Can Quality Metrics Become the Drivers of Machine Translation Uptake?
An Industry Perspective.

Katrin Marheinecke

text&form
Berlin
E-mail: katrin_marheinecke@textform.com

Abstract

Language service providers (LSPs) who want to make use of Machine Translation (MT) have to fight on several fronts. The skepticism within the language industry is still very high. End-customers worry about paying too much for translations that no human has interfered with. Translators refuse to get involved in post-editing activities because they fear that MT will take away their actual work, rendering themselves useless on the long run. And competitors try to outperform each other by either spoiling the market with low-quality MT offered on dumping rates or declining MT altogether for being inappropriate for commercial usage. This paper seeks to show that well-defined quality metrics can help all stakeholders of the translation market to specify adequate benchmarks for the desired translation quality, to use an agreed-upon consistent mark-up and to evaluate translation quality – MT and human translation output alike – accordingly. As a by-product, this professionally error-annotated MT output will help researchers to further improve MT quality, which in turn will help to make this technology more popular in the industry.

Keywords: Machine Translation, Translation Quality, Quality Evaluation, Benchmarking, Tool Integration

1. Introduction

The discussion on the usage of MT in daily life-scenarios of the translation business is still highly controversial. Whereas some sectors (e.g. the software sector) apply MT already for quite a while in their standard localization workflows, others tend to think that MT is not fit to end-clients’ requirements.

Among the group of the second important stakeholder in the translation market – the human translators – the fear of abolishing their own jobs by optimizing MT output through offering post-editing services is prevalent. The pool of adequately trained post-editors is very limited and the actual qualification profile for post-editors remains blurred.

Language Service Providers (LSPs) themselves, have to cope with tough and narrow margins. Some in the business help themselves by offering low-quality MT at rock-bottom prices without bothering about systematic post-editing. This again lowers the esteem of MT in general in the sector as a whole as well as among end-customers.

But what do we mean when talking about quality at all? How can quality be measured? And how can the parties involved contribute to reach the goal of higher MT quality? This paper seeks to provide answers to these questions.

2. Obstacles LSPs are Facing when Using MT

Although significant progress in field of MT within the language service industry is to be observed, many LSPs still reside from the broad usage of MT in their daily routines. This is for various reasons.

2.1 Lack of Understanding

Since only larger LSPs can afford to run professional MT divisions in their companies, most LSPs have to rely on ready-made MT solutions that are on the market. Often enough this means using untrained translation services that have no relevance to the domain and the end-users’ area of application. To many translation providers, these solutions equal black-boxes that deliver output they can neither affect nor properly evaluate. Not having universal evaluation criteria in place, LSPs are often forced to have a human editor proof-read the complete material without really knowing what kind of issues to focus on. This makes an objective quality assessment difficult, if not impossible. What is worse, the steps are performed in different tools, breaking the commonly applied translation workflows and causing additional manual pre- and post-production steps.

Taking the above said into account, for many smaller LSPs the use of MT seems inefficient and costly, instead of saving them time and money.

2.2 Common criteria

What would help to raise acceptance for MT in the translation business, therefore, is a better understanding of the processes and of the anticipated output. If the providers and the requesters of MT were on common grounds concerning evaluation criteria both parties would benefit – the requesters would know what to expect and the providers where there is room for optimization in their MT results.

The goal should be that both sides worked together more seamlessly, using the same vocabulary for common quality issue types in order to optimize MT engines accordingly, and thus resulting into improvements from translation to translation.

3. Quality Metrics: Why and How

The problem of how to evaluate the quality of translations is not new to the language service industry. Also for human translation, the question as to whether a translation is good or bad and by what this finding can be measured has been a matter of dispute as long as the professional translation sector exists. Despite many and improved ways of computer-aided checking methods the so-called “Four-eyes
principle” is still the method of choice. Even the latest version of the industry standard – the ISO 17100 – does not accept any other means of quality revision.

3.1 Learning from evaluation of HT
Evaluation performed by humans always harbors the risks of subjectivity and inconsistency. We cannot abandon these risks completely. By defining clear principles for error classes and by categorizing errors accordingly, these risks can be minimized significantly, though. The same goes for the evaluation of MT output. Therefore, a dedicated metrics system is key to a controlled quality assessment for both MT and HT.

