revista.tradumàtica
tecnologies de la traducció

455-463

Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics

Arle Lommel
Hans Uszkoreit
Aljoscha Burchardt

German Research Center for Artificial Intelligence (DFKI)
Language Technology Lab

ABSTRACT

In recent years translation quality evaluation has emerged as a major, and at times contentious, topic. The industry view on quality is highly fragmented, in part because different kinds of translation projects require very different evaluation methods. In addition, human and machine translation (MT) quality evaluation methods have been fundamentally different in kind, preventing comparison of the two. The lack of clarity results in uncertainty about whether or not a translation meets requesters' or end users' needs and leaves providers unclear about what requesters need and want. In response the EU-funded QTLaunchPad project has developed the Multidimensional Quality Metrics (MQM) framework, an open an extensible system for declaring and describing translation quality metrics using a shared vocabulary of "issue types"

Keywords: Machine Translation, Translation quality, Metrics, Issues types.

RESUM *(MQM: Un marc per declarar i descriure mètriques de qualitat de la traducció)*

En els últims anys l'avaluació de la qualitat de la traducció s'ha convertit en un tema rellevant i a la vegada que, de vegades, polèmic. La perspectiva de la indústria sobre la qualitat està altament fragmentada, en part perquè diferents tipus de projectes de traducció requereixen mètodes molt diferents d'avaluació. A més, els mètodes d'avaluació de la qualitat de les traduccions humanes i de les traduccions elaborades amb traducció automàtica (TA) són d'índole diferent. La manca de claredat provoca incertesa sobre si una traducció compleix amb les necessitats del seu promotor o l'usuari final, i deixa als proveïdors amb dubtes sobre el que els clients volen i necessiten. Com a resposta a aquest fet, el projecte QTLaunchPad, finançat per la Unió Europea, ha desenvolupat el marc denominat Multidimensional Quality Metrics (MQM), un sistema obert i ampliable per declarar i descriure les mètriques sobre qualitat en traducció utilitzant un vocabulari compartit de "classes de problemes".

Paraules clau: traducció automàtica, qualitat de traducció, mètriques, classes de problemes.

RESUMEN *(MQM: Un marco para declarar y describir métricas de calidad de la traducción)*

En los últimos años la evaluación de la calidad de la traducción se ha convertido en un tema relevante a la par que, en ocasiones, polémico. La perspectiva de la industria sobre la calidad está altamente fragmentada, en parte porque diferentes tipos de proyectos de traducción requieren métodos muy diferentes de evaluación. Además, los métodos de evaluación de la calidad de las traducciones humanas y de las traducciones elaboradas con traducción automática (TA) son de índole diferente. La falta de claridad provoca incerteza sobre si una traducción cumple con las necesidades de su promotor o su usuario final, y deja a los proveedores con dudas sobre lo que los clientes quieren y necesitan. Como respuesta a este hecho, el proyecto QTLaunchPad, financiado por la Unión Europea, ha desarrollado el marco denominado Multidimensional Quality Metrics (MQM), un sistema abierto y ampliable para declarar y describir las métricas sobre calidad en traducción utilizando un vocabulario compartido de "clases de problemas".

Palabras clave: traducción automática, calidad de traducción, métricas, clases de problemes.

Introduction

The Multidimensional Quality Metrics (MQM) framework is a flexible system for declaring translation quality evaluation metrics. Although focused primarily on evaluating the quality of translated texts, much of MQM is also suitable for evaluating the quality of source texts, a feature that can be used to diagnose problems in source texts and their impact on the quality of translated texts. MQM is also designed to be applicable to any sort of translated text (human or machine translated) and to any type of text. It does not, however, create a one-size-fits-all model for evaluating translation quality.

This article provides a general overview of the structure and principles of MQM. For a more detailed technical examination, readers are encouraged to consult the current formal definition of MQM (http://qt21.eu/mqm-definition).

