

A decorative graphic consisting of three blue circles of varying sizes, each with a lighter blue outer ring and a darker blue inner circle. The circles are arranged in a vertical line, with the largest at the top, a medium one in the middle, and a large one at the bottom right. Thin blue lines connect the top-left corners of the circles, creating a diagonal path across the page. The entire graphic is set against a white background with a thin orange border.

# Machine Translation: The Future or a Fallacy?

With a growing mountain of content being created within organizations and increased globalization, what is the best strategy to get this content translated and does Machine Translation finally hold the key?

**Gavin Wheeldon - CEO  
June 2010**

### **Introduction**

More for less is the cry! It is a common call across departments in organizations all over the world and it's no different in the translation arena. Couple this with the demands for immediacy that are now common place and the expectations that mountains of content can be translated within limited budgets - they want translation done quicker than ever before. So could Machine Translation finally hold the key to meeting these expectations?

### **A Short History**

Machine Translation officially established itself during the 1954 experiment carried out by Georgetown University and IBM, which undertook small numbers of translation between one language pair; Russian and English. Despite being based on a very small volume of sentences the experiment was nevertheless considered to be a success, generated lots of publicity and, on the back of that, secured Government funding. It was stated that, within three to five years, the machine translation problem would be cracked. There was also a similar experiment in the same year by Birkbeck College at the University of London using English to French as the test language pair which produced similar results.

Over ten years later in 1966 a US Government report from the Automatic Language Processing Advisory Committee (ALPAC) deemed the results to be much less positive than expected and therefore funding was greatly reduced. Momentum did not pick up again until the 1980's when interest was taken in Statistical Models for Translation at IBM, which was born on the back of the increased computational power available as Statistical Machine Translation, even today, is incredibly processor-resource hungry. This interest led to continued investment in rules based translation, particularly in Japan.

In the 1990's machine translation was freed from the mainframes and servers and found a home initially on personal computers and then on web-based services such as Alta Vista's Babel Fish.

In more recent years the majority of interest and research has been around Statistical and Example based models. Lots of interest has been driven by the Google Translate system and the success that this has shown.

### Rules Based vs. Statistical Translation

There are many different approaches to machine translation and the ultimate aim of them all is to take one language to another in an automated fashion. The two systems that have delivered the best results to date and are therefore deemed successful are Rules Based Machine Translation (RbMT) and Statistical Machine Translation (SMT) although there has also been some work and success around Example Based Machine Translation (EbMT) which is similar, in some ways, to Translation Memory.

If you're looking for successful implementations then you are much more likely to find them for Rules based Machine Translation (RbMT), simply due to the fact that it has been around since 1954 rather than the fact that it is necessarily a better system. It is also worth noting that RbMT does, at the moment, present better results in some cases for certain language pairs. Although if the level of corpora available for that language pair is high enough it is likely that equal or better results can be achieved using SMT.

**RbMT** – Rules based Machine Translation is, as the name suggests, an engine built on rules. It comprises a dictionary and a set of rules around that dictionary to construct sentences in a target language. The time and cost involved in setting up or customizing an RbMT system can be extremely high and prohibitive (which is why there are relatively few language pairs available) for most businesses, along with maintenance and tuning which can add large layers of additional cost and time. The other major disadvantages are that language doesn't follow a set of rules and it is not possible to incorporate context as the system looks exclusively at one system. Two key advantages I do see, however, are that there isn't a large computational resource required to train a model and the results that you get out are predictable.

**SMT** – Statistical Machine Translation learns from data both bilingual and monolingual to understand how to construct the target languages. It does this by looking statistically at the data and determining the most likely match, based on probability. The more data you have the better it can learn, although equally you are better off with a tighter set of high quality data rather than lots of 'dirty' data. The computational power in building, training and tuning a language model is high with some serious 'fire power' needed to do it in any sensible, commercially viable, timeframe. It is also a fairly complex process with many different options to fine tune output, so real skill and understanding of the tools is required. The massive advantage is of course that any language pair with enough corpora can be tackled in a relatively small timeframe and the output, when built for individual requirements, can be of a very high standard.

