QT21: A New Era for Translators and the Computer

Alan K. Melby
LTAC Global
akm@ltacglobal.org

Abstract

QT21 (see http://www.qt21.eu/) is an EU-funded project with several goals related to machine translation. This paper relates to the QT21 goal of "improved evaluation … informed by human translators", using a framework that harmonizes MQM (Multidimensional Quality Metrics) and DQF (Dynamic Quality Framework). The purpose of the paper, which expresses my personal views, is to obtain feedback on three claims I am making about translation quality evaluation of both human and machine translation: (1) Both automatic, holistic reference-based metrics (such as BLEU) and analytic manual metrics of translation quality are needed; (2) one metric is not sufficient for all translation specifications; and (3) widespread use of specifications and the harmonized MQM/DQF framework for developing metrics will have a positive impact beyond the QT21 project. If these three claims turn out to be true, we will see a new era in the relationship between translators and computers.

1 Introduction

One goal of the QT21 project (http://www.qt21.eu/) is to work toward "improved evaluation and continuous learning from mistakes, guided by a systematic analysis of quality barriers, informed by human translators". This effort will involve including professional translators, language service companies, and other stakeholders directly in the process of evaluating the quality of raw machine-translation (MT) output, using an analytic approach to complement the current *automatic*, holistic, reference-based approach. An *analytic* approach provides detailed information about errors as far down as the word level and does not require a reference translation, but it is manual; that is, it must be performed by a skilled human, rather than being automatic. Both approaches, *analytic* and *automatic* for short, will be used in QT21.

Over the past decade, research on statistical MT has, for various reasons, progressed somewhat independently from the practice of individual professional translators. However, the QT21 project goals indicate a belief that this needs to change. Human translations are used as reference documents in the automatic approach, but the translator who produced a reference translation will usually never see the output of a machine translation system. Instead, in the analytic approach, professional translators directly evaluate the raw output of machine-translation systems, using tools that allow specific errors to be identified and annotated by human evaluators. The results of this human evaluation can then hopefully be used by developers to determine what went wrong and how to improve the system.

Lest translators worry that they will be working themselves out of a job by helping researchers improve machine translation, I point out that for the foreseeable future, raw machine translation will be used "as is" in only very limited situations. See Figure 1 for various use cases along a spectrum of interaction between human and machine translation.

In the 1950s, some in the MT research community expressed optimism about the potential for rule-based MT to replace professional translators. Then, the first decade of the current century, some suggested that data-driven machine translation systems would eventually

Proceedings of the 37th Conference Translating and the Computer, pages 1–11,
London, UK, November 26-27, 2015.

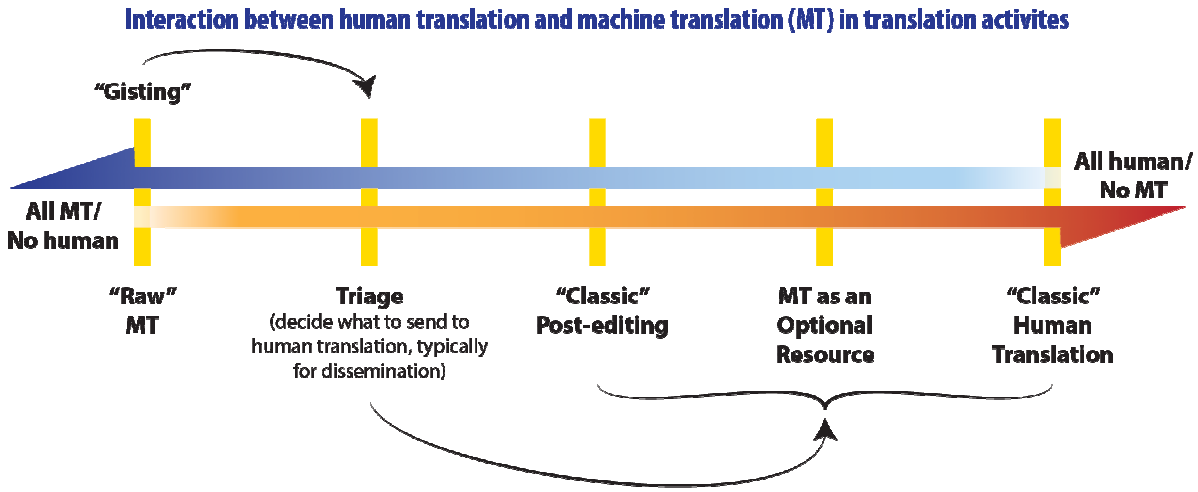

Figure 1. Use cases along the spectrum of interaction between human translators and machine translators

produce output as good as or better than human translation (see http://www.ttt.org/amta/), but the QT21 project does not take this position.

