Accurately Predicting Post-editing Time & Labor for Cost-Management

Carla Schelfhout cschelfhout@sdl.com
SDL International, Maidenhead, UK

Abstract

This paper will describe a way of assessing the post-editing effort for a specific project, language and engine combination. This serves as a tool for LSPs to estimate the necessary effort on the project and quote accordingly.

.

1. Introduction

Over the last decade, there has been an upsurge in the use of machine translation, both for the purpose of gisting and for the purpose of post-editing. There are various definitions of post-editing: the "term used for the correction of machine translation output by human linguists/editors" (Veale and Way 1997), "…checking, proof-reading and revising translations carried out by any kind of translating automaton" (Gouadec 2007) and "In basic terms, the task of the post-editor is to edit, modify and/or correct pre-translated text that has been processed by a machine translation system from a source language into (a) target language(s)." (Allen 2003) are among them. All definitions center around the notion that a human applies changes to machine translation output in order to create a final translation that reaches a previously agreed quality level. We will refer to this process as PEMT (post-editing machine translation).

More and more clients ask their Language Service Providers (LSP) to apply PEMT. Their underlying assumption is that PEMT is faster than conventional translation, as part of the final translation is already there. The demand to use PEMT is mostly combined with a request to reduce the financial rates paid for translation. LSPs have a vested interest in determining if they can afford to comply with such requests. This paper will outline one way of validation.

2. Productivity of PEMT

Various studies have shown (Laubli et al 2013, Plitt & Masselo 2010) that PEMT as a process can provide productivity benefits over conventional translation, at least for the conditions examined in those papers. However, the productivity in a specific case depends on a large number of factors. To mention just a few:

- The engine quality as such. This depends on the technology that was used, but also on the complexity of the source-target language combination (some combinations are harder than others).
- The applicability of this specific engine to this specific project. Was the engine geared towards this particular content in any way, or is it a generic engine? Is the project in question terminology-rich and specific, or very generic?
- How much experience with PEMT do the selected vendors have?

- How has the technical preparation of the files been handled? Any wrong segmentation will likely have a detrimental effect on the machine translation quality.

As the LSP generally needs to provide a quote to the client before starting the job, and needs to negotiate rates with its own vendors as well, it is vital to know in an early stage whether the productivity increase from using PEMT is sufficient to allow a rate reduction to both the client and the vendors. Doing this fully automated would be ideal, but technology is not quite there yet. A human factor in productivity testing is still needed, but the testing needs to be both cost-effective (the cost of testing should not negate the gain of the project) and time-efficient (the LSP needs to have the results in time to quote to the client within his set time limits). The next sections will discuss SDL's approach to this dilemma.

3. Approach

The recommended process has three stages: assessing the project, creating the test set and running the test.

3.1. Assessing the project

The first step, which is still entirely human, consists of an analysis of the project. Aside from the normal steps used for conventional translation, the assessment for PEMT adds a few more questions. The main ones are:

- Does the expected gain of this project justify the cost of PEMT testing? Only if so, the next questions come up.
- What languages are in the project? Do we need/want to test them individually, or would it be possible to group them – for example, assume that the performance of English>French is a good indicator for English>Italian?
- Is the project homogeneous, or does it have flows? For example: a large car manufacturer could offer the user guides for buyers, the marketing brochures for prospective buyers, the repair manuals for the mechanics in the garage and the assembly instructions for the workers in the factory. It seems likely that the linguistic characteristics of these documents will differ, and so will the MTPE productivity. Does the client offer the flows split, or all together? Does it make sense to test them separately?

3.2. Creating the test set

Once a decision has been made, a test needs to be set up for the intended language and content type. This test set needs to be representative and varied.

The representativity of a test starts from the project sample delivered by the client. If not done before, now it needs to be confirmed with the client that their sample is in no way exceptional. The sample has to mimic the total project in (stylistic and terminological) complexity, content and technical characteristics (markup and segmentation). Ideally it will also be large and consist of several outtakes of the total project. This gives a larger variety in topics and the related terminology. Any non-representative or invalid content needs to be removed from the client sample before taking the next step.

