which are displayed side by side and automatically aligned. Those so-called “bitext-based” tools are very useful because unlike the translation memories, they provide the source and target sentences in their complete context and allow translators to browse interesting reference documents to find more terminology and contextual information.

The problem with bitext-based tools is they have to align pairs of full-text documents. Various technologies may be used to do so, but one recent approach is to build so-called “alignment maps” which help find the relevant part in the target document. Alignment maps are built with the help of an SMT engine. The challenge for this integration is rather linked to the update process: each day new documents are added to the reference repositories so new maps must be built and stored. The map builder is integrated with the bitext-based CAT tool, which is itself integrated into the translation workflow.

e. Supporting search engines for cross-language search operations

A common issue in international organizations, where all users speak at least two languages, is to choose a language to perform a search operation. Some tend to search in their mother tongue first, and if there is no relevant result they start the search over in English. Others search in English first, and if they have trouble understanding the results (especially in a technical context), or if the result is not relevant enough (in case of a country-specific search, for example) they repeat the search in another language.

Thus there is a need for a “search expander” or “cross-language search tool” which would expand an initial request (whatever its language) into several others – and possibly translate back hit documents in the request language. For example, if someone enters a English request about the production of rice in China, he/she could find results in English, but it could happen that the most relevant document is in fact written in Chinese. The search expander can retrieve this document in Chinese, and if the user does not read Chinese, the tool can propose a rough machine translation of that document.

Since this cross-language search tool targets both the internal staff and the external members of an international organization (and not only the translators), it has to be integrated in the general document management environment so it can be called up by a manual click on a link or button in the CMS’ search page. This is an example of the SMT tool supporting a larger purpose within a general workflow dedicated to public use.

f. Providing statisticsto the Head of atranslation service to support decision-making processes

An SMT system allows indicating the number of already-existing segments and evaluating the overall quality of its own output. If properly compiled and presented in a meaningful way, this type of statistics can help a Head of translation service to decide whether to translate a document internally and externally, and to evaluate the time needed for the job.

Such a service can be integrated into the management’s workflow as a complementary decision-making tool. Alternatively, it may be a standalone service which is manually called when needed.

There are many other possible integrations of an SMT system with other tools, both for translators and for the other users. For example Moses can be used to support a terminology

extraction process, or as a simple dictionary for technical terms, etc. Those applications have not been explored yet by Simple Shift but they might be added to the translators' workflow at some point.

2. Building Blocks forthe Integration of an SMT system into a Translation Service

In the production environment of a translation service it is generally out of question to reshuffle the entiretranslation workflow to take into account the specificities of an SMT tool (even if some re-engineering might make sense, as mentioned in our conclusion). Therefore the implementation process should go the other way round and try to make the SMT system fit where it makes sense within the existing workflow.

Defining the main building blocks in integrating an SMT system into an existingtranslation workflow comes down to answering the following questions:

- Where are documents coming from and where are they going to?

Whatever the use of the SMT system within the translation service, all materials which will be submitted to the system for translation will come from the various document input sources. It is all the more important to identify those sources if the machine translation operation has to be performed automatically.

After the document is machine-translated, its destination depends on whether we are in a CAT or fully-automated scenario. In a CAT scenario there is a need to identify first the translator who will post-edit the document, and thus a need to develop a Graphic User Interface (GUI) allowing the translator to enter his/her email address (as was done at the WTO, see below). In a fully-automated scenario it could be considered, for example, to send directly the translation to a frame on the organization's intranet or web site.

- How are the documents currently processed to support CAT tools (building bi-text corpora, building translation memories, etc.)

Moses uses only plain text files. Thus if an existing CAT tool already converts the existing and new documents into text it might be useful to tap into this converted repository rather than convert the documents again from their original format. Besides, if a translation memory is manually or automatically built from those documents after they were translated, it is recommended to use the resulting TMX to improve the training of the SMT system, or at least to improve its segmentation capacity.

- What are the related formats (docx, xliff, tmx, etc.) and what is the effort needed to make the SMT compatible with them

Integrating the SMT system into the translators' workflow may require to produce the translation under a specific format to allow the exchange of text between the various CAT tools. For example, if some translators work under MS-Word while others use Trados Studio, the SMT system must be able to produce both a .doc or .docx format and an xliff format. In a more radical situation, if Trados Studio is systematically used by the translators and the SMT tool is dedicated to translating the missing segments, the translators won't even see the SMT system anymore:It will be called up by Trados through an API and thus it has to produce a well-formed xliff output.

