Machine Translation in an Enterprise Environment

John Dixon
Strategic Solutions Director
Applied Language Solutions
Greater Manchester, UK
john.dixon@appliedlanguage.com

Abstract

This paper aims to give an insight into some of the challenges and opportunities from implementing machine translation in an enterprise environment. This is written from a business perspective rather than a technical one and highlights how Applied Language Solutions has designed and rolled out a series of customer specific machine translation solutions within our Enterprise.

1 Executive Summary

There are some things that never change, regardless of the type of project that an organization is trying to deliver. Whether it be localization, construction or a software project, the fundamentals are the same through and through. It is of critical importance to the success of any project that you have a very clear and agreed set of deliverables in place, have checked the quality of your source materials, ensured that your resources are trained and properly aligned to undertake their assignments and lastly that you have built in a review from which you can implement those lessons you have learned for the next project. In twenty years of working in industry in a variety of assignments and positions these factors still hold true.

The building, training and operation of the machine translation engine has been a significant but methodical part of the process we have gone through in the last six months One of the big hurdles we have addressed however is expectation management across all parties and the education that is required to better manage this. In addition, we have encountered a barrier when dealing with another human element in the process as quite often the linguists have been reluctant to change their traditional style of working because of their misconceptions and fears over machine translation becoming a direct replacement to human translation. Once this has been cracked then we have some great converts who will hopefully spread the word within their network

2 Requirements

One of the key issues to address at the outset of any plan to implement a machine translation system is to identify the business objectives and work out the best way to meet them. The reason for this is that quite often one size does not fit all and therefore it is essential to design a system and a supporting process that meets the requirements of the business, possibly with some compromises in certain areas on the recommendation of the vendor.

The requirement of Applied Language Solutions was to build a scalable machine translation system and process that could provide a customer specific engine at low cost which was able to cope with variable and constant demands but at the same time provide a solution that integrated with our other services to provide a "blended translation approach". Designing this solution meant that we would be able to offer the same customer a set of solutions that ranged from pure un-edited machine

translation all the way through to human translation with two proof reading stages incorporated to provide a quality and cost sensitive service to match the requirements of the of the end use of the documentation.

Due to the dynamic nature of our business we wanted a technology solution that was able to be scaled at rapid notice and interchangeable between customers and this has, so far, proved to be possible.

3   Design and deployment

As Applied Language Solutions was looking to implement a series of customer specific machine translation models we also looked at the surrounding processes to see how we could standardize as much as possible and at the same time use existing tried and tested processes. At the same time we wanted to create a solution that was as automated as possible and allowed a blended approach with human intervention at the appropriate stages in the process. Therefore we built our systems around our existing TalkBase project management system with connectivity to enable automated tools. This system has been extended to free machine translations offered through our website for limited personal use.

Applied Language Solutions decided at an early stage that we would use a virtual machine design with specific engines built for each customer. This meant that we would be able to treat each set up as vertically aligned deployment to each customer. Due to the fact that some customers have varying document/content types we then decided to take this a stage further and, where required, build content specific engines that would be specifically tuned to the subject matter and style of the content. This resulted in Applied Language Solutions developing and deploying multiple engines which are ideally suited to our proposed technical IT architecture. These engines can then be deployed upon demand automatically within a matter of minutes.

The fundamental approach that Applied Language Solutions took to the design of the processes was that, moving forward, it should require minimal IT resource input in terms of file engineers, etc. This led us to approach the design from an end-to-end business process model perspective that made us examine what, from our own existing systems, we could incorporate into the new design In addition we didn't want to fundamentally re-train our project managers from scratch to cope with a new set of processes.

Combining this with the blended approach to solving the localization challenge (through a combination of machine translation, post edited machine translation and human translation), we considered the use of machine translation as just an extension of CAT tools, which the project managers are already familiar with. We have taken the opportunity to "black box" the elements to a certain extent and this has been well received both internally and externally. Most stakeholders are not specifically interested in how it's done - they are more interested in getting the right quality within the right timescale on a consistent basis.