MT is often used for tasks where professional human translation is impractical for one reason or another (e.g., instant, on-demand translation of low-value content, or translation where access to human translators is not feasible, or user-generated content where time-frames do not permit professional translation). When it is clear to all parties where MT is useful and where it is not, the immense value that professional translators provide can be better seen. Research and development in MT will hopefully enable professional translators to concentrate even more on the most challenging and rewarding types of translation.

In Figure 1, human translation plays some role in all use cases, and MT is involved in all but the *"Classic" Human Translation* use case. In the *MT as an Optional Resource* use case, translators use technology, but remain in complete control of which resources—such as a mix of terminology lookup, translation memory, and MT—are used in translating each particular segment of text. This point on the spectrum includes recent renewed interest in interactive MT (Green 2015). It is clear that translators will increasingly find themselves working in environments where MT is available to them on at least some segments. Hopefully, various interactions between MT and human translators will increase productivity, as has translation memory.

Varying types of professionals are involved with each of the five categories listed above:

- In statistical MT development, most of the work is done by software engineers, mathematicians, and computational linguists who use corpora of human translations as training data for their systems (therefore involving human translation as the basis for *raw MT*);
- For *triage*, the evaluation of MT output is typically done by monolingual subject-matter experts who decide which documents to send to human translators;
- *Classic post-editing* (where errors in raw machine translation are corrected from beginning to end) may be done by professional translators, but is often done by others, depending on the requirements (e.g., in some post-editing scenarios, minimal corrections are made by individuals trained specifically in post-editing, but who do not otherwise provide translation services); and finally,
- For the two rightmost use cases, *MT as an optional resource* and *"classic" human translation* (where MT is not involved), services are provided by professional (or paraprofessional) translators.

With increased interaction between human translation and machine translation, comes the need for methods of translation quality evaluation that apply to both. To complement existing automatic approaches, which are used only to evaluate machine translation, QT21 provides a framework (called MQM/DQF) within which metrics can be defined that can be used for analytic evaluation of either human or machine translation.

As used in this paper, a metric is a quantifiable measure. If what is being measured is changed, even slightly, a different metric is being used. Not all aspects of translation quality can be quantified, so metrics deal with those aspects that can be quantified.

I strongly believe that professional translators will benefit from QT21 because they will become better equipped to examine translation requirements, develop translation specifications, and provide a verifiable evaluation of when and how machine translation should be involved in a project, along the spectrum in Figure 1. This could help usher in a new era of collaboration rather than competition between professional translators and machine translation. There will be plenty of work for professional human translators.

One purpose of this paper is to obtain feedback from stakeholders in the language industry on the following claims I am making, regarding the implications of the QT21 goal of achieving improved evaluation of translation quality informed by human translators:

(1) Both automatic, holistic reference-based metrics (such as BLEU) and analytic manual metrics of translation quality are needed;
(2) One metric is not sufficient for all translation specifications (e.g., full vs. summary translation, overt vs. covert translation, and differing requirements for style and speed[1]); and
(3) Widespread use of specifications and the harmonized MQM/DQF framework for developing metrics (see http://www.qt21.eu/quality-metrics/) will have a positive impact beyond the QT21 project.

The rest of this paper expands on various points in this introduction.

2 Overall Focus of the QT21 Project and this Paper

A glance at the QT21 website (http://www.qt21.eu/) shows that the overall focus of the project is to develop machine-translation systems for "morphologically complex languages" with "free and diverse word order". As can be seen from the Introduction, this paper is not about techniques being used within QT21 to develop MT systems for these types of languages. There will be many papers published on this topic over the next several years. Instead, this paper is about the stated QT21 goal of "improved evaluation and continuous learning from mistakes, guided by a systematic analysis of quality barriers, and informed by human translators". There are other approaches to evaluation, such as task-based evaluation, that are beyond the scope of this paper.