**Hybrid** – So, is there a case for a hybrid approach to the problem combining the best of both? There have been implementations where SMT is used to smooth the output from an RbMT engine although. I am not sure how much benefit this offers over SMT on its own if you are starting from scratch although if a legacy RbMT engine is in place with the investment already done to tune this then it may be a good approach. There is however a strong case for borrowing the idea of rules to improve the level of output from an SMT system. This is different in that the rules would look at the text, both source and target, in an isolated fashion to see how it can be cleaned up. For source text, rules can be written to get it into the optimum structure for MT and output rules can be built to correct common problems and things such as date format, currency symbols etc. It is also possible to build rules using Natural Language Processing to apply key concepts from style guides for a target language.

### **Broad Based vs. Tightly Defined Engines**

Most people will be familiar with the broad based/general engines freely available on the web and most notably the Google Translate engine. Although these have always been the butt of many amusing jokes of what the systems produce when back translated, they are getting better, again most notably Google and do serve a purpose. Google is an SMT engine and is built upon billions of segments of corpora, mainly from scraping commonly available translations across the web and therefore has a very general view of the world of language.

Tightly defined engines and built for a specific industry, publisher or project and therefore have a limited and in theory higher quality and better targeted corpora for a given purpose. To build any kind of reasonable engine there needs to be a minimum really of 1Million words, preferably 2Million although the ideal is nearer 10Million with anything else a bonus. If there isn't that level of data available there are ways of getting corpora by crawling niche sites and using memory sharing from places such as the TAUS Data Association.

So where would you make use of the different types of engine. Well quite simply it would really depend on the content. If you have internally written documents around manuals or any other content appropriate for MT then the tightly defined engine is the obvious choice at all times. The area where a general engine could add value is on customer support or any area where User Generated Content is being written in a very generalist way as the broader engine is much more likely to get a good understanding of it. You may in these scenarios want to mix the general engine with pre applied

glossaries which would pick up any product or other specific references to achieve a good balance.

Overall a content strategy should be put in to make sure all content flows in the optimal way and this may involve having more than one tightly defined engine, translation memories and use of a broader engine where appropriate.

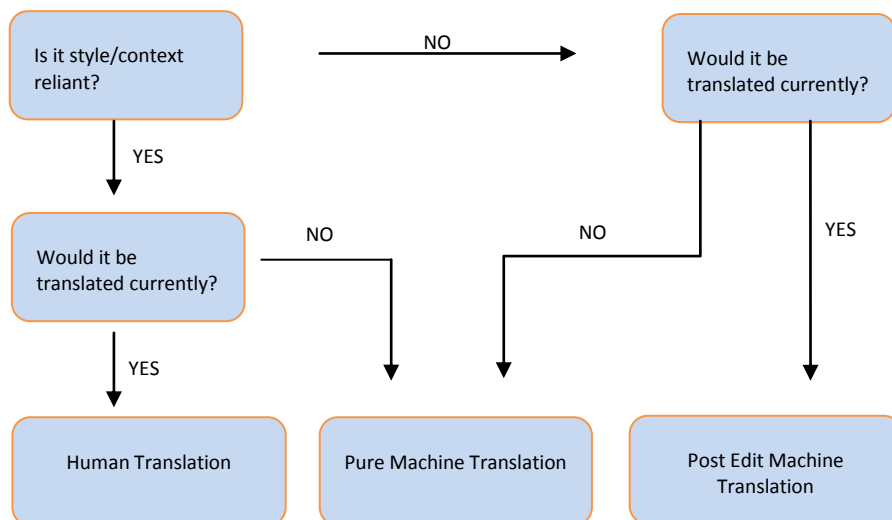
### **Content Types**

There is what we have termed a 'content mountain' in existence both generally and within most enterprises around the world. The rate at which content is generated by employees, customers and stakeholder parties is so great that it is currently impossible to have anything but the tip translated. It is estimated, in fact, that less than one per cent of content in the world ever leaves source language.

Context light or non style-reliant content, such as manuals, FAQs, knowledge bases, software strings and repetitive product descriptions, such as hotel listings, are ideal for machine translation. Equally, anything else that would never get the opportunity to be translated is an ideal candidate: something is better than nothing. Microsoft has a mixture of human and machine translation in their knowledge base articles, all of which have a yes/no answer for whether it solved the problem. They get exactly the same number of yes answers on MT vs. Human Translation.

