Combining MT and TM on a Technology-oriented Translation Masters: Aims and Perspectives

Mark Shuttleworth
Humanities Programme
Imperial College of Science, Technology and Medicine
Mechanical Engineering Building
Exhibition Road
London SW7 2AZ
E-mail: m.shuttleworth@ic.ac.uk

Abstract

This paper has two aims. The first is to report on a project which was recently carried out as part of the Imperial College MSc in Scientific, Technical and Medical Translation with Translation Technology. For this project students were given the assignment of translating a large extract from a medical information document using a combination of machine translation and translation memory technologies. The second aim is to discuss the potential of this approach for increasing understanding of 1) the complementary function of these two fundamentally different approaches to automating the translation process, 2) the possibilities of future convergence between the various technologies and 3) practical file format manipulation techniques for facilitating data exchange. Possible modifications to the project in future years are discussed, and in the light of the project conclusions are drawn regarding Masters-level training in translation technology.

Background and aims

The Imperial College MSc in Scientific, Technical and Medical Translation with Translation Technology (MScTrans for short) is a programme which aims to provide a qualification for students wishing to enter on a specialised translation career. Alongside a range of other practical and theoretical components, the programme also offers intensive hands-on training in a range of translation tools. While the course includes a significant component of training in machine translation, the main emphasis is on the theory and use of translation memory (TM) systems, with the programme covering IBM TranslationManager, Déjà Vu, TRADOS Translator's Solution, STAR Transit and SDLX in the course of the 2001-02 academic year.

The aims of this paper are twofold. The first is to report on a recent project which was carried out as part of the MScTrans programme and in which students were given the assignment of translating a large extract from a medical information document using a combination of MT and TM technologies. The second aim is to discuss the potential of this approach for increasing students' understanding of 1) the complementary function of these two fundamentally different approaches to automating the translation process, 2) the possible future development of TM technology and 3) practical file format manipulation techniques for facilitating data exchange.

The Project

On MScTrans we aim to supplement the practical training in the use of translation tools with a series of practical, team-based projects in which a number of teams (within which each student is assigned a particular rôle, such as project manager, terminologist, translator, MT operator, MT post-editor, etc.) have to use specified software packages in order to complete a large-scale technical translation project, working from English into a number of languages with a strict deadline imposed. On a fairly experimental basis, we decided to include a project which was to combine two major computer-based translation methodologies in the manner described above.

The principles behind this approach to translation automation are simple. The user employs the “Analyse” facility of a translation memory tool (in this case TRADOS Translator’s Workbench 3) to create a list of the new and/or unique sentences contained in a text by analysing the text – against a translation memory if there is one available, or else simply in terms of the amount of repetition that it contains. The sentences extracted in this way are sent to an MT system (SYSTRAN PROfessional Premium 3.0 was the one employed on the project) in a specified export format, machine translated and then reimported into the TM tool in the form of a translation memory all the segments of which have been assigned an automatic penalty to ensure that they are marked as the product of the MT system and are not accidentally accepted as perfect matches. The translator then “translates” the text sentence by sentence using TRADOS, post-editing the MT output as appropriate.

The project also offered the possibility of automatically extracting candidate terms for a word list, using the SyNTHEMA Terminology Wizard.

This technique of combining MT and TM is resorted to because of the theoretical productivity gains which may result. It has been around for a number of years, and different TM tools offer different possibilities for interfacing with MT engines. A number provide a ready-made data-exchange format for working with a particular tool. TRADOS Translator’s Workbench falls into this category as it offers a special SYSTRAN filter (which is complemented by SYSTRAN’s special TRADOS import format) and the company provides instructions on its website in how to use the two tools together. SDLX, on the other hand, both offers MT functionality from within the TM tool as an optional extra (using the Transcend engine) and also provides the possibility of configuring any available MT engine for direct data exchange. Failing that, it is in principle possible to combine any TM tool with a wide range of MT engines in an ad hoc manner by means of a series of complicated file manipulations based on segments exported in the XML-based TMX (“Translation Memory eXchange”) format (see Mügge 2001).