In order to select the most representative test set, it is recommended to randomly select segments that have the average segment length of the sample, give or take one or two words. This can be automated, which saves time, and it increases the chance that the linguistic complexity of the test set will mimic the complexity of the sample. A couple of longer and shorter segments can be added to test on the less frequent segments and to add variety to the test. In order to test the engine´s coverage of client-specific terminology, it is recommended to select a number of segments from various parts of the project rather than use running text, which will mostly cover only one or two topics and its related terminology.

For financial reasons, the smaller the test set can be while still giving meaningful results, the better. The smallest possible number of segments depends on the tool used for the testing and the margin of error this tool gives. While we recommend involving a statistician to assess the minimum for use with a specific tool, around 100 segments seems a good rule of thumb.

3.3. Running the test

The next step is to split the test set and have the two parts processed. One part will be done as conventional translation, the other part as PEMT. Both parts have to have the same average segment length to keep the times spent on them comparable. Both parts need to be done by the same resource(s), to ascertain the impact of the PEMT for this resource.

In order to increase the predictive value of the test for the project, it is advisable to use (some of) the resources who are likely to be used on the live project in the test. This will also help when it comes to negotiating rates for this particular project – having performed the test, they will have a better idea of its validity and of the MT value.

The key factor is the registration of times and actions to obtain meaningful results. Ideally, the interpretation of the test will consist of automated indicators to such a degree, that the test can be validated without having to read source or target language.

4. Tools

The more steps in the process can be automated, the cheaper each test becomes. Two steps are candidates for automation: the test bed selection and the hosting and analysis of the test. For the test selection, SDL has created a proprietary tool. It makes an automated selection out of one or more sample file(s), based on characteristics like segment length and linguistic characteristics like question/confirmation etc. Using this tool makes the test bed creation much faster, but as there is a development cost, it is only recommended if enough tests are needed in an LSP to recoup this cost.

The second step, running and analyzing the test, has been automated to a large degree in SDL. The ME tool is an online, proprietary tool, which has been used and further customized for a couple of years. In the meantime, similar functionality has been embedded in freeware like the qualitivity Studio plugin and the TAUS DQF framework. For the purpose of this paper, we will focus on the ME tool.

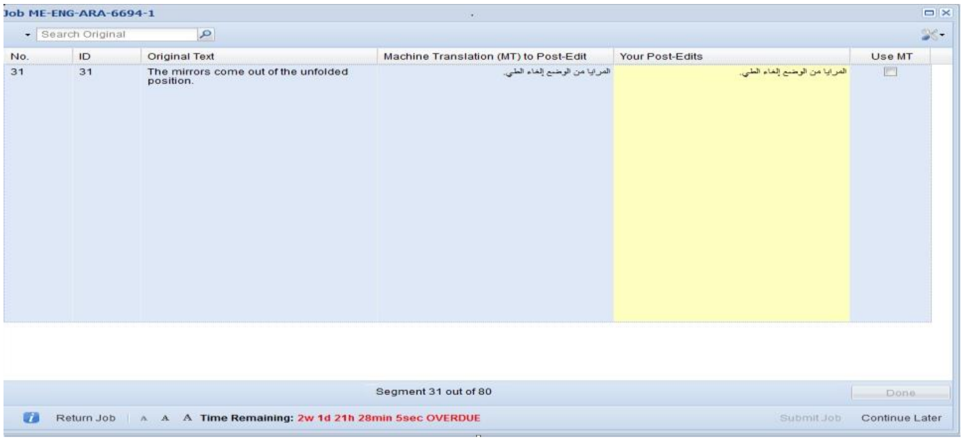
4.1. Characteristics of the ME tool

The tool is online, which means that resources can be onboarded by simply registering. This saves the overhead of resources downloading a tool, and prevents most of the complications of local PC setups causing incompatibilities. A test is uploaded as a tmx file. For the part of the test that is conventional translation, the source is copied into the target. For the PEMT part of the test, the MT is copied into the target field.