As to whether or not MT is a complete solution, I don't believe we are there yet and I think we are still a long way from the 100 per cent accurate sci-fi instant translator in all languages which would be the ideal. I believe overall there should be a blended approach to the content mountain to determine which approach is most suitable for each piece of content.

It is worth pointing out in figure 1 below that the flow shows that style reliant content that is not being translated currently could go to pure MT, the ideal is that budget saved elsewhere could be used to flow this through human or Post Edited Machine Translation. If that was not possible then serious consideration should be given as to whether this may cause brand damage and therefore nothing is actually better than something. Another option may be to clearly point out that the content has been machine translated and therefore comes with a clear health warning.



**Figure 1**

If you follow this model in its purest form you will get a lot more content translated for a lot less budget. What I believe many organizations will do is start looking at what content is flowing to the pure machine translation box and use budget to flow that to either Post Editing or Human translation, therefore getting huge amounts more content translated for the same or less budget if the overall saving outweighs the cost of the additional translation/PEMT.

## What about Quality, what is it anyway?

This is probably THE biggest question that surrounds MT and will ultimately dictate its success or failure in any organization. If people are setting the bar at the level of human translation then ultimately it will fail as currently raw output, even in a custom engine, is a long way from that ideal.

I think putting the words quality and machine translation in one sentence would shiver the bones of many a linguist. What needs to be realised is that if MT is really going to come of age then the two words need to become synonymous and that is a large leap from the recent past. I also believe that it is more a case of setting appropriate expectations of all involved.

I guess this begs a second question of what is quality or more importantly what is 'good enough'?

In my opinion 'good enough' quality is if it achieves the business purpose of the content. Every piece of content is there to do a specific job and the question is what does it need to be to do that job? Put another way you could argue that it is the emotional connection

## Machine Translation: Future or Fallacy?



that the content creates and what that needs to be based, again, on the objective of that content.

If you are attempting to sell something and building trust, brand image or trying to create a conversion into a sale then maximum emotional impact is essential. This would be true of many types of text such as press releases, marketing material, investor relations and many other examples of source content. If an individual, who is about to read the content, is likely to have a level of scepticism or needs 'converting' or 'selling' to, then achieving human level translation is well worth the investment as that person needs to feel as though that text has been written for them and for that market. They need an emotional connection that tips them in the direction you want them to go wherever they have 'choice'.

Once somebody has made the emotional jump and 'choice' is not part of the content objective then often meaning is the primary purpose and people have switched off their critical judgement mode and would either not notice, or be forgiving of a lack of style. In this scenario having MT post edited to a level where it is factually and contextually accurate with grammar and spelling correct should more than serve the end purpose.

*"In a recent usability study conducted in Germany, Dell observed that tech and internet savvy buyers were unperturbed by small translation errors, and in many cases did not even recognize them until they were pointed out. Their focus was on locating and purchasing the product. On the other hand, non-tech savvy buyers who needed to form an emotional connection as part of the purchasing process were both distracted and disappointed by translation errors."* Wayne Bourland, Snr Manager, Global Localization Team

If content is 'value add', user generated or it is something that wouldn't have been there before then most times anything is better than nothing. I say that with one big caveat and that is that it should still convey the correct meaning. If style, grammar and spelling are wrong that is forgivable but if the meaning has been lost then you would be better with nothing as the content has not achieved the objective and even worse may damage the brand. For this type of content with a good engine then you will achieve the objective, if the engine is not of a sufficient standard or in the early days you may consider a light post edit to just check that meaning is conveyed properly.

I'll give an example of the process and types of content here. I am looking to buy a new laptop and I am therefore investigating the different brands and options available to me. I already have a laptop from brand X and am happy with it but I have gone back into critical judgement phase so brand X still needs to impress. Anything less than human level translation is likely to lose me at this point on any style influenced text. I've now chosen my laptop and open the package when it arrives and my purpose is to get it set up and working, whatever text I read should serve this purpose in a clear and correct manner and style, if there, that I probably won't even notice, so post edited MT should be perfect. From here I go onto the forums and there are some tips from other users who have fine-tuned the laptop and I figure out how to do it myself from something a French user has written – this is a bonus as I never would have figured that out without the pure MT.