Translation quality evaluation

Accurate and objective evaluation of translation quality is a topic of considerable importance to both *requesters* and *providers* of translation. Requesters (sometimes known as *clients*) are generally interested in knowing if the translations they receive meet their quality expectations for a variety of reasons such as mitigation of legal liability due to factually incorrect translations, reduction in support costs, and preservation of brand reputation. Providers are interested in ensuring that they themselves meet quality expectations in order to satisfy their customers and also to mitigate any risks that come from providing problematic translations to their customers. Quality evaluation for requesters is complicated by the fact that, in most cases, they have no direct way to evaluate quality because they do not speak the target languages.

Evaluation of commercially produced translations as part of an overall translation quality management process may be the most common scenario in which translation quality evaluation plays a role, but evaluation is also involved in a variety of other circumstances. For example, developers of translation technology such as machine translation (MT) may need to understand how changes to their systems impact the quality of their output. Educators may evaluate translations produced by students in educational environments to identify problems and suggest improvements.

Traditionally translation quality is evaluated by bilingual reviewers, who examine source and target texts to determine whether the translation meets requirements. Unfortunately,

these evaluations are often subjective statements of taste, and reviewers can disagree, at times dramatically. For example, if a translation is requested from English (with a single level of grammatical register) into Korean (with many levels), it may not be clear which register should be used, and reasonable reviewers could disagree, resulting in a translation being considered very good by one reviewers and inadequate by another.

In the absence of specific guidelines, "quality" be quite quite nebulous in such cases, something that depends on the personal expectations of reviewers. Such problems suggest that a new approach for evaluating quality is needed. (For more on the difficulties of defining *translation*, *quality* in general, and *translation quality* in specific, see the two articles by Koby *et al*. and the article by Fields *et al*. in this issue of *Tràdumatica*.)

As a response to this situation, the notion of a single, universal, objective metric for translation quality is appealing, but such a metric would require that translation have universal features and objective, a notion that is highly questionable (see House 2008 for an overview of this issue). Even recognizing the unlikeliness of finding "the" translation quality metric, there was a shift in the late 1990s and through the first decade of the twenty-first century to use of "objective" lists of error types, exemplified by the LISA QA Model[1]—a quasi-standard released by the now defunct Localization Industry Standards Association (LISA) based on the best practices of a group of prominent localization service providers—and SAE J2450, [2] a standard developed by SAE International for evaluating the quality of automotive service manuals.

Although very different in their details, these two specifications shared certain core features. In specific, they provided a list of error types (twenty-five for the LISA QA Model and seven for SAE J2450) that could be correlated to specific errors in texts. In both metrics, errors are also ranked for severity: for example, a misspelling of *receive* as *recieve* would be considered less important than the omission of a *not* in a critical instruction. Errors are then counted to provide a quality score—normally presented as a percentage—that would, in principle, be more objective than the subjective opinions of reviewers. By setting quality thresholds (e.g., stating that translations must have a score of 99% or higher), requesters and providers would know whether translations meet requirements.

Both of these models saw fairly widespread adoption by localization providers, either in spreadsheets or in "scorecard" environments (scorecards provide a user interface that allows the user to quickly select an error category, as shown in Figure 1). They were often customized to reflect specific requirements. The wide-spread customization points to one of the major limitations of such models: they were essentially either one-size-fits all approaches (like the LISA QA Model) or limited in applicability (like SAE J2450), so any changes broke conformity with the specification and limited the interchangeability of evaluations (e.g., a 99% in one model might be a 93% in another).

Customization also led to the problem that quality evaluation practices differed between evaluators in an almost random fashion. In addition, in research conducted in the QTLaunchPad project, it was found that metrics developed for specific projects were often then adopted within an organization and survived based on institutional inertia, even when there were problems with the metrics they used.

Although customization seems to be necessary, maintaining some degree of consistency and interchangeability between different metrics is desirable, and industry was headed in this direction: before its insolvency in 2011, LISA had begun the process of developing a new version of the LISA QA Model that would have allowed for customization for specific needs.

[1] As the LISA QA Model is no longer available there is no normative public reference for it currently available.

[2] http://www.sae.org/standardsdev/j2450p1.htm. Note that the committee that created SAE J2450 never claimed it to be suitable for any type of translation other than service manuals, although others have used it for a wide variety of kinds of translations.