3.2 Relevant metrics
Using some kind of metrics for the quality estimation is not new to the industry, either. There have been several approaches to the compilation of error scorecards to support objective human quality evaluation models (LISA QA, SA J2450). The problem with these approaches was, though, that they were either restricted to one domain (SAE J2450), or that they followed a “one-size-fits-all” approach (LISA QA and its predecessors).

What was missing for a very long time was an approach that allowed us to compile domain- or even end-customer specific error profiles and – based on those profiles – standards that have to be reached in order to rate a given translation as acceptable.

What acceptable quality means for different environments must be defined by the industries and businesses themselves. That means that the industries or even companies must specify for their textual domains which error classes and categories are relevant in their respective use cases. Only upon these specifications error categorization and annotation can be performed. This is a distinction that automatic evaluation scores obviously cannot deliver.

3.3 Industry’s requirements towards MT
For human translation, a translation job that is rated as unacceptable will be returned to its producer in order to have it fixed. Post-editing will (under usual circumstances) not be performed on translated material that is rendered deficient in many ways. The same principle goes for MT output: If the MT output is too far away from what is needed for a given translation scenario a human translation from scratch will be performed faster than the post-editing of a machine-translated text. In other words: In such a scenario, the usage of MT for a translation company is economically nonsense. An LSP will not incorporate such a workflow on the long run. The judgement as to whether a given translation serves its actual market purposes cannot be performed by the means of automatic assessment scores but only by human specialists who have linguistic and domain-specific knowledge.

Another argument that is applied in human translation revision scenarios must be taken into consideration in the

MT context: After reviewing and – if necessary – reworking a translator’s work it is common practice to provide them with feedback on what they delivered. For the next assignment, the LSP will expect not to find the same kinds of errors again in the translator’s work. If they do – maybe repeatedly – it is very likely that the LSP will terminate the cooperation with this translator in the near future for obvious reasons: The translator seems unwilling or incapable to learn.

The same requirement is valid for MT output. If an LSP has to correct the same error types again and again in every MT workflow it is probably not worthwhile using it. Just as with the human translator, the LSP would expect the machine to learn from its previous mistakes i.e. have the MT engineers fixed what went wrong during the last translation circle.

4. MQM: A Recap
Before specifying categories for a quality metrics system we must define what we mean by “quality”.

4.1 The Idea of Quality
The underlying quality definition stated by the originators of Multidimensional Quality Metrics (MQM) assumes that a quality translation “demonstrates required accuracy and fluency for the audience and purpose and complies with all other negotiated specifications, taking into account end-user needs” (Koby and Melby 2013). What is important about this definition (and what sets it apart from other translation quality assessment theories) is that the end-users, applying the metrics determine the relevance of a given category, rather than the metrics itself.

4.2 The MQM hierarchy
The complete MQM master lists all issue types that different existing metrics models contain and results in a comprehensive but rather confusing hierarchy:

---

1 See https://www.iso.org/iso/catalogue_detail.htm?csnumber=59149
2 See https://www.sae.org/standardsdev/j2450p1.htm
3 See http://qt21.eu/mqm-definition
4 See http://qt21.eu/mqm-definition
For actual post-editing purposes this hierarchy would be impossible to handle. What differentiates the MQM model from other approaches, though, is the fact that only those categories will be considered that are needed in the translation scenario in place. Applying the concept of the end users’ requirements, requesters and providers can agree upon their relevant set of categories beforehand, making sure that only issue types are taken into account that matter to a given translation context. The MQM core (see figure 2) consists of 21 more commonly addressed issue types:

* = potentially automatable

![Figure 2: MQM Categories (core)](image)

4.3 Learnings from MQM annotation
The MQM system applied to error annotation of MT output serves the evaluation on several levels. For LSPs it makes error types countable and allows to classify them. It not merely counts errors, though, but it enables industries to choose only those error categories that are relevant to them. These can obviously vary from domain to domain and from text type to text type. By that, certain error profiles for various use cases and sectors can be compiled. Moreover, errors can be weighted according to their severity in a given context. An error that does not affect the overall quality of a text in one domain can be a show-stopper in the other. For example, the usage of the curly quotation marks (“ ”) instead of straight ones (" ") do not affect the text quality of a technical documentation whereas in a software localization project where straight quotation marks function as a marker for UI options curly quotation marks can break the software strings and ruin the whole translation.