The tool displays the segments to the tester one at a time, only offering the next segment if the user clicks « Done ». It allows users to interrupt the test by clicking a button

« Continue later ». They can pick up the test at any later time. This keeps the time registration free of disruptions like telephone calls which would otherwise create noise in the results. Besides these buttons, the conventional part of the test only contains source and target fields, where the target field is editable. The PEMT part of the test has an extra field for the MT output, which remains in view for reference. The target field starts containing the MT output and can be edited. The PEMT UI also contains a button « Use MT » to indicate that the MT is correct as is and needs no changes. The button « Done » is only enabled after either selecting this checkbox or making edits, so as to prevent accidentally skipping segments. A screenshot is shown in Figure 1 :

Figure 1. Screenshot of PEMT part in ME tool



Job ME-ENG-ARA-6694-1

No.	ID	Original Text	Machine Translation (MT) to Post-Edit	Your Post-Edits	Use MT
31	31	The mirrors come out of the unfolded position.	المرايا من الوضع إلغاء الطي.	المرايا من الوضع إلغاء الطي.	□

Segment 31 out of 80

Time Remaining: 2w 1d 21h 28min 5sec OVERDUE

The number of segments needed in a test depends on the exact tool and setup used. In consultation with a statistician, it was decided that for the ME tool 80 segments is the minimum. The tool allows for larger tests, and enables the comparison of up to 5 different engines using up to 5 resources per test.

4.2. Analysis of the ME results

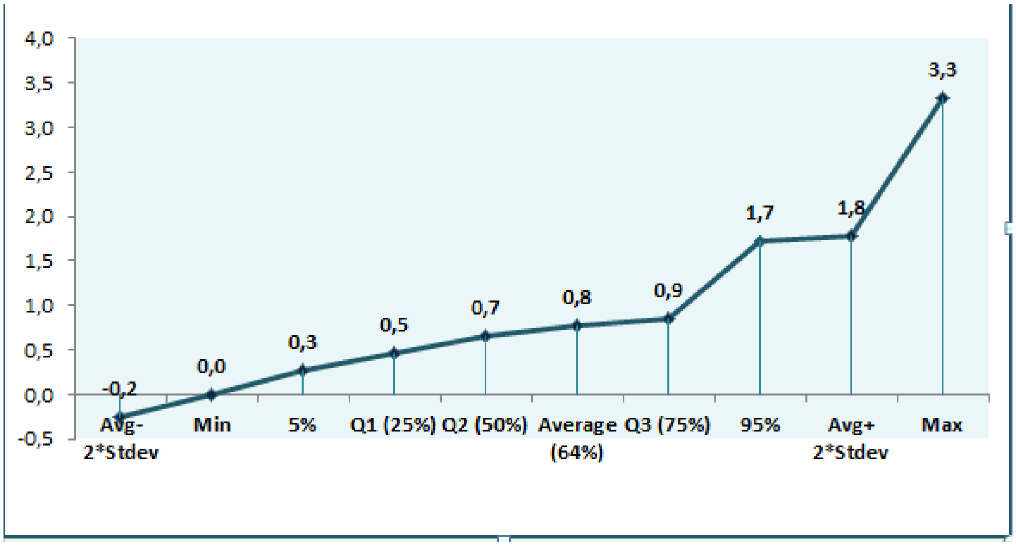
The tool delivers an automated analysis with a number of indicators. These serve to ascertain the validity of the test as well as the actual productivity increase. Among the indicators are:

- The translation speed for each resource in both conditions (conventional human translation, henceforth HT, and PEMT). This speed is not relevant for production as the tool is not mimicking the production environment, but extremely high or low numbers can point to a problem in the test.
- The actual increase in productivity as a percentage of the original speed for each resource. So if for example the HT speed is 800 words per hour and the PE speed is 880 words per hour, the productivity increase is 10%.
- The Levenshtein distance[1] per segment and on average for both conditions (human translation and PEMT).

[1] https://en.wikipedia.org/wiki/Levenshtein_distance

- Aside from the total time per segment, the tool also delivers the typing time per segment. The difference between the two is the time needed for reading source and target and thinking about the changes to make to the MT output.
- The number of times the resource interrupted the test. Depending on the size of the actual test, a low number could indicate that interruptions were accepted while the test was running, which would point to less reliable figures.
- The actions taken by the resources. Especially pasting actions without a copy-action are relevant, as they could point to copying from an outside source, like a Translation Memory. If so, the purpose of the test would be defeated.
- An overview of the time spent for the segments. An example is given in Figure 2.