3 Why Isn't There More Interaction between MT Developers and Professional Translators?

Twenty years ago, both statistical and rule-based approaches to MT were under consideration. As always in translation, both human and machine, there was discussion of how to evaluate

[1] These types are sometimes addressed under the rubric of "content correspondence". For example, is the target intended to be a *full translation* or a *summary translation*? Should it be an *overt translation* (i.e., it does not conceal that it is a translation) or a *covert translation* (i.e., it appears as though it were written in the target language with no obvious traces of the source that reveal it to be a translation) or an *adaptation* (a text that moves beyond "pure" translation to include substantial adaptations for the target audience)? Since translators generally assume covert, full translation, it is critical that other types be explicitly noted.

translation quality. In White et al. (1994), we see an early explanation of the terms "adequacy" and "fluency", which are sometimes respectively equated with "accuracy" and "readability". However, accuracy involves a direct comparison of the source text and target text, to see whether they correspond; adequacy, on the other hand, is an indirect measure of accuracy, based whether information in a reference translation is found in the raw machine translation by a monolingual evaluator.

Here is how White et al. describe these key terms:

> In an adequacy evaluation, literate, monolingual English speakers make judgments determining the degree to which the information in a professional translation can be found in an MT (or control) output of the same text. The information units are "fragments", usually less than a sentence in length, delimited by syntactic constituent[s] and containing sufficient information to permit the location of the same information in the MT output. These fragmentations are intended to avoid biasing results in favour of linguistic compositional approaches (which may do relatively better on longer, clause level strings) or statistical approaches (which may do better on shorter strings not associated with syntactic constituency).
>
> In a fluency measure, the same evaluators are asked to determine, on a sentence-by-sentence basis, whether the translation reads like good English (without reference to the "correct" translation, and thus without knowing the accuracy of the content). Their task is to determine whether each sentence is well-formed and fluent in context.

This approach was adopted, in part, because it allowed researchers to use readily available human resources for a task that was seen as not necessarily requiring the expertise of professional translators. About ten years later, automatic techniques for comparing reference translations and raw MT output, such as BLEU, began to appear (Papineni et al., 2002), which offered many apparent advantages over manual approaches.

During the past decade, *reference-based metrics* such as BLEU have been at the centre of evaluating the quality of MT output. In these approaches, one or more (seldom more than two or three) human translations of a source text are obtained. The raw output of the MT system is automatically compared with these reference translation(s), and a score is obtained, typically between 0.0 and 1.0 (or 0 and 100), where close to zero would indicate no overlap whatsoever between the MT output and the reference translation(s), and a score close to one (or 100) would indicate a nearly perfect match. The score is holistic in that it describes a property of the output text as a whole.

Human evaluation has also been used throughout the past decade to complement automatic evaluation, but it has been primarily holistic, for example using ranking (which segment or text is better?) rather than analytic error analysis. My first claim is that QT21 is correct to expand human evaluation to include an analytic approach using MQM/DQF.

The human translators who produce the reference translations typically do not see the raw machine-translation output, and the machine-translation developer who obtains the BLEU score may not speak either the source or the target language of the system being evaluated. The evaluation is purely mechanical. Furthermore, the BLEU score, being just one number, does not tell the developer what to do to improve the system. Often, developers tinker with the system, run it again on the same source text, and obtain a new BLEU score without looking carefully at the output. If the score goes up, it is assumed that the change to the system was a good one.

The MT development community widely acknowledges the limitations of BLEU and similar approaches; yet the field continues to use them because no cost-effective alternatives have yet appeared for scenarios where developers modify systems and need to see how their modifications affect the output. It would be impractical to run a change and then need to wait for days or weeks for evaluation of the changes. In particular, as Callison-Burch et al. (2006) document, one of the promises was that BLEU would correspond to human judgment (and thus serve as a useful proxy for more labor-intensive evaluations); yet the degree of correlation has proved to be less robust than had been hoped, with cases in which human judgment and BLEU contradict each other.