- How can the SMT system help a translator: Systematic vs. on-demand pre-translation

The next building block is related to the translation workflow as such: If the machine translation is systematically performed upfront there is no need for a GUI but APIs are

required. Various metadata must be passed on to the SMT tool, such as the language pair and required input/output formats.

If there is a human user a GUI must be developed to allow him/her to define the translation corpus (if there is more than one, so as to choose an SMT tool specifically trained for the document to be translated), the source and target languages, and the input/output formats.

- How can the SMT system help a referencing service (statistics, etc.)

Some organizations have an internal referencing service whose mission is to accelerate and simplify the translation process by performing upfront most of the referencing research work, and possibly providing a pre-translation of the document. An SMT tool can of course help by providing the pre-translation: Even if the output quality is poor, there is a good chance that the terminology chosen by the system is mostly correct so it can be re-used by the human translator. The SMT tool can also provide statistics to help decide whether to translate inside or outside the organization.

Here again, the integration tasks depend on whether the system will be used automatically (systematic upfront translation) or manually by the referencing service staff. In both situations the same metadata (language pair, input/output formats, etc.) must be passed on to the SMT system, either through an API or from the GUI.

- Integrating the SMT update processin the workflow

Any SMT system should be periodically retrained to be kept up to date, especially if new topics come up in the organization's activities. A well-designed workflow should include a final loop under which all source and translated documents are fed back into an organized repository where documents are classified by collection. This allows the SMT system to feed its training input from a single place, and possibly to train several SMT systems on different document collections (*e.g.* to build an SMT tool dedicated to a specific topic, see the CERN example below).

This process requires that incoming documents are tagged with the relevant metadata to allow for a correct classification, and that the SMT retraining program is able to discriminate the documents according to that metadata.

So far the retraining process is not able to operate on an incremental basis, *i.e.* it has to re-index the complete corpus to perform the training operation. Thus there is no need, at this stage, to take into account other metadata such as the name and date of the documents, but this could become necessary if an incremental system is developed later on.

3. A Practical Example : Integrating Moses at the WTO

a. Strategy and Goals

i. Support the pre-translation process by translating all the segments which are not available in the Translation Memory system (SDL Trados Studio 2009)

The WTO has an internal referencing service which started out with a clear strategy of supporting the Trados Studio tool with the SMT system. However, in a first phase it was decided that the referencing service would submit manually the text to be translated to the SMT system, which would provide the translation under the xliff format so as to allow a direct input into Trados Studio.

To that end, Simple Shift wrote a converter from xliff to plain text format (which is the standard output of Moses) and back to text. Various issues were met in that process, essentially due to the fact that some tags were difficult to process, and also because Trados Studio 2009 is not entirely compliant with some programming conventions. However, the

resulting output of the SMT system is mostly compatible with Trados and the remaining tags which can not be automatically processed are currently manually removed by the referencing service staff.