The design of our solution has meant that we able to deploy and on-demand an enterprise solution to a single customer. The customer no longer has to potentially develop their own solution which may sit idle from time to time as demand for the service varies across the business
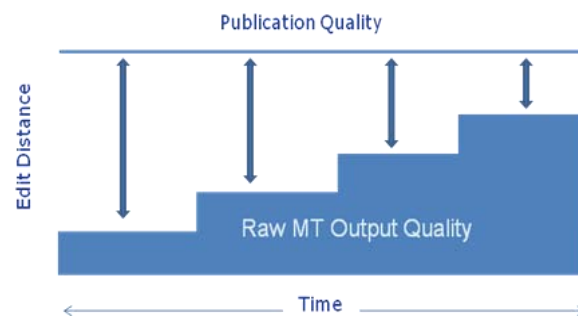
4   Challenges

4.1   Source material quality

One of the major challenges we have seen is the availability of a high quality, subject specific, translation memory and glossaries available. This, coupled with getting definitive style guides, has

meant that the ability to tune and educate the machine translation system has either been enhanced or limited with the availability of existing content from customers. Through our investigations and testing we have found that getting the best quality information up front means that we can provide better quality output. There is an old adage which holds true that "garbage in equals garbage out" and this is of major relevant when considering the implementation of an MT engine. Therefore one of tasks we undertake with each customer is a spot check on the quality of the translation memory provided.

4.2 Industry-relevant corpora

One challenge we have encountered is the quality of corpora available from industry sources. On one major project it took a week of concerted effort to remove poor quality data, including software strings etc. This has been a challenge and therefore will be more reliant on our own and customers data as we progress. As the demand for our PEMT solution increases however the in-house corporate build will become greater so in theory this element should improve significantly over time.

Another challenge from certain customers, who do not have a controlled or managed authoring system or process in place, is the quality of the original source files. This problem can manifest itself in poorer quality output as the machine has been trained on certain qualities of input and collateral. As with any quality issues we investigate the course and rectify the issues with process or training changes to improve the situation We have complimentary services which we are looking at developing further around controlled authoring, uses of simplified dictionaries and even source language proof reading. All of these services used in the right way can enhance the quality of the output by increasing the quality of the input.

Of course as most people will know, one of the major challenges that all developers of these systems face is the speed at which you can create the engines from scratch. This is one of the areas we are looking to address in future developments as time is of the essence for most enterprises.

Another old chestnut that was easy to predict was getting the linguists trained in post-editing and for experienced post-editors getting them trained in customer specific requirements. Amongst some linguists there was a reluctance to undertake this work as they thought of it has a threat to their current revenue stream. Once we had explained that this opened up another revenue opportunity for them and allowed them to take part in our education program regarding post editing some of the fears and anxieties were removed. We faced a similar resistance when translation memory was introduced many years ago and we therefore believe that the benefits need to be clearly communicated to linguists who can provide a solid editing resource as machine translation becomes more and more popular for context light material requiring translation.

4.3 Post edit impact and associated costs

I have described how the quality of existing translation memory and corpora has a major impact on the effort involved but to bring this to life it is worth highlighting the extend of the impact on our business and on relationships with post editors in relation to a recent project we undertook for a software application company. The organization in question required 1.3 million of words post edited after machine translation. We had issues with the machine translation and hence post editing in the first few phases. After investigation it was found that the quality of the translation memory was poor and in addition was not necessarily subject matter specific, despite assurances from the customer that it was appropriate to the material requiring machine translation.

As a result the tuning of the machine translation engine was not as effective as we had expected it to be. To resolve this we changed tack and trained the engine based upon the new collateral we were post editing, which provided improved output. In some cases however it still required a new human translation as the machine translation quality was so poor it could not be edited to a sufficient standard.

The outcome was that we invested additional internal costs in going through the process and undertaking investigatory reviews which cost us

approximately six man months in additional time. In addition our linguist costs were higher due to the amount of translation as opposed to post editing we had to do. We estimated this to be approximately $30,000 of additional costs. However, it also had a detrimental impact on the post editors who became frustrated with the initial quality of the machine translation and therefore we had to do a hearts and minds exercise to win back the trust.

It would have been naive of any vendor to assume that a significant time investment wouldn't be required up front however more detailed quality checks on the TM and on samples of the glossaries being supplied to the editors would have reduced this initial time investment which obviously equates to an additional labor cost that hadn't necessarily been budgeted for.

Longer term however the quality of the relevant TM available, combined with the ongoing tuning of the engine means that the edit distance needing to be covered by the editors will reduce significantly which will obviously make role of the post editor less labor intensive and more profitable.

5 Opportunities

One of the major opportunities presented to ALS and its customers is that we now have the technology and the process in place that shows that the customer has a trusted quality solution that has allowed an extension of the service offering into new areas of the business. This has allowed customers to consider translating content that was previously not translated for a number of reasons such as cost and speed of completion. This means that customers can now consider machine translation for reference material; user generated content, etc and then consider a more high quality solution for other high value content, such as marketing collateral.