A perusal of papers presented at recent instances of the Workshop for Machine Translation (WMT) shows that BLEU is widely used as a proxy for "quality", along with human ranking of segments. However, additional methods of evaluation, besides automatic comparison with reference translations and human ranking of output, are starting to gain traction. At LREC 2014 in Reykjavik, a workshop was held that explored alternative methods of assessing translation quality; it included hands-on experimentation with analytic error-annotation methods (Miller et al., 2014). In both 2014 and 2015, WMT hosted a shared task on quality evaluation that used data annotated for errors using the MQM framework (discussed in Section 6) as references for training systems to predict specific error types.[2] Although the results of these shared tasks were not conclusive, considerable work is being carried out in this area.

It must be pointed out that the automatic approach has the distinct advantage of being practically instant and completely reliable. If a BLEU metric is re-applied, it produces exactly the same result. However, manual analytic evaluation, because it involves humans making judgments, is not perfectly reliable. Different human judges may come up with different results applying the same metric. This problem is encountered in quality management across all industries but it can be addressed. Achieving an acceptable level of reliability in the analytic approach involves fine-tuning of the training materials and testing the evaluators.

An interesting question for further study is what specifications have been given to the human translators who produce reference translations.

In last year's ASLING keynote address (Prószéky, 2014), it was noted that neither the purely statistical approach of recent systems nor the hybrid approaches currently being tried have produced raw-machine translation at hoped-for levels of quality. So what comes next? I suggest that one thing that comes next is work on the QT21 goal of "improved evaluation … informed by human translators", despite the difficulties of achieving high levels of reliability in manual analytic evaluation, and further emphasis on translation specifications.

4 Large-Scale Involvement of Human Translators in Analytic Quality Evaluation

Previous MT research efforts have involved translators, often productively, but on a relatively small scale. The QT21 project appears to be increasing the scale and nature of this involvement. In the QT21 proposal submitted to the EU, we find the following observations:

> [M]ainstream MT quality assessment methods based on automatic metrics are incompatible with the methods used for professional human translation, and typically do not reflect the needs of actual users of translation.

[2] See http://www.statmt.org/wmt14/quality-estimation-task.html and http://www.statmt.org/wmt15/quality-estimation-task.html .

> [In addition to] its utility for diagnostic purposes, putting humans in the loop also marks a significant change in the current MT development/maintenance paradigm.
>
> [E]xplicit error annotations could be used to pinpoint specific issues that happen systematically. Such information, disregarded by pure data-driven methods, would help to develop advanced diagnostic tools, as well as to trigger and drive focused (error-specific) improvement techniques on different aspects of the MT process.

In evaluating professional translation, except in an educational or testing environment, a reference translation is not available. Instead, translation is evaluated in various ways, most frequently by the identification of "errors". By including *analytic* evaluation techniques that involve manual identification of specific issues in a translation (rather than *holistic* approaches, either automatic or manual, that evaluate a translation as a whole), often analyzing right down to words or phrases, without a reference translation (rather than automatic estimation), the same techniques can be applied to both human and machine translation. This analytic approach has generally not been undertaken in the past because of concerns about cost and time, but work in the QTLaunchPad project (http://www.qt21.eu/launchpad/) showed that manual analytic analysis, when properly focused, is sufficiently promising to merit further exploration in the QT21 project.

However, work on analytic evaluation raises the question of which error typology to use. While various proposals have been made for error typologies (e.g., Flanagan, 1994) and even tools developed to assist with error annotation (e.g., Nießen, 2000), none of these has gained traction or widespread adoption. As a result, most error-annotation efforts to date have used ad hoc typologies that prevent the direct comparison of results and have remained largely isolated efforts. The use of post-editing analysis (e.g., in Hjerson, a system for automatic classification of MT errors based on reference translations (Popović, 2011)) is beyond the scope of this paper.

QT21 includes a plan to extend analytic error annotation to thousands of segments in many languages, and to correlate the results with other quality-evaluation methods. Exactly how the results of analytic evaluation will be used to improve a particular MT system is beyond the scope of this paper.

5 Why One Translation-Quality Metric is Not Sufficient

Assuming that a given translation quality metric can be applied to both human and machine translation, there is still the question of whether metrics vary according to the type of translation that is required. Initially, it might be tempting to look for one translation-quality metric that can be applied to all translation projects. At a very general level, there is one metric: a translation should be accurate and fluent. That is, it should correspond to the source text, according to the type of translation requested, and it should read well in the target language, independent of whether it is a translation or an original composition. However, simply expecting "accuracy and fluency" is not a sufficient guideline to evaluate all translations in a useful manner, irrespective of the purpose and the intended audience of the translation.

