MT@EC: Working with translators

Markus Foti
DG Translation
European Commission

The European Commission’s Directorate-General for Translation (DGT) is currently building a new machine translation system, dubbed MT@EC, to be used by the Commission’s Directorates-General and ultimately by other EU institutions, some public-facing services, and public authorities in the EU Member States.

The project is being built using statistical MT technology and the open-source MOSES engine in combination with DGT’s vast database of translated documents covering the 23 official languages of the EU (Euramis), and the expertise and effort of its 1750 translators.

Work is ongoing in three strands: *data*, focussing on cleaning the data available and adding other sources, *engines*, building and improving the MT system itself, and *services*, building future portals for other Commission users and other Institutions.

To ensure that the final product will meet end-users’ needs, the involvement of DGT’s translators was sought as early as possible in the project. A Machine Translation User Group (MTUG) was set up as a forum for translators to provide feedback not only on the output of the project itself, but also on its structure and focus. Its 51 members from 23 language departments and the IT unit serve as contact points for communication with translators and for organising feedback and testing within their respective language departments.

The first task carried out by the MTUG was to assess the quality (defined as usefulness for translators) of the system’s output. This so-called “maturity check” kicked off by giving translation staff with an interest in machine translation access to a first generation of engines from English into their languages. These volunteers used the output from the engines as a translation aid, then assessed the usefulness of these baseline engines for translation work. This test was completed in May 2011.

As was to be expected, results varied across the various language pairs. Nonetheless, translators deemed 10 of the language pairs (EN into BG, DA, ES, FR, IT, NL, PL, PT, RO and SV) to be sufficiently useful to include them in the next stage: the “real-life trial”, where machine translation is integrated into existing workflows through the automatic provision of tmx files for use with the CAT tool (Trados Translator’s Workbench 7) in day-to-day production.

One year on, MT as a tool has gained greater acceptance. Even prior to the introduction of the second generation of engines, a further seven language departments (CS, FI, LT, LV, MT, SL, followed by EL) elected to have machine translation provided as a matter of course.

A new generation of engines was built for English into the remaining 22 languages, and vice-versa and became available in June. BLEU scores are used as an automatic metric, and, as was to be expected, the scores improved simply through the addition of an additional year’s data. Nonetheless, although the incremental change according to BLEU was always positive, it ranged from a low of 0.98% (EN-RO) to a high of 4.25% (EN-GA). The latter is an outlier, as Irish has much less data available for building engines and a very low score to begin with, so a larger amount of data would be expected to produce significant improvements. The next highest improvement was that of EN-EL, at 3.79%.

Although the engines team felt confident that the gains were significant, confirmation was sought through a second round of testing. Apart from the obvious benefit of obtaining human confirmation of the trends evinced by the flawed but useful BLEU score, additional testing would afford an opportunity for the engines team to devise and test a modular tool for comparing the output of two engines.

A simple binary comparison was sought, as the issue to be assessed was straightforward: were there any surprises undetected by BLEU that would prevent the new engines from being deployed?

There were proponents of more-involved testing, which would require the volunteer testers to assess the usefulness of each sentence on a scale of 1 to 5, and provide feedback regarding typical errors, but as there were no plans to address such issues at that time, and it is very labour intensive, this approach was not taken.

To carry out the binary comparison, a simple web-based evaluation tool was set up. It was designed in such a way as to make future surveys easier to design and implement, and is thus an important building block in an ongoing cycle of feedback for the various engines.

The MT@EC evaluation tool presents a blind binary test comparing the engine currently in use with its proposed replacement. Translations are presented in a random order, along with the original text and translators were asked to select which of the two options they deemed better.

No show-stopping errors were reported, and the second generation of engines was deployed.

Binary comparison of engines' output is also supplemented by other forms of feedback to allow for more finely-grained identification of problems specific to languages and to text processing itself, and thus, it is hoped, targeted solutions.

The first, rather simple, one of these was a typographical errors survey carried out in April 2012.

