Machine Translation engine selection in the Enterprise

Heidi Depraetere
Cross Language NV
Woodrow Wilsonplein 7
9000 Gent, Belgium
heidi.depraetere@crosslang.com

Pablo Vazquez
Cisco Systems, Inc.
150 West Tasman Drive
San Jose, CA USA 95134
pavazque@cisco.com

Abstract

The focus of this article is to show how client centric Machine Translation (MT) evaluation can assist in identifying the best MT solution in an Enterprise context.

Examining extensive lists of opaque scores (BLEU and others) does not bring value to the Enterprise market unless it can equate the results with its own daily commercial challenges.

Armed with the right set of directives the decision process is made simpler as Cisco recently discovered. From the results of an external evaluation program they were able to identify a clear process for engine selection as well as creating a matrix for future needs.

1 Introduction

Enterprise evaluation of MT is often assigned to people and teams that have little knowledge or experience of how MT is deployed in their busines environment. This paper aims at showing how client centred evaluation leads to defining minimum acceptance criteria for MT engine selection purposes.

The findings are based on quantitative and qualitative data gathered during an evaluation process carried out for Cisco, by an external partner.

The issue of return of investment and where MT can be used is outside the scope of this paper. These subjects although very important and critical for the adoption of a technology like MT should be done prior to the engine selection and may be the subject for another paper.

The starting point of this article is the Enterprise with a clearly identified MT application context including content to be translated as well as an explicit MT workflow.

2 Parameters

Clearly one needs to evaluate the potential of a new technology and understand what it can do for the Enterprise before one considers using it.

MT is no different; except that once one goes through this process and decide that it can bring a benefit to the Enterprise one is then faced with an additional challenge: identifying which MT system(s) on offer will best meet one's specific needs. This is well documented by (Lehrberger & Bourbeau, 1988): [1] "The objective of an evaluation is of course to determine whether a system permits an adequate response to given needs and constraints."

It is therefore essential to define a set of parameters which will allow the main stakeholder and potential users in the Enterprise to make an objective business decision.

3 Defining Cisco's needs

Cisco has been using MT for translating support content for nearly a decade. They translate technical documentation and troubleshooting guides to aid their customers and partners in solving problems without the need to directly contact Customer Support.

All the material that has been tagged translatable is currently published in Spanish, Portuguese, Russian, Japanese and Chinese. Before the launch of each new website the content is

[1] /Lehrberger/ J., /Bourbeau/ L. 1988. Machine Translation: Linguistic Characteristic of MT. Systems and General Methodology of Evaluation, John Benjamins *...*

processed through several highly customized MT systems that have been updated and customized with current enhancements and assets.

However, as new languages were added it became clear to Cisco that they needed to create a process to select specific engines. It also became apparent that linguistic quality, as important as it is, is not the only factor in a successful deployment of an MT engine.

The first step was to create a “master plan” that allowed Cisco to evaluate several MT engines and compare the features. For this a strict definition of their needs and a review of the most important features were required.

Eventually they came up with an evaluation process. The goal of the evaluation process was to select an engine for a new language and perhaps more importantly, establish a systematic approach for future and further engine selection.

After reviewing the past issues and anticipated problems they came up with a final list of six (6) master criteria.

The decision process was made easier than expected when they analysed the output of the test engines using this criteria.

Here are the descriptions of the six (6) categories:

3.1 Translation quality

Though translation quality is only one consideration in the decision process to implement MT, for most Enterprises it is the both the most important and most difficult criterion to value. End users will always set their own expectation levels for the quality of the MT output and judge it in their given context.

It is normal practice that when we plan to measure something we first try to find something to measure it against, identifying a “gold standard” [2] (Hovy et al, 2002). The same thinking process can be applied to measuring the quality of the MT output but as “perfect” translation is an elusive concept a different set of rules will need to be applied. Exact quality measurement of MT output is very difficult as the context is full of variables which need to be understood before any evaluation program has a chance of success.

This leads to in-context quality evaluation. It is important before starting any examination to ensure that expectation levels of both end users and stakeholders within the Enterprise are set and aligned correctly.

Regardless of the application context of MT a first step within the translation quality assessment will always involve a human linguistic evaluation as this initial measurement will be a good indicator for more focused application specific evaluation. MT vendors have a tendency of referring to output in terms of BLEU (Papineni et al., 2002) [3] scores. However meaningful these computed scores may be in a development environment they are very hard to relate to in an Enterprise user environment. Stakeholders and users want to find out whether a translation is accurate enough to be understood or good enough to be post-edited.

In a translation production environment the Enterprise wants to find out in how far MT will speed up their localization process. A productivity assessment will demonstrate productivity increase estimates and their potential associated per word cost savings for translating new words. A post-editing evaluation will expose how much effort it takes to correct the MT output and will showcase the typology of corrections.

In a support context the Enterprise may want to establish whether MT can help users solve their problems. Can MT enable high-level English speaking support specialists provide solutions for their local support engineering colleagues. A usability assessment involving active participation of the targeted end users will shed light upon this. User surveys are set up where end users are given comprehensibility tasks to perform.

As a conclusion it is given that the quality of the translation will always influence how usable the MT output is for the intended usage. However instead of linguistic evaluation being the sole metric used to determine this, the application context plays a major part in the calculation.

3.2 Customization capabilities

Why does Cisco need a high level of customization? Generally most applications for MT have a lower customization factor. However, when Cisco defined their workflows a conscious decision was taken to position the MT engines at the center of the translation and generation process rather than just as a plug-in of the CMS systems.

[2] Hovy et al, 2002. Principles of Context-Based Machine Translation Evaluation. Machine Translation, 16, pp. 1-33. Springer.

[3] Papineni et al. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 311-318.