One thing that nearly everyone in the translation industry agrees on is the importance of translation *project specifications* (sometimes called a *project brief*) that include full details about expectations, including audience and purpose, target language, expectations for terminology, and many other aspects. Suppose the specifications call for only a short

summary translation of less than three hundred words, but the translator produces a beautiful full translation three thousand words long (about the same length as the source text). That translation will receive a negative evaluation. Perhaps the most obvious specification is the target language. If someone requests a translation into "SL" (Slovenian), but it is delivered in Slovakian ("SK") because a project manager misinterpreted the language codes, it will be rejected. Likewise, a highly accurate and fluent translation of a technical-support item that is delivered a week after it is needed to solve a problem will not be given better ratings than a less fluent, but useable, translation that is delivered in time to be useful in solving a time-critical problem. Not meeting the agreed-on specifications is problematical. Thus, it is also important to define the specifications carefully. A metric tied to inappropriate specifications is useless.

A translation-quality metric must be linked to a set of appropriate translation specifications to be valid. Since there are many widely differing sets of translation specifications, there must also be many translation-quality metrics. Metrics differ in many ways:

- *Which error-category hierarchy* they draw on;
- *What is checked* (e.g., a piece of external marketing material might be checked carefully for style, which an internal service manual would generally not be);
- *How errors are weighted* (the relative importance given to kinds of errors);
- *How granular* (detailed) *the categorization* and annotation of issues are; and, very importantly,
- *What is considered to be an error* (e.g., a deviation from the source text might be considered an error in an overt translation, but an appropriate adjustment to the target culture in a covert translation).

Thus, metrics must be applied according to the specifications they are based on. For example, a quick "acceptance test" metric of a progress report might ask evaluators to provide a simple rating for accuracy, fluency, and style for the entire text, while a final-review metric of the translation of a legal document might require detailed annotation of every single error.

6 Specifications and Metrics in QT21

Rather than developing an ad hoc system for developing translation specifications, the Multidimensional Quality Metrics (MQM) format for quality metrics developed in the QT LaunchPad project uses an existing international standard, ASTM F2575. Section 8 of F2575 (2014) explains how to develop structured translation specifications using a standard set of 21 translation parameters, which include the obvious parameters of target language, delivery deadline, and *content correspondence* [3], but also many other parameters established empirically through collaborative standards development involving many stakeholders. The QT LaunchPad project had some influence on the 2014 version of F2575.

Once a set of structured translation specifications is established, a comprehensive hierarchy of error categories is needed. Over the past several years, two such hierarchies have evolved in parallel: the Dynamic Quality Framework (DQF) from TAUS (www.taus.net), and the MQM framework. (See Lommel et al., 2014 and Lommel et al., 2015). As part of the QT21 project, these two hierarchies have recently been harmonized, with DQF as a fully compliant MQM subset that is recommended for use in machine translation, general professional translation, and localization scenarios. Already, various tools are emerging that are based on the harmonized MQM-DQF hierarchy of error categories, some free and open-source, some fee-based.

[3] Content correspondence (full/summary, overt/covert, etc.) addresses the relationship of the source and target texts.

One metric is insufficient for all specifications, but all metrics can now use the same error categories, with standard names and definitions. (As noted above, however, the application of