However, this development had not proceeded beyond internal drafts before LISA ceased operations in April 2011.

In addition to the models described above, a number of translation quality checking tools have been developed over the years that are used for the automatic detection of errors. These tools generally focus less on providing a quality score than on identifying specific errors so they can be rectified. The errors are often formal in nature (e.g., a mismatch in formatting tags between source and target, content that remains in the source language, extreme difference in segment length, and dates or numbers that may be in the wrong format), but also include items such as improperly translated terminology.

BLEU (Papineni et al. 2002) and METEOR (Banerjee & Lavie 2005, see also http://www.cs.cmu.edu/~alavie/METEOR/) are based on completely different principles than are applied to human translation and the results have generally been incomparable between the two. MT evaluation methods have generally involved examining how similar MT output is to one or more human reference translations, meaning that MT quality is usually evaluated only for previously translated texts and these methods cannot be used in production environments.

Multidimensional Quality Metrics (MQM)

The Multidimensional Quality Metrics (MQM) system was developed in the European Union-funded QTLaunchPad project to address the shortcomings of previous quality evaluation. It has its roots in the efforts to update the LISA QA Model (the authors of this article were the primary developers of MQM and one of them worked on the uncompleted revision of the LISA QA Model) and therefore takes many of its principles from the LISA QA Model. Unlike previous quality metric models, however, MQM was designed from the beginning to be flexible and to work with other standards in order to integrate quality evaluation into the entire document production cycle. It was aimed to avoid the problems of both a one-size-fits all model and the fragmentation of unlimited customization.

The central component of MQM is a hierarchical listing of issue types. The types were derived in a careful examination of existing quality evaluation metrics and the issues found by automatic quality checking tools. The issues were restricted to just those dealing with language and format, leaving out issues from the LISA QA Model that deal with *project* quality (e.g., issues related to on-time delivery or project completeness). The resulting list of issues can be visualized as shown in Figure 2.

The current version of the MQM hierarchy (as of October 2014) contains 114 issue types and represents a generalized superset of the issues found in existing metrics and tools.[3] It is generalized in the sense that some tools go into considerable more detail in particular areas, such as Okapi CheckMate, which has eight separate whitespace related categories, all of which are generalized to MQM *Whitespace*; rather than try to capture all possible detail, some of the fine-grained details are abstracted away and if a particular application needs to use them, they can be implemented as custom MQM extensions. The issues types are defined in the MQM definition (http://qt21.eu/mqm-definition).

The hierarchy depicted in Figure 2 is a crucial part of MQM. Every node (including very high-level ones like *Accuracy* and *Fluency*) can serve as an issue type, and children of an issue represent specific cases of the parent issue. As a result an MQM metric can be declared at various levels of granularity. For general-purpose quality evaluation, for instance, a relatively coarse metric is likely to suffice while for diagnosis of problems with an MT system a

[3] Although MQM was derived based on an examination of industry practice, many of the issue types (not surprisingly) show similarity to the sorts of issues identified in House (1997).

much more detailed metric that focuses on specific issues relevant to system development would be appropriate.

In other words, MQM addresses the problem seen in one-size-fits-all metrics (such as the LISA QA Model) by defining a model to declare multiple metrics rather than one single metric. By providing a standardized vocabulary for declaring issues, however, it allows metrics to be compared and, where there is overlap, the results can also be compared. It is also important to note that, with limited exceptions,[4] MQM is intended to be language neutral and therefore applicable to any language pair.

Note that although MQM supports a very fine-grained approach to quality evaluation, best practice for human evaluation would dictate that metrics be only as complex as needed to support a particular evaluation task, in part because it is more difficult for evaluators to distinguish between fine-grained categories than higher-level ones. For example, if the evaluation only needs to know if there are grammatical errors in a translated text, it would not make sense to include all the subtypes of *Grammar*; on the other hand, it would make sense to address the subtypes if MQM is used to evaluate the output of an MT system and it is important to understand the precise types of problems found with the system. (For automatic error-detection tools, however, it makes sense for them to be as precise as the system allows and to allow subsequent processes to decide whether or not to convert their results to more general types.)