By the time the project was undertaken the students had acquired a working knowledge of three major TM tools (IBM TranslationManager, Déjà Vu and TRADOS) and one major MT system (SYSTRAN PROfessional Premium) and had also attended four or five lectures on the theoretical aspects of MT. The text selected did not contain significant amounts of repetition so that the use of other technologies besides TM had the potential to add greatly to the automation of this particular translation job, provided the task of post-editing did not prove too onerous. The project was written up as part of the assessment for this course component; as a result, multiple and detailed student evaluations of the project are available for analysis.

A number of the students did express reservations concerning the usefulness of this approach for this particular project. The most negative comment was from someone who stated that “we would all agree on omitting the SYSTRAN part”. However, this seemingly representative statement was belied by the majority of other comments, as most participants did in fact see at least some potential benefit in the approach, particularly in terms of the time-saving potential which it offered for larger-scale translation projects. Interestingly, some viewed the terminological help provided by the SYSTRAN glossaries as a kind of glorified automatic look-up facility. Finally, the most positive comment was that using the tools in tandem constituted “a powerful combination”. All in all, it was clear that it had been a highly stimulating experience for most participants, and had led to unexpected insights for many of them.

The rest of this paper will be given over to evaluating the potential of this approach for increasing students’ understanding of translation technology. As already stated, by the time of the project students had acquired a broad (if not deep) experience of working with TM systems: they knew how to operate several such systems, although were probably not fully expert users of any of them. With MT on the other hand they were less familiar. However, in terms of theoretical knowledge they had had a more thorough grounding in MT than TM. In the light of this one of the main aims of the project was to raise students’ awareness of how TM fits into overall patterns of translation technology use – in terms both of how it relates to other technologies, most notably MT, and also of its own strengths and weaknesses as a methodology in its own right. By the end of the project all participants – either through direct personal involvement or through interaction with other team members – had had ample chance to reflect on these two different technologies and how they differ in terms of their potential and their possible drawbacks.

Three main potential areas for student learning were identified at the beginning of this paper, and these will be considered in the following three sections. At the same time I shall include a number of other relevant issues which were not directly touched on during the project.

Complementary function of TM and MT

The expectation amongst many new students – possibly reinforced in some cases by listening to a daily humorous slot on Radio One – is that MT is little more than an amusing plaything. Seventy-five percent of them, on the other hand, will not have heard of TM – which is a slightly harder concept to grasp, and is certainly hardly known at all outside professional translation circles. Both methodologies are introduced fairly near the beginning of the course, with outline descriptions of what each strategy involves and where it is most appropriately applied being discussed, although at this stage the possibility of combining the two techniques is not mentioned.

One of the problems with the way TM is often presented is that it can be sold as a panacea: buy our product and you will cut translation time by at least 50%, goes the sales pitch, so that, I suppose, one of the advantages of the way the technology is presented on Masters courses is that students are able to compare the pros and cons of several products and reach their own conclusions as to what the software is realistically capable of. MT, on the other hand, tends if anything to undersell itself, and there is a definite risk that students will end up convinced that it is only worth using it in real life if they are able to buy into a serious, top-end MT system.

While the text selected for translation in the project may not have been ideal for such a combination of methodologies, the whole point of the project was to enable students to become clear in their own minds about their respective potential and to come to an understanding of how their combined power can be harnessed. Combining the two approaches does of course throw a number of issues into sharp relief for the students, encouraging them to reach their own conclusions on a wide range of questions. Exactly how much help can be gained from the use of technology? What are the strengths and weaknesses of the two approaches? What is the maximum possible level of translation automation – and to what extent does it aid the translator? When could this combination be used to its greatest effect? What kind of help can MT provide for the freelance translator? What are the types of text where neither approach is wholly suitable? What are the benefits and problems of automatic terminology extraction?