Now there are also many more moving parts of to determine whether or not quality is 'good enough' and, if I was a mathematician, I'd write a really impressive formula which highlighted exactly what type of translation should be used based on the many moving parts. Other things that might be considered are factors such as if you are not in a market is any translation better than none as you may get some sales, or is this going to damage brand too much? At what point does the ROI on any given content outweigh the cost of any given method, and are there any multipliers of using for example human translation over post editing i.e. would the level of sales if human translated justify the extra investment? That formula may be one for another paper.

A major content publisher once joked with me that the more linguists he removed from the quality process in MT, the better the quality got. I think this gets back to the problem of measuring results against human translation as opposed to whether the content is 'good enough' and this is the biggest question.

### **What is the difference between Translation and Post Editing?**

I've been asked many times who does the post editing and is it somebody different to the normal translators that we work with. The answer is that it could be but it isn't as they are by far the best people to do the work. I think people in this industry sometimes want to change everything because they think it should be, rather than for any real logic. There is lots of talk around whether post editing should/could be done by native language speakers of the target language who have no knowledge of the source text. I would ask a simple question – why? We have a massive pool of highly talented linguists who merely need to have a shift in mind set. And I still believe that an understanding of

the source text is essential. In fact the only difference I see between translation and post editing is that the skill is different but not something that can't be learned quickly by a trained linguist. .

Another common question is what training a post editor needs over a translator, if any? I believe that there are distinct differences between the two and there is a definite education process for a translator to understand what is expected from them rather than an education in skill set. It is often more about what not to do, rather than what to do and this is one of the biggest reasons that there is such resistance in the translator community, and quite honestly goes back to the question of what is 'good enough'. If a translator spends their time trying to get the quality up to the same standards of their full translation work then it is a bad post edit and understandably they will question what the point is. The vast majority of the work we do in training on this skill is about understanding what is expected and how to make the decisions on any given segment as to whether or not that has been achieved.

### **Light Edit vs Heavy Edit**

Standards and best practice are settling around two versions of post editing and it is worth understanding the difference between the two. Broadly speaking a light edit will be correct and understandable information, whereas a heavy edit should achieve the same results as the translation element in a more traditional all human approach. If you add a proofread to heavy edit you should expect as good, if not better quality than you would from a human translation process. In this scenario it is simply a productivity tool similar to Translation Memory.

The following was produced by TAUS with Sharon O'Brien from Dublin City University and CNGL and I think are a good guide.

### **Light Edit aka "good enough" Quality**

"Good enough" is defined as comprehensible (i.e. you can understand the main content of the message), accurate (i.e. it communicates the same meaning as the source text), but as not being stylistically compelling. The text may sound like it was generated by a computer, syntax might be somewhat unusual, grammar may not be perfect but the message is accurate.

- Aim for semantically correct translation.
- Ensure that no information has been accidentally added or omitted.
- Edit any offensive, inappropriate or culturally unacceptable content.

## Machine Translation: Future or Fallacy?



- Use as much of the raw MT output as possible.
- Basic rules regarding spelling apply.
- No need to implement corrections that are of a stylistic nature only.
- No need to restructure sentences solely to improve the natural flow of the text.

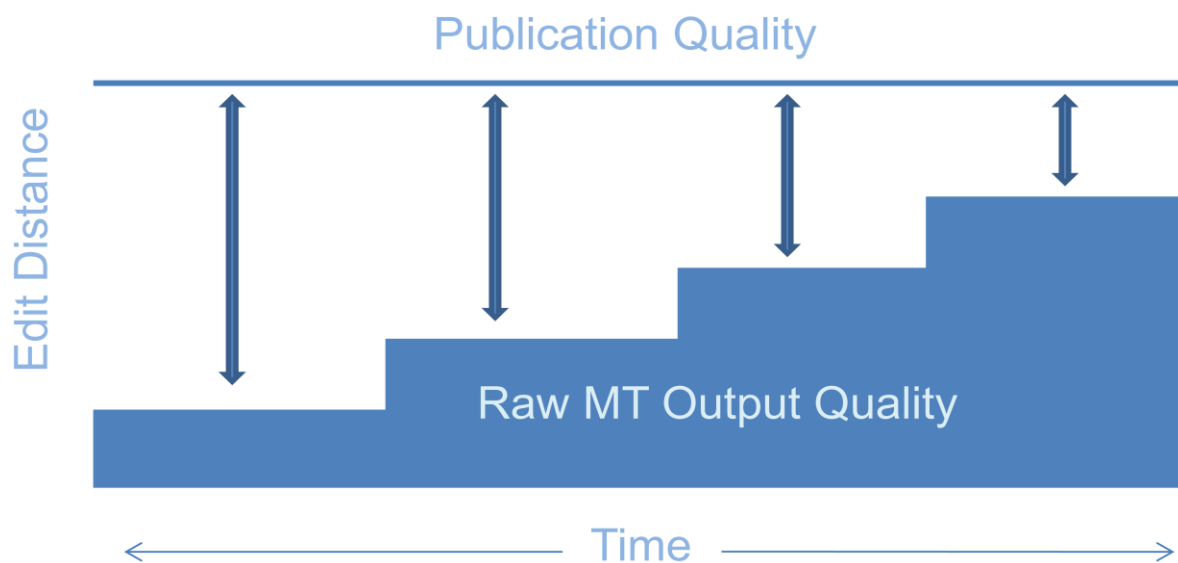
### Heavy Edit aka “human equivalent” Quality