Figure 2. Times spent on all segments



The Q-values indicate what percentage of the segments was done in less time than this value. So in Figure 2, the first quartile of the segments was done in less than 0.5 minutes. Half of the segments was done in 0.7 minutes. The relevant bit is the far right of the figure. If there is a sharp angle upwards, it indicates that one or very few segments took far more time than the others. This may indicate that the resource was distracted, or that the segment in question was very difficult. Such segments require further attention and if the segment is deemed unreliable, it needs removing from the test. The tool will automatically recalculate all values.

4.3. How to use the ME results

The approach discussed above was designed to give a prediction of the productivity of PEMT for a project. The careful selection of the segments is meant to enhance representativity, while the ME tool will give precise numbers. However, the selection of segments remains just a spotcheck of the total number of segments in the project. The productivity change coming out of the test will not be replicated exactly on each and every job in the project, even if the overall productivity is likely to be similar. For this reason, it is recommended to interpret the productivity figures in bands. For example: a gain of 20%-40% indicates a decent productivity gain for the project and could therefore give an LSP grounds to reduce the rate to the vendor by a commensurate amount.

When vendors have been introduced to the testing process, and have taken part in some tests, they will be better able to interpret test results and any associated rate discounts. Depending on the local vendor market, this can save quite some overhead on discussions about how valuable MT will be in this case and what rates vendors are willing to accept.

Please note that the PEMT process is only part of the overall translation delivery process. While PEMT may increase the productivity of this one step, compared to human translation, it will not have any beneficial impact on overhead like the downloading of files or engineering and desktop publishing effort. Also the step of reviewing translations is not sped up as such. The LSP will need to assess for every individual project what impact the productivity increase in PEMT will have on the total project before deciding on any rate reduction.

5. Impact of the tool

SDL have found that using the two tools described above (select data tool and ME tool), the time spent on creating and analyzing tests was reduced to one-fifth of what it was when humans built the tests, while following the same selection guidelines. Of course, developing both tools came at a cost as well. For the amount of tests SDL processes, this cost was recouped in a year.

The main advantage is that tests can be built, sent out and analyzed within one working day. Of course, in real life there may be delays for practical reasons – without sufficient heads-up, resources may not be available on demand, and depending on the languages, some resources may live in different time zones. But the amount of work to be done within the LSP is restricted, and that allows a fast turnaround time for any tenders or sales opportunities requiring a fast response. In a highly competitive market, this offers a huge benefit, as the following example will illustrate.

5.1. Example of ME usage

In order to illustrate the point : SDL was asked to quote on a PEMT project. The deadline was a week. ME testing indicated that the existing off-the-shelf engine would not offer sufficient productivity gain to bid with reduced rates. SDL requested the client TMs and built a new engine, which was tested on ME as well. This engine offered sufficient productivity gain. SDL was able to offer reduced rates and return the tender within the given timeframe and won the bid. Thanks to the tool, we could run two tests as well as build a new engine within the week.

6. Conclusion

In a market where PEMT is increasingly becoming a standard part of translation workflows, LSPs need to be more and more aware of what they can quote for a particular PEMT project. This paper has described how the ME tool, and the processes around it, can help determining the future gain for a particular project in an acceptable timeframe, thus giving LSPs safer ground to quote to their clients and a firmer stand with their resources.

References

Allen, Jeffrey (2003). Post-editing. In *Computers and translation: a translator's guide*, edited by Harold Somers. Pages 297-317.

Gouadec, Daniel (2007). *Translation as a profession*. John Benjamins Publishing.

Läubli, Samuel, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk (2013): Assessing post-editing efficiency in a realistic translation environment. In *Proceedings of the 2nd Workshop on Post-editing Technology and Practice (WPTP),* pages 83–91.

Plitt, Mirko and Francois Masselot (2010): A Productivity Test of Statistical Machine Translation Post-Editing in A Typical Localisation Context. In *Prague Bulletin of Mathematical Linguistics, 9.* Pages :7–16.

Veale, Tony and Way, Andy (1997). Gaijin: A Bootstrapping, Template-driven Approach to Example-based MT. In *Proceedings of the Recent Advances in Natural Language Processing.*