Verity

One of the contributions of MQM is a concept called *Verity*. Verity refers to *extra-linguistic* truth correspondence in texts. It refers to issues with the text that relate to the broader world. For example, an employment contract written in the United States (a common-law jurisdiction) and translated into Spanish for an audience in Spain (a civil-law jurisdiction) might need to be substantially altered by a highly skilled professional with considerable expertise in both legal systems, with legal constructions parallel to but distinct from those used in the U.S. version for the Spanish text. By contrast, if the contract is translated for Spanish in the U.S., the appropriate translation will look *very* different from a version created for Spain. Such differences go well beyond the locale appropriateness of terms and language. In a recent informal survey of language professionals, we found *substantial* disagreement about whether these sorts of change should be considered part of translation or not; nevertheless, MQM takes a broad definition of translation and includes these issues.

Because it is extra-linguistic, Verity deals with issues that are outside of the traditional notions of Accuracy and Fluency; in other words, Verity in many cases cannot be evaluated simply be reading the translation or comparing it to the source, but instead requires the evaluator to consider the text in its intended environment. Using the above example, the quality of the translated text could not be evaluated unless the intended audience and purpose of the text is known, and a fluent and accurate translation might be inappropriate (perhaps critically so) when translated for Spain versus a U.S.-based Spanish-speaking audience because it contains provisions that apply in the U.S. but which are contrary to the law in Spain. (Although similar to House's notion of "intervention," Verity is more practically focused on the suitability of the text for a given audience and context, rather than with ideological shifts in translations. See House 2008:16.)

While a legal contract might be an extreme example, such problems emerge frequently in other contexts. For example, a Chinese translation of a text that refers to a U.S.-based hotline number staffed by English monolingual support staff would be inappropriate for the translated

[4] The category *Single/double-width (CJK only)* is specific to font considerations for East Asian languages and does not apply to other languages.

text, even if the text is accurately and fluently translated. Similarly, legal notices, descriptions of available product options, descriptions of promotional offers, and other such aspects of texts may need to be changed dramatically during the translation process.

Because Verity falls outside of the normal range of translation quality, it must be noted that MQM identifies issues/errors in texts, but does not assign blame. A Verity problem may exist even though the translator had no way of knowing it was a problem. In other cases the translator may detect a problem but have insufficient information to repair it (in which case the translator should bring it to the attention of the requester). Regardless of *how* Verity problems are to be addressed, MQM provides a way to identify them.

Evaluation methods

MQM is agnostic concerning the evaluation method. Most of the focus on MQM has gone into defining *analytic* methods that seek to identify specific errors in a translation in order to quantify them. They thus focus on portions of the text and rely on the positive identification of specific errors. In such analytic methods the errors are seen as deviations from a text that fully complied with specifications and "errors" that do not lead to a deviation from specifications are not counted: for example, if specifications for a service manual state that stylistic concerns are not important, poor style would not be counted as an error, even though incorrect terminology would be.

In contrast with analytic measures, *holistic* measures focus on an evaluation of the text as a whole, often based on questions, such as "To what extent the text meet expectations for accuracy?" with appropriate scalar values. Holistic measures are advantageous for situations where evaluators must quickly determine how well texts meet overall specifications, but do not identify specific errors for remedial action. (By contrast, analytic methods are comparatively labor intensive but result in the identification of specific errors that can be addressed.)

Although the focus of most MQM usage to date has focused on analytic methods, an analytic metric can easily be adapted for holistic usage by taking the issue types in the metric and converting them to questions about the text with accompanying scalar values. For example, consider the following simple MQM metric:

- Accuracy
  - Mistranslation
  - Omission
- Fluency
  - Stylistic
  - Grammar

In this metric, there are six issues, with *Accuracy* and *Fluency* serving for any issues not covered by their children. Conversion to a holistic metric would result in something like the following:

	No	Some	Yes
Accuracy			
[1a] Has any content been inappropriately *omitted* from the translation?			
[1b] Does the text exhibit any *mistranslations* that change the intended meaning of the text?			
[1c] Are there any other *accuracy* issues that affect the translation?			
Fluency			
[2a] Are there any *grammar* problems in the translated text?			
[2b] Are there any *stylistic* problems in the translated text?			
[2c] Are there any other *fluency* problems in the translated text?			