Based on the relevant categories industries can define benchmarks that function as a delimiter for different quality levels. If an annotated test sample of a given output falls below the defined threshold then the translation is not suitable for an MT + post-editing workflow without making improvements on the engines that produced the output.

For MT engineers MQM helps to understand where their engine does well and where it fails. Whereas that may be true for automatic evaluation methods, this information alone leaves the improvement of engines to a large extent to the field of trial and error.

For an interpretation on why the engine fails in certain contexts and which patterns these errors show, a more detailed analysis will be necessary. This analysis (see figure 3) can be performed only by a trained linguist who has a deep understanding of both source and target language. The reliability and thoroughness of human annotation compensates for the higher effort compared to an automatic evaluation method.
5. Easing the Collaboration

During the last years, not only the lack of common metrics and standards impeded a collaborative approach between language service business and language research. In order to make a cooperation between both parties work they also need work environments and tools that integrate well into the processes in place.

5.1 Tools and Applications

Since the times of LISA QA much has happened. Whereas in those days QA professionals had to fill in Excel spreadsheets with exact reproductions of found errors or to shoot screenshots to prove inconsistencies in translations, nowadays technological support by adequate applications is available. But, although there are many standalone tools in the market – open source and commercial – that offer useful functionalities that LSPs need for reasonable revision stages, one huge problem remains: Most of them break the industry’s common workflows for the handling of translation projects, HT and MT alike. That means existing translations have to be exported from the translation environment in use to be imported into the revision tool. After revision is done, the reworked material as well as the error descriptions have to be returned back from revision to translation tool. And metrics and evaluation results are most likely to be managed in yet another system like a translation management application or a translation resources database.

Every working step aside from the dedicated workflow path, though, costs the LSPs time and money and cause a vast management overhead. This renders the application of the given method user-unfriendly and uneconomic.

For MT researchers and engineers on the other hand it is important that the results from annotations and error markups can be fed back easily into the MT engines in order to optimize output during the next optimization and translation round.

Future advances in the field, therefore, must not only focus on assessment methods but also on the development of suitable tools where methods interlock with translation and revision workflows. Only if functionality and accessibility is well integrated into the translation and post-editing environment and if engines can “learn” easily from post-editors’ feedbacks added value for all parties will be generated.

5.2 Development and Progress

Fortunately, advances for a more feasible MT usage is underway, and huge progress has been made, recently. The TAUS Dynamic Quality Framework has developed a range of tools that support evaluation and benchmarking within the industry. The underlying DQF Error Typology that has been harmonized with the MQM model in 2015 and that represents a subset of the MQM specifications provides a means to quantify translation errors. It can be integrated via customized plugins into many commercially available translation tools or will be in the very near future according to a TAUS press release as of March 2016.

Old and new application vendors like SDL, Memsource or MateCat have brought up new functionalities and CAT tools that combine MT and translation memories for traditional computer-aided translation into well-integrated workflows. Many of them function in the cloud and offer real-time processing and interactive post-editing of suggested MT segments. The result is a “self-learning environment” that not only measures editing distance and errors but also incorporates required changes for future similar occurrences.

Although not all open questions are answered yet, these forward-looking developments in the field of language technology are encouraging and propose a real change to come in the translation market.

6. Conclusion

The usage of MT in many professional translation contexts bears many chances and future prospects for the translation industry. MT researchers, on the other hand, need large amounts of domain-specific data to train engines and qualified expert feedback that serves as a basis for further optimization.

If both parties bundle their knowledge and leverage it for the sake of high-quality MT not only the language service

---

5 See https://evaluate.taus.net/evaluate/dqf-tools


Proceedings of the LREC 2016 Workshop “Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem”, Georg Rehm, Aljoscha Burchardt et al. (eds.)
sector but also the field of MT research will have their merits. The analyses from real-life scenarios offer valuable insights into common error patterns and necessary approaches for the improvement of MT engines. By using common standards, consistent benchmarks and integrated tools all players will benefit from each other’s work in order to reach better results.

7. Acknowledgements

This work has received support from the EC’s Horizon 2020 research and innovation programme under grant agreement No. 645452 (QT21).

8. Bibliographical References


9. Websites


LISA QA Model 3.1: https://www.w3.org/International/O-LISA.html (consulted 31.03.2016)


TAUS: https://www.taus.net/ (consulted 31.03.2016)