ii. Allow users to manually submit a whole document

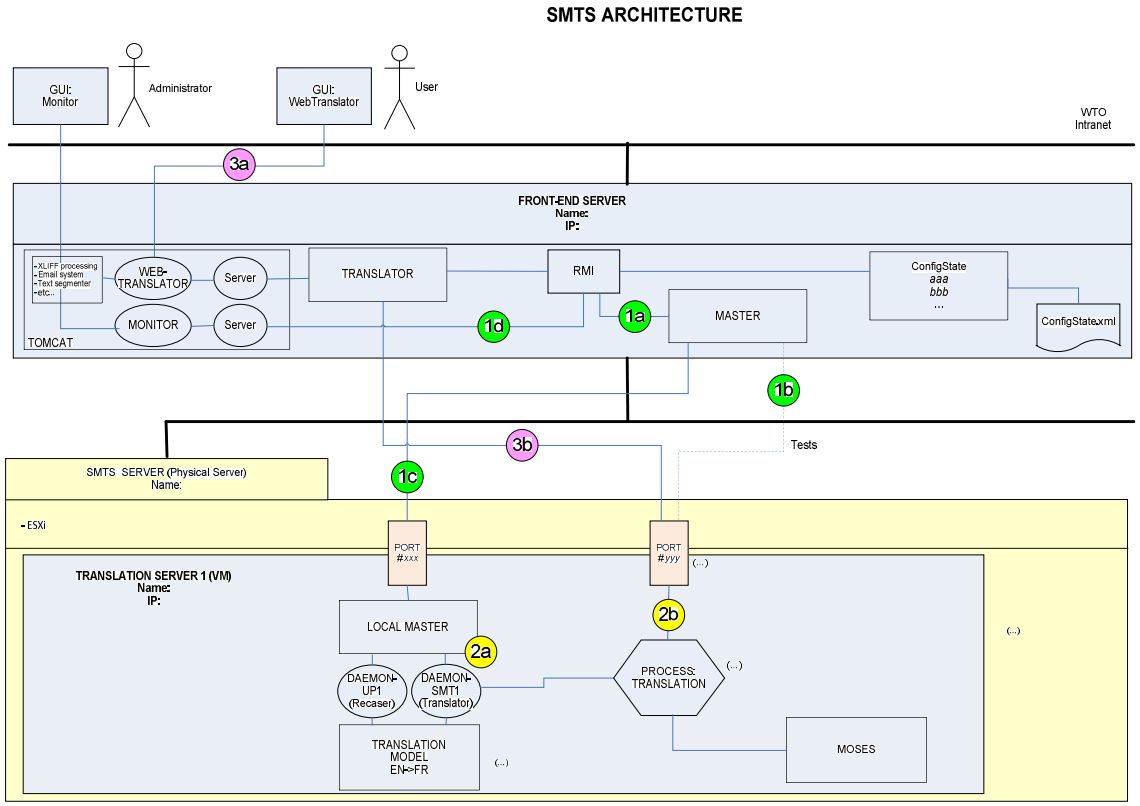
Over time however, it appeared that Trados Studio was not systematically used and a number of users wanted to directly submit a complete document to the SMT system. Thus Simple Shift is now developing a converter for the .docx format: Users will manually upload a .docx document which willbe automatically converted into text and then translated. The translation will be converted back into the .docx format. In that last process most of the layout information (such as titles, bold and italics, numbering, etc.) will be restored. An issue remains with the footnotes, because the system does not know which word in the target sentence should be linked to the footnote number; so each footnotewill be placed at the end of the relevant target sentence.

iii. Support the bitext-based CAT tool by improving the automatic alignment of full documents

The WTO plans to upgrade its current bitext-based CAT tool and may use the SMT system to build alignment maps. A separate module would haveto be developed for that purpose: it would use the same SMT engine but would operate in back-office mode and would store the alignment maps to be later used by the CAT tool. While an API would have to be developed so the CAT tool could use those maps, no further integration in the workflow would be necessary.

b. What is Already Implemented

Currently the SMT system is integrated at the WTO under the following architecture:



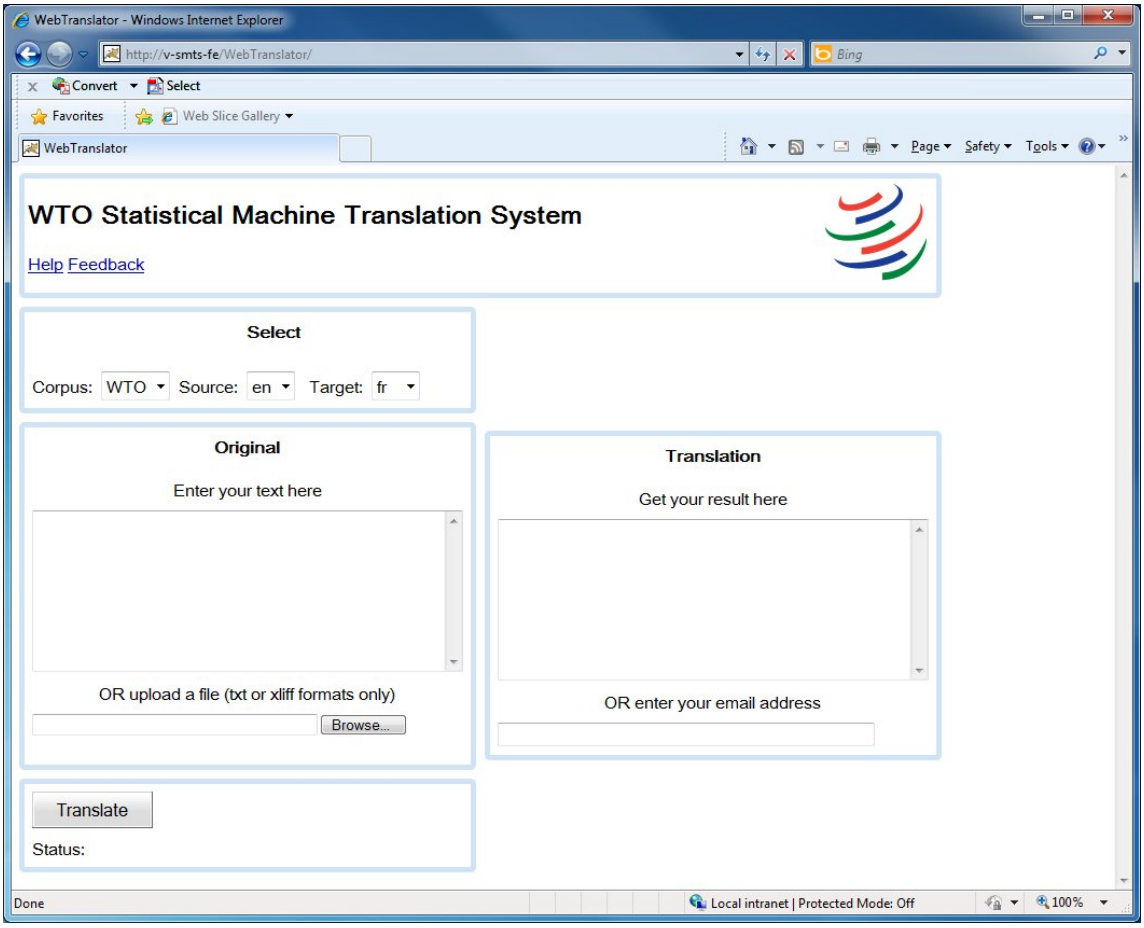
- **1a:** A Master program, located on the Front-End server, readsfrom the ConfigState parameters (through the RMI utility, which allows programs to be provided as online services) which translation nodes should be available on which ports (for the various languages pairs and corpus collections).
- **1b:** TheMaster tests the related ports to check out whether the translation nodes are responding or not (“are up or down”).
- **1c:** If a translation node does not respond the Master requests from the Local Master (located on each translation server and accessed through port no. *xxx*) to re-launch it.
- **1d:** The Master sends information to a web interface called the Monitor, which displays the list of all the translation nodes with their respective availability status.
- **2a:** When the Local Master receives from the Master a request to re-launch a translation nodes, it sends to the local Translation Daemon the instruction to launch a translation process for the collection and language pair concerned.
- **2b:** The translation process initiated by the daemon is then made available on the network through a port number which is automatically generated (ex: *aaa*).
- **3a:** When a user submits a text to be translated for a given collection and language pair, a program called the Translator checks out the port numbers on which the related translation nodes are available, and sends them the text segmented in sentences.

- **3b:** The translation nodes reply directly to the Translator, which centralizes the translated sentences, rebuilds the target text and displays it in the GUI or sends it through the email system.

This architecture is a bit complex but it is quite robust and scalable. It is designed to be flexible enough to affect the translation power required for each language pair according to the translation demand.

Upstream, the SMT system has access to a repository of documents where the administrator can copy a new corpus if he wants to update the translation models. The documents can be copied in their original formats; the supported formats are doc, docx, xls, ppt, pdf, txt, htm, html, odt, ods, odp, csv, wpf, rtf.

Downstream, the system has two GUIs, one for the users and one for the administrator (le latter will not be detailed here). The user GUI looks like this:



The corresponding workflow is the following: The user chooses the automatic translator (currently there is only a complete “WTO” model and a smaller “test” model, but collection-specific models could be built later on) and the source and target languages. Then the user pastes the text to be translated in the text area and clicks on the Translate button. Alternatively, the user can upload a plain text or xliff document (the .docx format will soon be supported too).

If the “Original” text area is used, the SMT system displays the translation into the “Translation” text area. If a file is uploaded, the user must enter his/her email address in the

corresponding field; the translation will be sent back to this address through the WTO's messaging system, to which the SMT system is allowed to access.

If the uploaded document is an xliff file, its source column must be populated with source segments; the SMT system fills up the target column and returns the completed xliff file.

c. What Remains to be Implemented& Related Challenges

The next step is to integrate the .docx add-on described above; this should be completed by the end of 2011.

Later on the WTO would like the SMTS to be automatically called up by Trados whenever there are segments to be translated. This requires using an API available under Trados Studio; a close cooperation with SDL will be necessary to get the relevant documentation and some technical help.

Currently Trados highlights in different colorsthe translations which are produced by the translation memory andthose which are produced by their MT engine. The Trados API may have to be adapted for the same service to be used with the Moses engine.

4. Further Projects : Integrating Moses at CERN

a. Different Tools and Challenges

CERN uses WordFast for Translation Memory management. They have a similar goal : to automatically translate all segments which could not be found by WordFast, and to feed back the TM content produced during the training phase into the TM system.