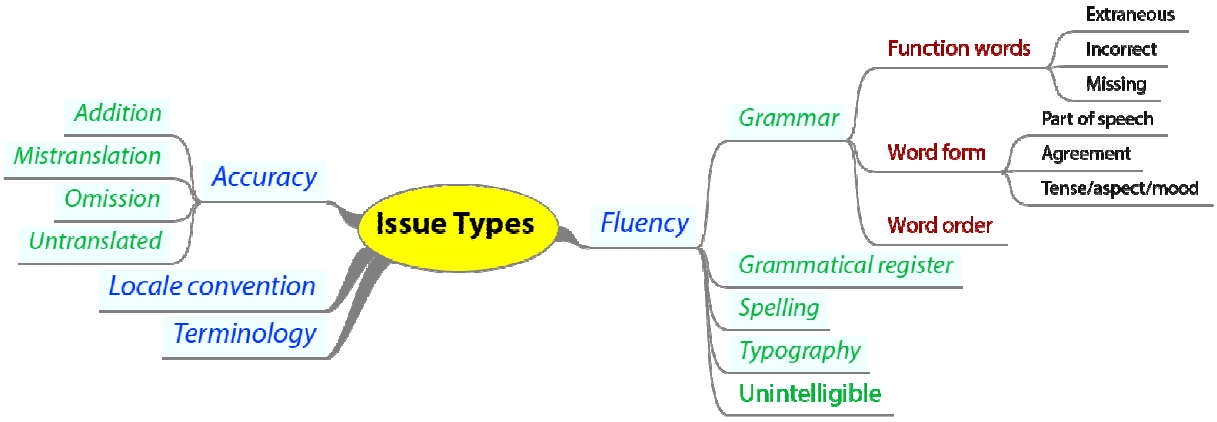

Figure 2. Graphical representation of the MQM metric used in QT21 for evaluation of MT output

an error category is relative to the specifications, in particular with respect to "content correspondence".)

Figure 2 shows a graphical representation of the issue types in one such MQM metric, adapted specifically for working with MT output.

Figure 3 shows an implementation of this particular metric in a "scorecard" tool, developed in the QTLaunchPad project, which allows for tagging issues at the segment level.