Note that the order of issues within the sections has been reversed in order to present more specific issue types before less specific ones (e.g., *Mistranslation* is presented before the more general *Accuracy*). Other than the change in order, the holistic metric corresponds precisely to the analytic metric. However, for complex analytic metrics with many issue types, some simplification is generally in order when converting to an analytic measure: questions such as "Are there any problems with the usage of *function words* in the translated text?" may not be particularly useful in a holistic evaluation to determine whether a text meets specifications (although they may be in some circumstances).

Relationship to ITS 2.0

Published in 2013, the Internationalization Tag Set (ITS) 2.0 specification contains a *localization quality issue type* data category.[5] The values of this data category were derived from an early version of MQM that contained fewer, less granular, data categories. It is possible to losslessly map the ITS 2.0 values to a subset of MQM and to map MQM issue types to ITS 2.0. The MQM-to-ITS 2.0 mapping may result in some data loss, depending on the MQM metric in use since ITS 2.0 cannot fully represent the granularity of MQM. Nevertheless, for the most common error types, MQM can be represented in ITS. Part of the ongoing work on MQM will be to further develop a formal representation that better integrates MQM with ITS-compliant systems.

Relationship to TAUS DQF

A frequent question asked about MQM is how it relates to the TAUS Dynamic Quality Framework (DQF).[6] Developed in roughly the same time frame, MQM and DQF may seem to be in conflict at first. However, as of late 2014, the two initiatives are in contact and working to harmonize their efforts. A close examination reveals them to be largely complementary. MQM provides a way to describe arbitrary metrics in a standardized fashion but does not provide guidance on the interpretation of the results. The DQF, by contrast, does not seek to describe all possible translation quality metrics but does provide guidance on interpreting quality evaluations for specific scenarios. Part of the ongoing work is to bring the two frameworks together such that any DQF metric can be described in MQM. This harmonization is planned as a key aspect in a forthcoming EU-funded project.

[5] http://www.w3.org/TR/its20/#lqissue-typevalues
[6] http://evaluation.taus.net

Implementations and Future Development

MQM is currently used by three translation-oriented tools: the open-source translate5 system[7] (which added MQM support with funds from the QTLaunchPad project), the commercial XTM system, and an open-source server-based "scorecard" tool developed as a partnership between the QTLaunchPad project and the Brigham Young University Translation Research Group (TRG) with additional development at the German Research Center for Artificial Intelligence (DFKI) in Berlin funded by the Caribbean Regional Information and Translation Institute (CRITI), based in Paramaribo, Suriname. A number of localization service providers have adopted MQM-based approaches to quality evaluation and the Mozilla Project is in the process of moving quality evaluation processes for Firefox to an MQM-based approach.

The next step in development is to turn MQM over to the wider community. Currently DFKI and other stakeholders are working closely with TAUS, the Globalization and Localization Association (GALA), and other interested parties to develop MQM further as a standard within an appropriate international standards body. This development should enable MQM to be sustained by interested parties in the translation community beyond the European Union-funded project that gave rise to it.

Acknowledgements

Development of the Multidimensional Quality Markup framework was supported by the QTLaunchPad project with funding from the European Union's Seventh Framework Programme for research, technological development, and demonstration under grant agreement no 296347.

Special thanks are due to the anonymous reviewers for their suggestions for clarifications and corrections.

References

Banerjee, Satanjeev & Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL).

House, Juliane. 1997. *Translation Quality Evaluation: A Model Revisited*. Tübingen: Gunter Narr.

House, Juliane. 2008. Beyond intervention: Universals in translation? *Trans-Kon* 1(1):6–19.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 311–18.

[7] See http://www.qt21.eu/launchpad/content/resources for more information.