This particular set-up assumes a high quality MT output and as a consequence very demanding efforts on customization. Hence, customization is very extensive including customization of the assets such as Translation Memories, dictionaries, glossaries as well as linguistic rules. These assets are the key to improve the quality of the final translation output.

If deployed correctly with mature engines this process does not usually require post-editing, or if it is then the post-editing effort involved will be minimal.

3.3 Standard format support

Due to the highly customized workflow described above as well as the continuous MT technology changes Cisco needs to have the flexibility to change the MT engines.

In such a scenario it will be important for Cisco to be in a position to get the assets, dictionary customization, TMs, and even linguistic rules in and out of the various engines. Portability from one system to the next is critical as Cisco does not use a specific engine but applies a workflow driven approach.

LISA standards (www.lisa.org) such as TMX, TBX and SRX, are good examples and Cisco "forces" their engines to follow those standards.

3.4 Integration potential

In the past MT vendors often considered their technology as a unique component operating stand-alone and as black boxes.

However, in the past years this situation has changed dramatically due mainly to competition from emerging vendors, MT technology additions and a better general awareness of MT technology. The existence of open source engines gave the Enterprise new "leverage" with the MT engines vendors and providers.

As MT activities grow in the Enterprise and start playing an important role in its operations the Enterprise has added MT experts and lexicographers with industry knowledge to their workforce.

As a result the Enterprise infrastructure now adopts a multi-engine approach as it is reluctant to rely on one vendor or technology for all its needs. This requires MT engines that are able to connect with other environments and repositories.

Open API's are no longer a "nice to have" but instead they are a "hard requirement". For enterprise implementation black boxes are no longer an option.

3.5 Scalability

On the lifecycle of the MT engine there is a critical point when it is ready to go to production.

At this point it is vital that the MT system can perform to the expected level.

Cisco had experiences with some "supposedly" fast engines that when highly customized and loaded with large amounts of data become slow and moreover unstable – in some cases even unusable.

If the MT engines are not able to translate, they have memory leaks or other impediments that impact on the speed or the throughput of the engine. It is vital that these issues are assessed and addressed prior to the adoption phase.

Another common issue is the scaling capability of the MT engines – by adding processors, speed, clustering, cloud type environments, etc. it may become a problem at a critical moment of the deployment if not tested correctly and discovered in advance.

To avoid these problems Cisco tests the candidate engines in several scenarios that mirror both the production environment and the assets loads. Stress tests are a must when looking at a new MT engine, but are sometimes very difficult to simulate.

3.6 Cost

Enterprises very often look at the monthly cost as being the only investment in a license or a service charge in hosted set ups for commercial MT systems.

Initial customization budgets are very often underestimated because little attention is given to what it takes to lift customization up a level from a pilot context to a deployment level. It is important to gauge correctly the efforts from a small-scale test environment to a large-scale production environment.

It is easy to see why Enterprises make this mistake as MT vendors often sell up the fact that their system is easily customized but play down the complexity and cost in doing it.

Customization maintenance is another element which very often is understated. Enterprises need to budget for maintaining the output levels of their MT engine(s) through professional services or internal resources.

Are upgrades included in the offering or are they additional cost elements?

In summary, at your peril underestimate the importance of transparency in cost assessment.

An alternative approach is the very active Open Source market which provides the attractive benefit of no licence or commercial fee.

However, there is no such thing as a free MT engine – even Open Source alternatives such as Moses[4] have a cost factor. Assuming that you have the right resources available they need development and customisation to work correctly. And once working they need the same regular support and maintenance as propietary systems do – in some cases even more so.

However, once they are functioning the technology is yours for the keeping and the knowledge levels built up remain in-house for further projects.

4 The matrix

Having established a matrix with evaluation categories Cisco measures the performance of an engine by applying minimum acceptance criteria for each category.

Once each category is within accepted levels test and scores are assigned for each one of the candidates.

This created a transparent and easy to replicate process where the rules are clear to the vendor as well as for Cisco. This approach is similar to the "user data chart method" described by Nagao at an AMTA panel discussion (Muriel Vasconcellos 1994). [5]

5 Conclusion/summary

A comprehensive in-context evaluation scheme is a pre-requisite for engine selection.

Choosing an MT solution is more than just a quick assessment of translation quality. Defining an evaluation matrix and applying it for any new target language will ensure building an adequate MT solution that fulfils the Enterprise particular needs.

Whereas in the past Enterprises were fully relying on MT vendors they are now taking more control of the decision process as they have come to realize that there is no one MT engine or technology that fits all their needs.

The Enterprise has always felt more comfortable when buying and implementing new technologies to base their decisions on solid researched information and sound business case assessments – then they will make investments and fully embrace the technology.

References

Hovy et al, 2002. *Principles of Context-Based Machine Translation Evaluation.* Machine Translation, 16, pp. 1-33. Springer.

Koehn et al. 2007 *Moses: Open Source Toolkit for Statistical Machine Translation*. ACL demonstration session, p. 177-180

Lehrberger J. and Bourbeau L. 1988. *Machine Translation: Linguistic Characteristic of MT. Systems and General Methodology of Evaluation.* John Benjamins.

Papineni et al. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation.* Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

Vasconcellos Muriel(ed.). 1994. *MT evaluation: basis for future directions.* Proceedings of a workshop sponsored by the National Science Foundation, 2-3 November 1992, San Diego, California. Washington, DC: Association for Machine Translation in the Americas.

[4] Koehn et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. ACL demonstration session, p. 177-180.

[5] Muriel Vasconcellos (ed.). 1994. MT evaluation: basis for future directions. Proceedings of a workshop sponsored by the National Science Foundation, 2-3 November 1992, San Diego, California. (Washington, DC: Association for Machine Translation in the Americas, 1994)