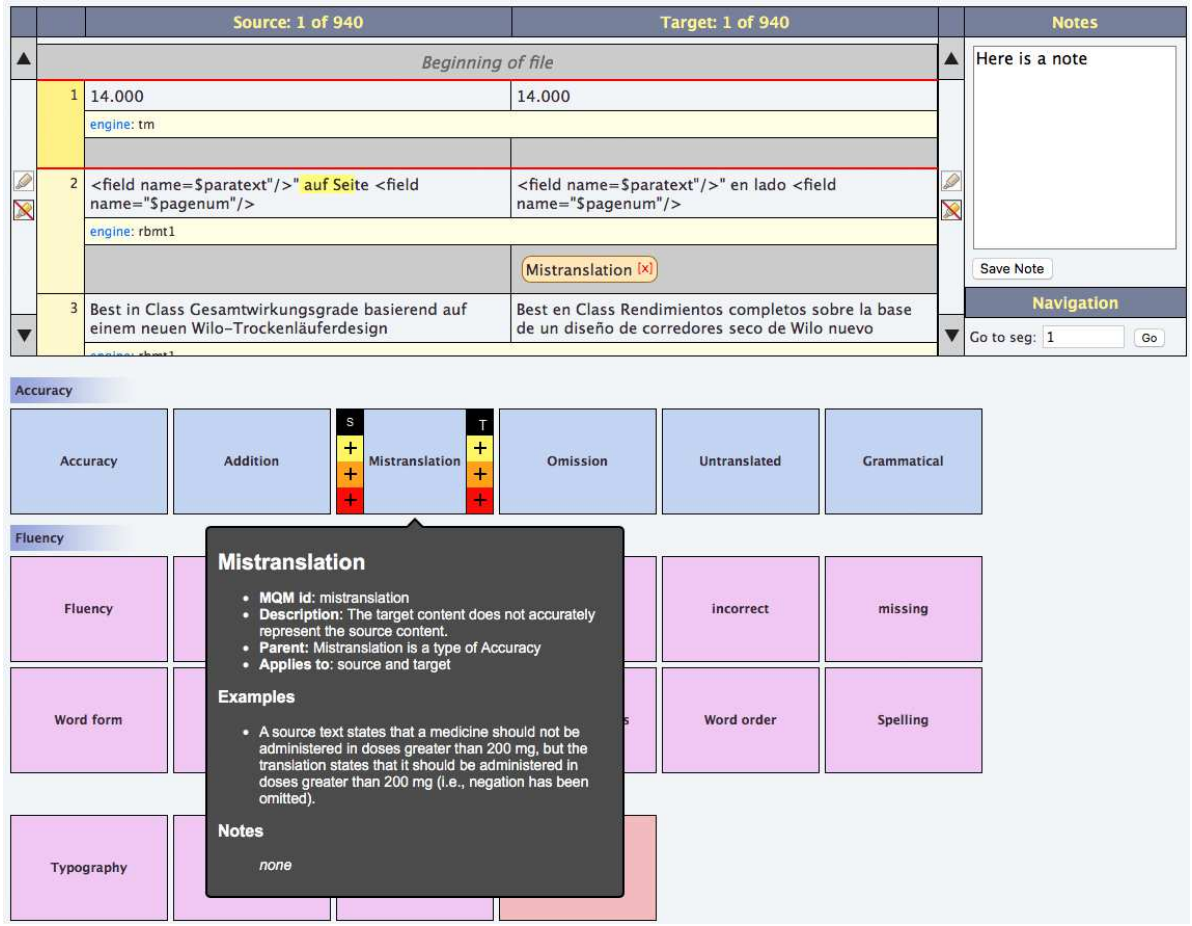

Figure 3. Implementation of a metric in a free and open source "scorecard"

Figure 4 shows a much simpler selection of issues compatible with the DQF subset. This is thus a different metric from that in Figure 2. This selection of issue types might be suitable for the evaluation of Word documents that have been processed using translation memory (it allows "improper exact matches" from TM to be flagged), and addresses *Design* (formatting) at a broad level, with special attention to cases where text is truncated due to text expansion.

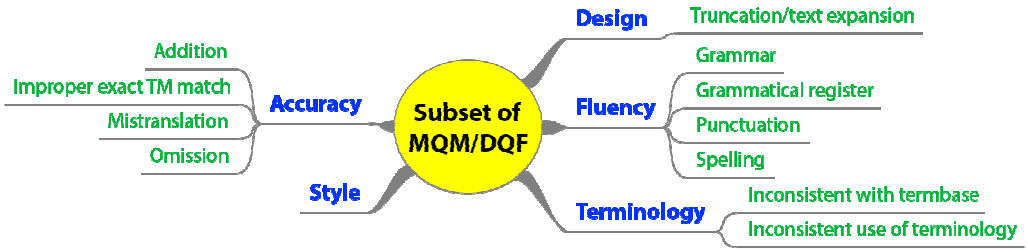

Figure 4. Possible metric from a subset of MQM/DQF (for evaluating human translation)

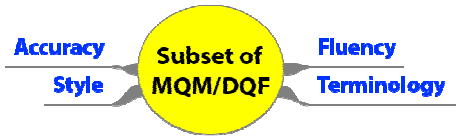

Figure 5. Selection of issues for a very simple metric

Figure 5 shows a very simple metric (also compatible with DQF) that might be sufficient for a "quick and dirty" assessment of human translation where only general types of errors are needed (i.e., if the source and target convey different meanings, then *Accuracy* is used; if the text is linguistically malformed, then *Fluency* is used; *Terminology* is used to mark incorrect terms; and *Style* is used to mark violations of the style guide.).

As can be seen from these examples, the approaches taken in quality evaluation in MQM are flexible for specific needs, but they are consistent in treating human translation and MT using the same methods.

7 Why it is Beneficial for Professional Translators to do Analytic Evaluation

Section 5 indicates why the QT21 project claims that the machine-translation community needs the involvement of professional translators; namely, to provide actionable diagnostics regarding specific problems in raw machine translation, rather than to depend on only a single "quality" number from an automatic metric such as BLEU, or even a manual holistic evaluation.

The Introduction also touched on why this is beneficial to all parties involved in the language industry: they will be able to provide verifiable evaluation of when to use raw machine translation, when to use classic human translation, and when to use some mix of the two. Professional translators should now therefore embrace MT. It will not replace them, but it can provide high-level consulting work to translators.

I believe that, so far, MT has tended to increase the amount of interesting work available to professional translators and other language professionals. I cannot prove it, but it would be interesting to launch a study on this question. I suggest that professional translators seek to better understand MT (including its strengths and limitations, and how to evaluate it relative to requirements) in order to profit from it. They also need to be able to counter scenarios in which upper management might suggest to translation department managers that they could reduce costs by simply replacing human translators with raw machine translation.

If faced with the question of whether a particular text should be translated by professional translators, MT, or some combination of the two (as in Figure 1), the real question is, what are the specifications for the translation? [4] Does an appropriate MT engine already exist? Does the engine deliver translations that meet the specifications? If not, can an appropriate engine be created, within time and budget constraints, that meets the expectations? How can the raw output of the MT system be used on the spectrum in Figure 1? These and other similar questions are the beginning of those that need to be asked to determine what role, if any, MT will play in specific scenarios. Only when professional translators can discuss specifications and actual results with respect to specifications can they make a convincing case for their work.