Due the approach that we have taken, using specific engines built for specific customers and niche content matter and the use of virtual machines we have had two benefits. The first is that we do not need to deploy vast amounts of technology that could potentially sit idle between projects. The use

of virtual scalable machines has allowed Applied Language Solutions to adopt a cost benefit model of only having machines deployed when needed. It also means that due to our multiple engine models we do not have to undertake extensive re-tuning of a master engine each time we deploy the customer and content specific engine. This has significantly reduced the amount of input required from expensive IT resources. The use of niche engines has allowed us to fine tune the parameters probably more than a generic engine from other vendors. This means that the quality of product going to post-editors has increased.

These opportunities have enabled us to more effectively look at the return on investment that is available to use for future developments which are outlined in the section below.

With the introduction of our end to end integrated process we are now able to automate the pure machine translation of approximately 500,000 words per week across a number of projects and languages. Prior to the introduction of this automated system this would have required the input of approximately three full time equivalents to handle these pure machine translations projects which has now been saved. As this line of business increases, due to the steep upturn in customer demand that we are now experiencing, then our cost benefit realization will continue to grow over the coming months and years.

6 Stakeholder and Expectation Management

As with any project, the main issue you may encounter is the management of expectations of individuals or teams involved in the outcome. This is because they may all have different ideas of what the solution is delivering. In order to manage this effectively Applied Language Solutions has undertaken a series of communication strategies with customers in order to make sure they know exactly what is and is not being delivered at each stage. We also educate them on how the blended approach to the use of machine translation means that we are picking (or they are picking) the most ap-

propriate solution for each content set. Despite this communication during the planning phase of each project we have undertaken, there have been cases where the outcome still doesn't sit in line with the expectation in the customer's mind. This repeat education is something that we expect to have to continue with for some time until there is a greater general understanding of what can be achieved via machine translation.

With Google more publically promoting its translation tools we have a number of customers that may have had good or bad experiences. Our philosophy is to create a solution that is very much tailored to a specific content and therefore our overall solution provides and excellent fit with the needs of that content type.

We work very closely with the customer to engage them in defining the rules that we will use in the blended approach that will decide the method of processing for each content type. In addition, we will then modify the processes with information from the quality assurance feedback loop that will ensure that the process and machine translation engine/rule set is continually being updated with dynamic factual input.

7   Next Stage of Development

Now we have undertaken a number of exercises both for external customers and internally with major test projects we have identified a number of potential improvements that we are looking to develop in the coming months.

Firstly we are looking to fully integrate our solution through our TalkBase system, with customer based CMS and TMS systems. This should make the solution more real-time and reduce the amount of time needed from a project manager to manually handle content and documents. This should also make the process more seamless and for the customer, potentially reducing the amount of time required by the customer.

Secondly, we are looking to develop a pre-editing solution which will identify any issues with the source text that might impact on the quality of the

target content. This will allow for checks against controlled authoring, glossary and style guide parameters specified by the customer.

The third development stream will be to look at using the machine translation tools to automatically identify the type of content of the document and make an automated recommendation of how it should be processed. As an example, if the machine identifies the document as a piece of marketing text then the process rules would send this content for translation by the appropriate human linguist and any further processing. If the document was a user manual or reference document then this would be sent to a specific machine translation system which would then process it and then identify a post- editor for further processing.

The fourth stream of potential development is possibly the most exciting of all in terms of the dynamic tuning of the engine by gaining real-time feedback from the post editors. Applied Language Solutions is currently looking at how we integrate our online in context review tool for use in a post editing environment with a link to the tuning parameters of the niche engines. On large projects with multiple concurrent editors and proof readers we are examining the possibility of how changes made by post editors can be fed back to the tuning process in a dynamic fashion to then update the engine, tune it and make those changes in real-time available to other post-editors. Applied Language Solutions considers this is vital on large content projects where due to the volume of content and the timescales, multiple post editors have had to share the work.

The acceptance of machine translation by the language industry is something that we need to encourage as it presents fantastic growth opportunities for vendors and allows greater content translation volumes for organizations worldwide. Communication, however, will be key as we move forward. The limitations and the opportunities of post edited machine translation need to be fully explained, fully understood and, more importantly, reiterated to all stakeholders so that expectations around return on investment and quality will always be met.