Improving MT productivity within the cloud

Anna Rubina

Tremendous technological breakthroughs in the past two decades have revolutionized the translation industry. The creation of computer-aided translation (CAT) tools was only the beginning. The integration of translation memory (TM) and terminology management have vastly increased translators' productivity, but there is still much more to be done.

The emergence of machine translation (MT) technology and its integration into CAT tools has been a potential boon, but also a contentious issue. The topic of professional use of MT has been in the spotlight for the last several years. The proper implementation of MT technology is critical to ensure real gains in productivity. Without proper implementation, there is the potential to get bad fuzzy matches, which would simply not improve productivity. Perhaps more alarmingly, bad MT proposals at the segment level run the risk of negatively impacting productivity, since the translator will have to read through multiple variants to determine which one, if any, to choose. But now, cloud technology is changing the entire landscape of the translation process, unlocking great potential to increase productivity, quality and consistency.

So what do we hear from the production side? The two main concerns with MT implementation are how to measure productivity and how to assess MT quality. The first is often addressed by measuring the performance gain of a human working with MT compared to a human working alone. There are a variety of metrics for the second such as BLEU and METEOR. Nevertheless, the correlation between the automated MT quality metrics and post-editing effort is not clear. A poor metric does not necessarily imply that post-editing time will be high, and vice versa.

Anna Rubina is responsible for ABBYY Language Services activities in Europe, promoting language technology solutions in the translation and localization industry. She holds an MA in European regionalism from the University of Graz. Thanks to Artem Ukrainets, ABBYY's head of R&D, for his contribution.

The evidence clearly shows that integration of MT into the translation process increases productivity. The challenge we face is showing both translators and customers that MT technology is useful for them. Our solution to this challenge is simple – we want to make productivity metrics available in a cloud-based CAT tool, following our experience in integrating MT engines into the translation environment, as well as experiments in measuring the productivity of post-editing within this environment.

Experiments

We conducted some experiments to compare the performance of six different MT engines translating from English to Russian. We tested AsiaOnline, Bing, Microsoft Translator Hub, Google, ProMT and our own ABBYY MT. We tested performance in four different subjects: IT, oil and gas, law and energy. Texts in these subjects were translated by each MT engine and each generated translation was post-edited. The post-editing time was compared with the time spent translating the same text without MT assistance. We selected a team of 30 post-editors, linguists, developers and managers. Over 150,000 words were translated over a period of two months.

The maximum productivity gain across all topics was 80% and the minimum was 15%. The difference between two engines in one subject could reach up to 20%.

Up to 60% of post-editing time was spent correcting and checking terminology. Therefore, terminology management is a decisive factor in achieving good MT results. All participants of our experiment claimed that they spent most of their time searching for terms and editing them –sometimes over-editing them, as we discovered.

Our experiments vividly demonstrated that productivity depends on a large number of factors: the MT engines and settings used; the level of professionalism and personal qualities of post-editors; the source text, its topic, quality and volume; the desired quality of the target text; the translation environment; and other human factors.

Precise work time measurements and tracking the time spent on specific tasks by post-editors are absolutely necessary to obtain numbers that can be trusted. For the automated metrics, we used BLEU with several reference translations. It was established that the difference between two reference translations for the same MT engine is often higher than between two different engines for the same reference. This suggests that different engines should be selected for different subjects. However, the difference in numbers is often in the double digits, which makes BLEU rather inconsistent for evaluating translation quality for our purposes.

Thus we learned a good deal about the difficulties of post-edited MT measurements. We also made an important observation regarding MT's time-saving potential: with a properly customized engine and skilled post-editors, we managed to nearly double translation productivity compared to translation without MT assistance.

There are two important requirements for the methodology of such experiments. The first is that the experiments be conducted in the same environment in which translators work on a daily basis. Tracking time, editing effort and automated metrics are essential for understanding the effectiveness of post-editing. Nevertheless, the effective incorporation of MT into a translation work environment in a practical way should not be neglected. To avoid typical problems that occur when measuring post-editing time and effort, one needs to trust data (such as timesheets) received from post-editors, or alternatively perform measurements on large-scale projects, where the discrepancies of manual measurement are negligible.

The second requirement is that various MT engines be used and metrics be taken for each one of them. This is crucial to achieve objective results that can also be implemented in practice because it grants the opportunity to use the strengths of each system in translating texts in different formats or on different subjects.

Terminology is critical

As previously mentioned, we discovered that terminology is a decisive factor in achieving good MT results. All the post-editors participating in our experiment stated that they spent the majority of their time searching for and editing terms. Further analysis showed that sometimes post-editors even changed terms that had already been translated properly, meaning they did useless work that increased the overall post-editing time.

Terminology is always the most time-consuming part of the work in a project, especially large projects. To address this, the first essential step is extracting and translating terminology, thus creating a term base. This term base should be supported by the terminology management function of your translation tool. It serves as the skeleton and will support the whole body of translation.

Automatic quality checks (which should be present in the tool you are using if good results are to be achieved) would be sufficient to ensure correct terminology usage. We have discovered that it is extremely helpful to highlight correctly used terms to protect them from over-editing and communicate the correct terms to the post-editor or translator without taking up extra time.

We propose that such a dynamic approach would increase both productivity and consistency. The text for translation is first analyzed and the terminology is extracted, saving the context for the translator's reference. Then the terminology is incorporated into the MT customization parameters, and once the post-editing is done, the automated quality check-up validates the term translations. We suggest that this approach would work most efficiently in one integrated instrument in the cloud. This way, translators would have constant access to the system, in all stages: creating the term base, customizing the MT engine and post-editing.