Machine translation and professional translation are not interchangeable. Professional translators should never be expected to produce less than their best effort. Machine translation should not be expected to produce professional levels of accuracy and fluency.

Instead of telling buyers of translation services that they need professional human translation because it is simply "better", translators and organizations that provide translation services should engage in a process of identifying requirements, developing specifications based on those requirements, selecting an appropriate translation environment and method, and then evaluating whether a translation meets the requirements or not, based on a suitable metric (presumably using the MQM/DQF framework) and trained evaluators who can apply the metric reliably.

8 Conclusion

I have endeavored to support the QT21 plan to add manual, analytic metrics to current evaluation methods. It is not yet clear how the QT21 goal of using improved evaluation to guide the improvement of MT output will evolve. However, it is clear that there is an urgent need for professional translators on the one hand, and translation buyers on the other, to enter into dialogue and cooperation regarding MT, rather than ignoring it or, worse, taking an antagonistic attitude towards it. Antagonism can unintentionally encourage the confusion and damage caused by buyers who sometimes purchase "bad translations".[5]

I believe there will be a very positive consequence of QT21, as indicated in the third claim. What is the positive impact of QT21 of this claim from the Introduction? I believe that a key to constructive dialogue is the use of translation specifications based on ASTM F2575-14, as discussed throughout this paper, especially in Section 6, in conjunction with the MQM/DQF framework for defining translation quality metrics. F2575-based specifications, paired with the MQM/DQF framework in QT21, will provide valuable tools to professional translators when they engage with translation buyers to decide, based on specifications, not emotion, what mix of human and machine translation is appropriate in a particular translation project (refer back to Figure 1). I boldly suggest that the specifications+metrics approach from QT21, regardless of how it impacts MT development, could usher in a new era for translators and computers.

[4] Defined per the 21 standard translation parameters in ASTM F2575-14 (see www.astm.org)

[5] Another important topic, outside the scope of this paper, is the downward price pressure felt by professional translators today. Bad translations (i.e., translations that do not meet specifications) might be cheaper, but this harms all stakeholders. I believe that this downward price pressure comes not from technology itself, but from translators who unwisely offer services at unsustainably low prices, from buyers who are unable to distinguish between translation that does and does not meet their requirements, and from unfair practices, such as those that assume that translation-memory matches require no human review, and/or expect humans to work without sufficient context (see http://www.ttt.org/context/).

I invite feedback on the various claims in this paper. I do not expect everyone to agree with everything I have written, but I do ask for civil debate.

Acknowledgments

I thank Arle Lommel, who is part of the QT21 project, for multiple discussions and many contributions to this paper. I also thank Kim Harris for explaining alternative perspectives that I will probably encounter. I am a Professor Emeritus at Brigham Young University (BYU), and a member of the Council of the International Federation of Translators (FIT), but the opinions expressed in this paper do not necessarily coincide with the official position of QT21, BYU, or FIT.

References

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. *EACL* 6: 249–56.

Mary Flanagan. 1994. Error classification for MT evaluation. In *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 65–72.

Arle Lommel, Aljoscha Burchardt, Alan K. Melby, Hans Uszkoreit, Attila Görög, Serge Gladkoff, and Leonid Glazychev. 2015. *Multidimensional Quality Metrics (MQM) Definition*. http://qt21.eu/mqm-definition/

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica* 12: 455–63.

Keith J. Miller, Lucia Specia, Kim Harris, and Stacey Bailey. 2014. *Automatic and Manual Metrics for Operational Translation Evaluation*. LREC. http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-MTE%20Proceedings.pdf

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. *LREC 2000*. http://hnk.ffzg.hr/bibl/lrec2000/pdf/278.pdf

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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*, 311–18.

Maja Popović. 2011. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 96:59–67.

Gábor Prószéky. 2014. Almost Fifty Years after the (First?) ALPAC Report. *Translating and the Computer*, 36: 24–36.

Spence Green. 2015. Natural Language Translation at the Intersection of AI and HCI. *Comm. of the ACM*, Vol. 58, Num. 9. Pp 48-53 (also available at: http://queue.acm.org/detail.cfm?id=2798086)

John S. White, Theresa O'Connell, and Francis O'Mara. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas*. 193–205.