This level of quality is generally defined as being comprehensible (i.e. an end user perfectly understands the content of the message), accurate (i.e. it communicates the same meaning as the source text), stylistically fine, though the style may not be as good as that achieved by a native-speaker human translator. Syntax is normal, grammar and punctuation are correct.

- Aim for grammatically, syntactically and semantically correct translation.
- Ensure that key terminology is correctly translated and that untranslated terms belong to the client’s list of “Do Not Translate” terms”.
- Ensure that no information has been accidentally added or omitted.
- Edit any offensive, inappropriate or culturally unacceptable content.
- Use as much of the raw MT output as possible.
- Basic rules regarding spelling, punctuation and hyphenation apply.
- Ensure that formatting is correct.

### How is MT and Post Editing Priced?

It is first worth understanding what you can expect from investing in MT and what the landscape will look like over time. The reason it is worth outlining this is that many people are very nervous about putting any kind of pricing around both MT and Post Editing because they are afraid of being caught out.



**Figure 2**

As the diagram in figure 2 demonstrates, the raw MT output should, over time, increase in quality and therefore the amount of work (edit distance) involved in getting it to publication, or 'good enough' quality will reduce. This should be achieved by many different methods of retraining the engine and improving the source and output text before it arrives on an editor's screen.

Due to this variable return in benefit the pricing models being discussed are varied to the extreme. Most language vendors are trying to figure out how to price the benefits as they are achieved, and this uncertainty adds to the reluctance of publishers to take this up as there are no guaranteed benefits.

My opinion is that for this to move forward vendors need to be bold and offer all the benefits of the end goal up front. Yes this will hurt the pockets of many to begin with but will at the same time develop strong, trusted partnerships that over the long term will ensure solid revenue streams from a mutually beneficial partnership. I know few, if any, vendors within the translation industry who will agree and I can see where their logic would come from with this when looking at cold hard facts of margin profitability. Another option where publishers are amenable is to had specific milestones agreed, time or otherwise where step changes in discount levels are agreed which may work better for both.

### **Conclusion**

## Machine Translation: Future or Fallacy?



I think the translation industry is currently at a tipping point, much as it was with the advent of Translation Memory in the mid 1990s. Machine Translation is very much the future and we will witness this in greater and greater volumes over the coming years. There is no doubt that MT has been the expected golden child of computing for many years and continued to disappoint. I think technology has made serious progress, however, over the past few years and, with appropriate expectation levels, will find a very clear part in the broad mix of language output across enterprises and in general in society.

I am very excited about what is happening in this area both within Applied Language Solutions and in the wider academic and technology arena in general. If you have large amounts of content or content that is left un-translated then this is something that should be seriously considered. The corpses of MT trials can be counted in the hundreds, if not thousands and they have all no doubt been very, very expensive. I believe we are now at the point with technology where a serious trial can be conducted for no more than a few thousand pounds/dollars/euro and should yield solid results.

So in answer to the question; future or fallacy? Definitely the future!



[enquiries@appliedlanguage.com](mailto:enquiries@appliedlanguage.com)

Tel: +44 (0) 845 367 7000 - UK

Tel: +1 (800) 579 5010 - USA

[www.appliedlanguage.com](http://www.appliedlanguage.com)