Further improving MT productivity

We believe that an ideal translation platform should centralize all linguistic resources — source files, TMs, glossaries and MT — in one location. Furthermore, a selection of engines should be available from this platform that perform the translation in the cloud. The platform would thus display the different variants

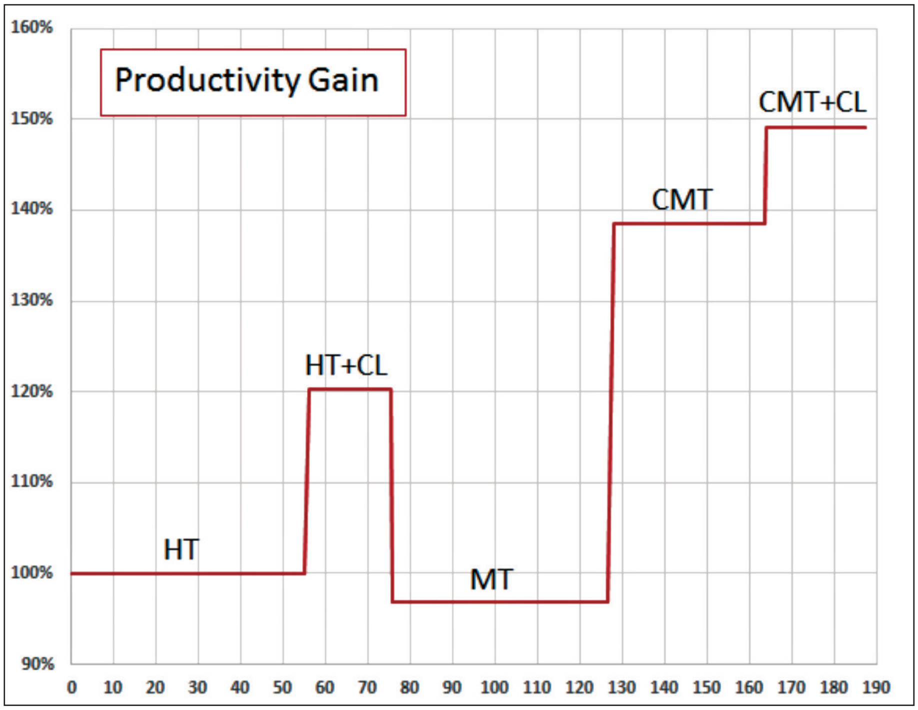

Figure 1: HT stands for human translation, CL stands for controlled language, and CMT stands for customized machine translation.

suggested by various MT engines for each segment. The cloud platform also allows multiple people to work on the same file at the same time to increase productivity. The fact that post-editor, editor and terminologist can work on the pool of translations simultaneously gives the system more flexibility and agility.

Different MT systems have different strengths and weaknesses. For example, a model-based engine delivers better results for longer phrases, while statistical machine translation (SMT) is more efficient for small or segmented sentences. This means that choosing only one system for a project is often a suboptimal solution. Furthermore, an MT engine's performance also depends on the topic of the text. Offering different MT outputs can further reduce post-editing effort, but we must be very careful in the implementation. Simply suggesting the results from three or even five different engines won't work, as the post-editor will inevitably have to spend more time choosing the best option rather than actually editing the text. This brings us to a specific procedure that needs to be established before the start of any MT project.

The first essential step is extracting and translating terminology, which is then used for MT customization and uploaded to the translation environment. Then, during the project setup phase (for a single large project or for continuous flow of small standard projects) a set of experiments is performed as follows: MT results from different engines are presented to trusted post-editors with good performance records. They post-edit the texts for a certain amount of time to get statistically relevant data within the CAT tool that will be used later in the production process. Post-editing preferences and post-editing time are tracked and compared for different engines. The best performing engines are selected to be used during the translation phase in order to pare down the choices the post-editor will face.

Along with measuring time, we also track the internal engine metrics. If the engine doesn't have a metric, we can use a target language model derived from the TM or other sources to evaluate the fluency of the translation. We then assess the correlation between these metrics and the actual post-editing time, and set a threshold, below which MT output from the engine is not fed to the CAT search. We also estimate the impact of terminology on a particular segment and run a glossary consistency check using advanced morphological technologies. If the machine translation for the segment doesn't match the glossary, it is not displayed, and only the term and its translation are shown in the CAT search. This allows us to reduce the work of post-editors when MT results are not helpful.

After the setup phase, the default translation option for each segment is determined based on fuzzy match percentage and the selected MT engines. Then the MT engines are rated based on the average productivity for each segment by assigning a corresponding fuzzy match percentage.

In carrying out post-editing experiments, we are able to measure productivity at the segment level. This makes the process transparent for post-editors, the customer and ourselves. It allows us to compare the productivity gain generated by each particular type of linguistic technology used. For example, see Figure 1 for the average productivity gain over time as compared to translation from scratch (where the X axis is the length of the experiment in minutes and the Y axis is the productivity gain compared to human translation from scratch). The format and subject of content were the same for all stages. This graphic could facilitate the negotiation of translation rates, and also supports the argument that this technology works.

Plans for further development

We are currently planning to introduce various advanced automated metrics into the process, as well as make them easy to understand and accessible to users. We are currently evaluating different approaches, including using a language model to estimate the complexity of the source text. Once this is implemented, we will be able to correlate automated metrics with the actual post-editing efforts at the segment level. Our goal is to bring the metrics from the domain of theoretical knowledge into the real translation environment, from experiments into real-life production. After the background calculations described above have been made, the system will suggest the best MT output available for a particular segment — or if all MT outputs are not good enough and it would be faster to translate manually, nothing will be suggested. We are transitioning from editing a pretranslated text in an Excel or RTF file with no other options or suggestions (such as other MT engines or a terminology base) to an agile, integrated, adaptable system designed to automatically maximize translation productivity in the cloud. M