Possible future development of translation memory technology

The TM approach has of course long been subject to a number of serious and substantial criticisms. The most important of these are summarised in Multicorpora R&D Inc. (2002:8) as follows:

“1. Dependence on whole sentence repetition

“2. Loss of context

“3. Building a TM database is prohibitively labor-intensive”

The first in particular has become a notorious shortcoming of most commercial TM tools, which as it stands are generally unable to provide matches for segments which are less than an entire sentence in length. (It must be stated, though, that this criticism does not take proper account of the concordance function, which will be discussed briefly below.) Hence, the argument goes, such tools are of very limited usefulness for translating texts with a relatively low level of repetition. Secondly, loss of context is of course a problem, since typically the matching segment located in the TM is presented without any indication of the context in which it was previously used. The third major drawback of the technology is the fact that a TM product is generally sold as an “empty box”: no ready-made TM is supplied, so that the user has to either develop his or her own over a period of time, or go through the long and painstaking process of aligning previously translated material using the alignment tool which most systems now provide.

Thus combining TM with MT can be seen as a way of breaking the log-jam on points one and three in particular, since the ability to receive input from an MT system frees the user from both of these constraints (even if the possibly poor quality of the MT output might present the translator with a new set of problems). And if the approach is presented to the students in this light, it provides an excellent lead-in to talking about the nature of TM and possible ways in which it might be improved.

The fact is that there are two quite distinct types of translation memory system. According to Macklovitch (2000), the first, narrower type consists of the standard range of commercially available products. The second type, which he describes as “interactive bilingual concordancers” (2000:2), involve placing at the user’s disposal a fully searchable “enormous virtual example-based dictionary” (2000:3). In many ways this resembles the concordance function which most commercial TM tools offer, although according to Multicorpora R&D Inc. this facility as offered by most such systems is too slow to be of any realistic support, while the examples which it supplies will be likely to lack a suitable level of contextual information (2002:8). The solution offered by Multicorpora R&D Inc. in the white paper quoted above is in effect an example of a system based on Macklovitch’s broader definition of TM.

Another approach to solving the problem of the “sentence-level only” matching is by means of the AutoAssemble function which Atril Software incorporate into the Déjà Vu TM tool. According to the Déjà

Vu manual, using this facility allows you to "[squeeze] the last drop of information from the lexicon and the databases" (Benito et al. 1999:85) – and indeed, a long-term user of the software will be able to derive huge productivity gains from using this function by recycling phrases and other "portions" from previous translations. In the event, however, these "smaller pieces" are only frequently repeated groups of words which the user has previously considered it sufficiently worthwhile to add to the lexicon, so that there is in fact no question of Déjà Vu being a semantically-enabled system capable of supplementing its TM hits with suggestions from a built-in example-based machine translation (EBMT) engine. Once again, the project should permit such concepts to begin to fall into place.

As previously mentioned, SDLX offers an alternative solution to the same problem by making an MT engine available as a fully-integrated optional add-in to the standard TM package.

It is certainly to be hoped that by the end of the project students will have become highly receptive to learning about the future of TM, and will understand the issues involved in the following types of questions. Is TM technology as presented in most commercial tools likely to continue unchanged? In what sense are TM and MT likely to converge in the future? What is the possible rôle of EBMT within TM systems of the future?

It goes without saying that the project also served to bring to the fore the nature of the processes involved in MT as well as their limitations, although such matters are beyond the scope of the present paper.

File format manipulation

The procedure for combining TRADOS and SYSTRAN as detailed on the TRADOS website consists of some seventeen steps which involve using the TRADOS Analyse facility to identify unknown segments, exporting these in a special SYSTRAN format, using SYSTRAN to translate them and then reimporting them into TRADOS for post-editing into a finalised form. Segments thus pretranslated will appear as matches, although they are assigned a nominal penalty so that the fact that they are the product of MT is clearly indicated in order to inform the user that they will probably require post-editing.