A recurrent complaint among translators working with MT@EC's machine translations was the repetitive and mechanical nature of certain corrections that were frequently required. These are not corrections of grammar or style, but almost akin to proofreading. They give rise to a sense of frustration on the translator's part, and have a disproportionate negative impact on perceptions of the machine translation's quality.

To unearth these, a survey was carried out using the members of the MTUG as the contact points for their various language departments.

58 errors were reported, ranging from extraneous spaces around quotation marks to the corruption of characters in LT and LV. Probably the largest complaint was due to a spelling reform in Portuguese. Since MT@EC's engines have been built on bilingual data going back eight years or more, Portuguese MT included both variants, causing Portuguese translators to waste time and effort on small corrections.

As a result of the survey, scripts were written to process the data and bring Portuguese spelling up-to-date before engines were trained.

35 of the errors reported will be fixed in the engines now being built for the end of the year, while another dozen should find their way into the following generation.

A number of the errors were out-of-scope, with some translators reporting issues with grammatical cases or erratic use of turns of phrase. Others, such as erratic bolding or formatting changes, were artefacts of using MT through a tmx file in TWB. Very often, a linguist's mindset was apparent, with suggestions for rules that could be incorporated, which may be useful in future as some attempt is made to add some grammatical knowledge to a hybrid system, but it was clearly outside the scope of the survey at hand.

But DGT's translators are not merely there to provided answers to developers' questions. They can also propose avenues of investigation.

One of the recurrent requests was to make MT@EC's terminology output more reliable by incorporating data from IATE (Interactive Terminology for Europe, available at http://iate.europa.eu). It was suggested that adding the terms from this database would improve the MT system, mirroring translators' working patterns, whereby validated terms in the database are considered to be more reliable than Euramis or other sources. Tests were done using the EN-HU language pair, and, disappointingly, BLEU scores actually dropped.

Further investigation showed that forcing terms to be used according to IATE introduced errors because one of the principles underlying the inclusion of a term in that database is that the meaning recorded must not be the primary one from an EU terminology perspective. The most telling example was that "Commission", as in "European Commission", rather than being translated properly into Hungarian as "Bizottság" was now being translated as "jutalék", that is, a payment for services rendered.

Further experiments are also being undertaken within the engines team, before testing is be broadened to the members of the MTUG and then to translators in the language departments concerned.

One of the most promising of these is word-reordering of the English source for translation into German and Hungarian. Stemming and adding case endings after translation is being looked at for the Finno-Ugric languages.

A corpus is being prepared, and the language departments concerned will soon be asked to test sentences generated by the current system against those which apply the experimental approaches.

A binary system will again be used, largely because, although a large pool of testers is available, MT evaluation is always competing against other priorities, mainly tight translation deadlines.

Nonetheless, evaluation methods which provide a more objective measurement of MT@EC's output are being sought, with the focus on capturing time spent and editing effort.

With a view to this, the Post-editing Tool devised by Wilker Aziz, U. of Wolverhampton and Lucia Specia, U. of Sheffield, has been distributed to members of the MTUG's Evaluation Methodology Task Force before distribution to a broader audience.

This measures time taken and keystrokes, and HTER can be applied to its output file to measure translation distance. This would provide a much more meaningful assessment of the usefulness of the output of different engines, both of subsequent generations of the same language pair, and across language pairs. Unfortunately, considerable adaptation is needed for DGT's production environment, as the tool works on flat files, to which DGT's translators do not have direct access, and its output is an xml file which is rich in terms of data, but is not directly usable as a translation.

Given that DGT already has a cycle in place whereby translators are provided with MT, edit it, and the completed translations are stored on a segment by segment basis in Euramis, the holy grail is to record which segments began their life as MT, obtain the resulting final translation, and measure the distance between the two renderings.

As DGT will be updating its CAT tool in 2013, work on integrating it is currently underway. The engines team has submitted a request to record when MT is used. It is hoped that a system whereby translators are automatically providing steady feedback simply by doing their work could be put into place. This would in no way preclude other forms of direct feedback, but, in future, these would be refocused to address specific issues.