CERN's main aim, in the medium term (coming 2 years), is for its CERN-specific SMT system to reduce, by a significant margin, the time taken by in-house translators to produce their first-draft translations, being fully aware that much of the acceptable MT output may not ultimately "survive" into the final published version, due to the inevitable reordering and reorganisation performed by human translators in order to render a translation readable in the target language (SMT strictly follows the source language sentence structure).

For the longer term, and depending on the success of the general SMT system currently under construction, CERN is considering building a dedicated SMT system which would be exclusively used for translating Vacancy Notices for the HR Department. This system would be trained primarily on previous vacancy notices as well as on HR documents and other administrative texts (Staff rules, Competency Model, etc.) The ultimate goal would be to launch the SMT service automatically whenever a new Vacancy Notice is written in English so as to provide the HR Department with a French version immediately, with a view to simultaneous publication of the Vacancy Notice in the two official languages.

There are no plans, for the time being, for the SMT system to be made more widely available to the CERN community or beyond, and thus no need to integrate it into a workflow outside of the translation service.

b. Where We Are

So far the SMT system works in standalone mode only. However it seems WordFast Pro has an in-built MT tool (just like SDL) so Simple Shift is currently investigating the possibility of using a WordFast API to call Moses instead of WordFast's MT engine.

WordFast seems to support XLIFF (at least as a standard XML file) so the XLIFF feature of the SMT can be used by CERN translators (although at this stage this format is not currently used internally).

Various statistics will be compiled so as to indicate which proportion of the text to be translated already exists in the reference corpus, and to rate the average quality of the machine translation. This could help the Head of the Translation Service to decide whether to translate the document internally or externally, and to adapt his estimate of the time required to perform the translation.

Simple Shift has made the SMT solution available in web service mode so it is now possible to call automatically the translator whenever a document in source version is made available for publication. Thus it will be possible to publish documents on CERN's website in both languages at the same time (this feature could be used for example if the HR-specific SMT tool is deployed in production).

5. Further Projects : Integrating Moses at the UPU

a. Using Moses to Support a Search Engine

Most staff members and country representatives in international organizations speak at least two languages (and often more). Thus they are able to search for documents in more than one language. However they currently have to perform an initial search in a first language (typically English, which has the largest content) and then perform the same request in another language (typically their mother tongue).

Moses will be used at the UPU to extend a single-language search to one or several other languages. Besides, since the search extension produces a lot of noise (non-relevant documents), an automatic classifier will be used to select the most relevant hit documents and to order them by relevance.

b. Integrating the Search Solution in a Content Management System

The cross-language search engine does not cover the Internet at this stage, but it does target the internally-produced documents, which are stored under a CMS application (SharePoint 2010). The SMT system is called by SharePoint through a webservice to expand the search to additional languages. This integration is currently being performed by Simple Shift and the company which integrates SharePoint at the UPU.

Moses will also be used to propose a bilingual display, with automatic alignment, of documents found through a single-language search. This feature is so far limited to the Translation Service of the UPU and turns the SharePoint search engine in a simplified bitext-based tool.

c. More Challenges to Come

More languages will have to be added to the cross-language search tool, because it currently only supports English, French and Spanish. It is foreseen to add Arabic, Portuguese and Russian in the future.

So far the cross-language search tool expands a query to several languages but does not translate the hit documents back in the original query language. This feature will have to be added later on because it would allow users to search into any available language, including a language they don't speak.

Conclusion

Most of the efforts currently deployed to integrate the SMT solution aim at fitting that solution into the existing translation workflow. This implies considerable efforts in terms of adapting the SMT system to the other CAT tools,to the content formats and to the security rules at each stage of the translation process.

However some translation services have recently shown that it may be wise to re-think partially or entirely the translation workflow in the light of what can be done with an MT system. The new workflow which thus emerges places all pre-translation work upfront with no human operation. The quality of the pre-translation output becomes the key to the rest of the translation workflow: the pre-translation can be used as such and directly published, or sent to an internal or external translator for quick proof-reading or for deeper post-editing. Thus machine translation and translation memories become much more than a help for human translators: they determine the entire translation workflow downstream.

Such a re-organization seems promising but it is heavily dependent on two elements:

- The accuracy of the machine translation; and
- The efficiency of the integration between the SMT and the TM systems.

Although the first track has been and continues to be extensively researched, we believe that further efforts should also be dedicated to the second track because it is one of the most promising ways to integrate SMT systems into current and future translation workflows.