TAPTA4UN: Collaboration on machine translation between the World Intellectual Property Organization and the United Nations

Cecilia Elizalde, Bruno Pouliquen, Christophe Mazenc and José García-Verdugo

This paper presents a collaboration on machine translation (MT) between the Global Databases Service (GDS) of the World Intellectual Property Organization (WIPO) and the Documentation Division (DD) of the United Nations Headquarters (UNHQ). The collaboration started in November 2011 after informal conversations held during the ASLIB Conference. In the context of this collaboration, DD provided 11 years of human translations in bitext format (around 60,000 bitexts for the first language combination, English-Spanish, and slightly more for English-Arabic, English-Chinese, English-French and English-Russian). Using these bitexts, WIPO developers trained its existing machine translation tool, called TAPTA, which was developed using Moses technology and adapted for the translation of patent abstracts. Then, WIPO developers installed the resulting systems in cloud servers procured by the United Nations (UN) and delivered a one-week training session to our IT staff to teach them how to maintain and retrain the existing systems and to develop new ones using Moses.

The performance of the prototype trained using UN documents was evaluated as satisfactory or very satisfactory in all the language combinations, as the system produces workable drafts that can be post-edited by professional translators, in particular for some categories of documents. Currently, the system is used within five Translation Services in DD (Arabic, Chinese, French, Russian and Spanish). Translators access the system through a web interface for gist translation and post-editing, or through a plug-in for Studio 2011. Given the successful acceptance of the tool among UN translators and its promising initial results, machine translation was added as a component of a global UN project called gText aimed to modernize the documentation workflow of the UN. In the near future, TAPTA4UN will be integrated to an in-house computer-assisted translation tool (code name Mercury), developed by the United Nations Office in Vienna (UNOV).

General information

The Documentation Division at UNHQ in New York translates around 90 million words in six official languages (Arabic, Chinese, English, French, Russian and Spanish). Most documents are publicly available through the Official Document System (ods.un.org) in monolingual format (.doc and .pdf). In 1994 the UN signed an agreement with the Linguistic Data Consortium (LDC) of the University of Pennsylvania to make its corpus available for research purposes. This agreement was renewed in 2011 and an updated UN corpus is expected to be released by LDC in 2013.

The UN has also an extensive collection of bitexts (parallel text) covering from 2000 to date, for all language combinations. Bitexts are automatically generated several times per day using Align Factory. However, the quality of the alignment is far from perfect due to formatting discrepancies between the documents. WIPO developers applied another sentence

alignment tool (tailored for machine translation) on this set of bitexts, which was then used by to train TAPTA4UN.

The UN documents submitted for translation in New York deal with a great diversity of subjects, including 10%-15% of documents relating to budgetary and administrative issues that are good candidates for computer-assisted translation because they contain around 30% of repetitive language. For the full production, the rate of revision is around 50%, while the other 50% of the translations are done as self-revision (these percentages vary among Translation Services). About 80% of all the output of the Division is subject to proofreading or scoping (verification of main elements of the text, such as titles, paragraphs, numbers and names) at the text-processing units of each language.

Input methods vary depending on the Translation Service and are correlated to demographics. In general, younger translators or translators who have worked in the private sector before joining the UN tend to favour keyboarding and computer-assisted translation (CAT) tools, while senior staff who have spent most of their careers at the UN or other international organizations usually prefer dictation. However, dictation is gradually being phased-out for budgetary reasons and voice recognition is not available for all the UN official languages.

Use of machine translation at UNHQ

UN translators in New York have been exposed to machine translation through Google Translate and Bing Translator (either directly or through CAT tools) and have found that the output quality, for the purposes of the translation of some categories of UN documents, is good enough for post-editing. However, the quality of machine translation, especially in Google Translate, has decreased over the years as documents from other organizations were added to the system. For this reason, the UN was interested in developing its own tool with its own documents. Regarding confidentiality, it is important to keep in mind that the overwhelming majority of documents translated at the UN are for general distribution and are available on the Internet through the Official Document System.

In automatic evaluations for all the language combinations, TAPTA4UN has consistenly obtained better BLEU scores than Google Translate, using as reference 1.000 segments of human translations not included in the training. The same goes for human evaluation. In this respect, the only system that was the object of a structured, blind evaluation was the English-Spanish system, which rated roughly 4/5 for fluency and 4/5 for accuracy. Scores vary widely depending on the categories of documents processed with machine translation. For some cyclic documents produced by the UN Secretariat, the scores and the quality are very high, while for documents submitted by Members States the scores and quality are less good. This variation is correlated with the categories of documents contained in the training set (if a similar document was included in the training, the machine translation system will produce better results).

In some categories of documents, the output of TAPTA4UN allows translators to speed-up the translation process. However, no productivity analysis has been conducted so far as to measure efficiency gains. Currently, TAPTA4UN is available and translators are free to decide when they want to use it and how they want to use it (through its web interfaces or through Studio 2011). An aspect that is unique about the implementation of machine translation at UNHQ is that translators, and in particular senior revisers (P5), have been

enthusiastic about adopting TAPTA4UN as an additional tool[1]. Some revisers are not used to touch-typing and they appreciate the fact that machine translation gives them a typed draft. As UN translators have been very systematic over the years in the use of terminology and style guidelines, the output of TAPTA4UN is very consistent. Senior revised are also used to correct and evaluate the work of junior translators, so they are comfortable with post-editing. It is interesting to note that most revisers report that post-editing requires a different intellectual effort than translation, an effort that is similar to revision but that is still more challenging, as in some cases the system might deliver sentences with high fluency but low accuracy (sentences that are gramatically correct, but where the meaning has not been fully reflected or is totally wrong).

Given the positive user acceptance of TAPTA4UN among DD translators, the scope of gText, a current global project to develop an integrated and web-based suite of terminology, referencing and CAT tools for all UN duty stations, was expanded to include also machine translation. As a first step, the existing language combinations of TAPTA4UN will be integratedto the gText CAT tool. Then, new language combinations will be developed and new services will be created (for instance, to translate MS-Word documents). Also, we would like to conduct structured, blind human evaluations for all language combinations besides English-Spanish.Translators are closely collaborating with WIPO and UN developers and have proposed to add to the system a terminology verification feature, where the output of TAPTA4UN will be automatically checked against the UN main termbases to ensure compliance with official terminology. As today, there are no plans to open the system to other UN departments for gist translation, so TAPTA4UN remains a machine translation tool for professional UN translators.

For technical information on TAPTA4UN, please contact Bruno Pouliquen (bruno.pouliquen@wipo.int) and Christophe Mazenc (christophe.mazenc@wipo.int) (WIPO).
For information on the TAPTA4UN plug-in for Studio 2011, please contact Michal Ziemski (ziemski@un.org) (UNHQ).
For information on usage of TAPTA4UN by UN translators, please contact Cecilia Elizalde (elizalde@un.org) and José García-Verdugo (garcia-verdugo@un.org) (UNHQ)

[1] The current translators' toolkit at UNHQ includes dtSearch and dtSearch web (indexing and full-text search), Studio 2011 and Multiterm 2011 (computer-assisted translation), AlignFactory (bitext generation), LogiTermWeb and Logitrans (referencing and pre-translation), UNTERM and the Global Terminology Portal (terminology), WorkShare Compare (text-comparison), Dragon Naturally Speaking (voice recognition), Express Dictate (digital dictation) and a a very large set of MS-Word macros organized in toolbars, among other tools. The aim of the gText project is to offer a new set of integrated tools, internally developed, and accessible through a web interface.