One problem which has not been mentioned up to now is that – with TRADOS 3, at any rate – this TRADOS/SYSTRAN interface did not in fact work as expected, as the output from SYSTRAN was not recognised as it should have been on reimportation into TRADOS. This meant that the already complex procedure needed to be extended by the addition of some nine extra steps in order for the two systems to work together successfully. The TMX format (see above) had to be used. Following export the data needed to be opened in Word and manipulated into table format, and the text for translation isolated in a single table column and saved in a separate file which was then passed to SYSTRAN. After translation this whole complex process needed to be reversed before the data could be reimported into TRADOS. No doubt to their great relief, the students were not asked to come up with the resulting 26-step procedure themselves, although even so some of them found this type of complex format manipulation highly challenging.

In our opinion this type of file manipulation – including such actions as converting between Word tables and tab-delimited text or performing multiple global Find and Replace operations – represents an important skill which can easily be overlooked in training programmes, but which can come in extremely useful in real-life situations.

Future development of the project

Although certainly experimental in nature, we consider that the project as organised last year provided an important learning experience for all the students, and we certainly plan to run similar projects in future years. However, the project could obviously be modified depending on the precise outcomes intended. In any event, a different text will probably be selected this coming year since SYSTRAN did not perform very well with the relatively large number of direct questions which it contained. Besides this, however, at least two more major modifications could be put in place for future years. Firstly, if Déjà Vu rather than TRADOS were used, the limitations of this former tool's AutoAssemble function – the only attempt by a major TM tool to cope with matches below the segment level – would very quickly become apparent when contrasted to the MT approach. Secondly, if the specific outcomes intended were appropriate for this it might be worth investigating using an example-based MT system as the MT engine, as it is arguably here that the crossover potential is greatest – even though the need for a huge bilingual corpus (such as a large TM that has been built up over time) would probably make this impracticable.

Conclusions

One point that we have taken for granted so far is that by the time that participants have completed about half a Masters degree they will have understood the difference between MT and TM. Sadly, experience on such programmes has proven that this is not always the case, as for a very small minority of participants a certain amount of confusion between these two fundamentally different approaches has been known to linger for quite a long time. However, although the project would no doubt clear up any remaining doubts there may be on this score, this is by no means one of its main intended outcomes. The project is more concerned with pushing at the frontiers of usefulness of TM, rather than with defining basic concepts.

At some point it is worth considering how we can justify devoting so much energy to TM systems on training courses of this type. Just what is the point of providing training in the use of TM at Masters level if, as has been suggested, learning how to use TRADOS is about as difficult as learning to use Excel and can presumably be left until the student arrives at the workplace? Surely there is a point, though, as it is through providing a great breadth of coverage of the various tools, backing up practice with theory, placing the use of TM tools firmly in the broader context of other types of translation technology, and perhaps trying to think about possible future developments in the field that we can start to offer something of real value to our students. And of course it is partly through the implementation of projects such as the one described in this article that at least some of these aims are promoted.

TM has existed in its present form for some ten years or so. It has saturated some areas of the translation industry, although within other sectors – most notably amongst freelance translators – its uptake has remained relatively limited. In all likelihood current systems will sooner or later be replaced by something rather different; while not an explicit aim of the project, I consider it important to prepare participants on such technology-intensive training programmes to play an active part in developing the next step, whatever that may turn out to be.

Bibliography

Benito, Daniel, Celia Rico & Tilo van der Berge (1999) *Déjà Vu – Productivity system for translators: User manual*, Madrid: Atril Software.

Macklovitch, Elliott (2000) "Two Types of Translation Memory", in *Translating and the Computer 22: Proceedings from the Aslib conference held on 16 & 17 November 2000*, London: Aslib/IMI. (Pages in this publication are unnumbered; the page numbers cited above have been provided for the sake of convenience and relate only to the pages of this particular article rather than the whole publication.)

Mügge, Uwe (2001) "The Best of Two Worlds – Integrating Machine Translation into Translation Memory Systems: A universal approach based on the TMX standard", in *Language International* 13:6, 26-29.

Multicorpora R&D Inc. (2002) *The Full-Text Multilingual Corpus: Breaking the Translation Memory Bottleneck* (A Multicorpora White Paper); available at http://www.multicorpora.ca/papers/WhitePaper_1.pdf. (Accessed 13 October 2002.)