MQM Scoring Model

Contents

Tables

Overview

Contents

Stages of Translation Quality Evaluation

Contents

Translation quality evaluation (TQE), as specified in this standard, comprises one manual stage—error annotation—and three automatic stages—error compilation, quality analysis, and quality rating. They proceed from an evaluation text to a list of errors to a matrix of error counts to calculated quality measures and on to organization-specific derivative quality ratings. This document makes passing reference to all four stages, but it focuses on the quality analysis stage. The MQM Scoring Model drives the quality analysis.

The MQM Scoring Model presented here presumes that the translation product that is the subject of the quality evaluation has gone through an error annotation process. During error annotation, a quality evaluator uses an error annotation tool to examine the evaluation text carefully and identify errors from an active MQM-compliant error typology. Each error is annotated with an error type, an error severity level, an error root cause, and other relevant error data.

Error annotation is followed by error compilation, which is an automatic process. Error compilation produces two aggregate formats of the errors identified in the error annotation stage, in files that are independent of the error annotation tool. The first is an error list, which is a sequential list of all of the errors found, including key descriptive data for each error.

The second data format is an error summary. The error summary includes a matrix of error counts for each combination of error type and severity level. Filtering based on error root causes controls which errors are reflected in the error counts. Duplicate error conflation may be applied

to reduce the error counts of specific error types if they are tied to iterations of the same error string. Root cause filtering and duplicate error conflation are not elaborated further here. The MQM Scoring Model, in any case, operates the same way irrespective of any error count modifications that may have been applied because of root cause filtering or duplicate error conflation.

A second value in the error summary is an evaluation word count. The evaluation word count is the number of words in the source text of a translation product that is the object of a quality evaluation. The error counts and the evaluation word count are outputs from the error annotation and error compilation processes and inputs to the quality analysis process.

There are other inputs to the quality analysis process, namely, the scoring parameters, which are central to the scoring model. The five scoring parameters fall naturally into two distinct groups—determinative parameters and scaling parameters. The determinative scoring parameters contribute in a substantial, discriminative way to define a quality measure. They include a range of penalty values applied to severity levels and a range of weights applied to the accumulated penalty values for each error type. They capture the relative contributions of different severity levels and error types to the overall quality measures.

The three scaling parameters combine to scale and position final quality measures. The reference word count provides a standard document length for comparing accumulated penalty values to. The maximum score value sets an upper limit for the final quality score and sets a range against which the quality scores are scaled. The penalty scalar expands or contracts the target range for quality measures. The scoring parameters combine with the error compilation outputs and the formulas in the scoring model to produce the targeted quality measures.

Quality analysis is the third major stage in quality evaluation. In many evaluation projects it is also the last stage. It may be followed by a fourth stage, a quality rating stage. Quality ratings are custom measures that are defined by organizations based on the quality measures generated in the quality analysis stage, but they may also reflect any aspect of a translation product, translation process, or quality evaluation process.

Introduction to Quality Measures

Contents

MQM is an analytic translation quality evaluation standard. This means that individual errors are identified and tabulated, and quality measures are calculated from the bottom up. This process is described in mathematical detail in the MQM Scoring Model Architecture section that follows. Suffice it here to say that the accumulated penalty totals for each error type are summed and then divided by the word count of the evaluation text, and this number is combined with the scoring parameters to generate the quality measures.

The MQM Scoring Model recognizes two quality measures—the overall normed penalty total (ONPT) and the overall quality score (OQS). The overall normed penalty total is, first, a penalty

total. It is based on the grand total of penalty totals for the error types that are active in the error annotation stage.

To make these penalty totals comparable across evaluations, the penalty totals are normed. First, they are normed by being divided by the evaluation word count. Second, they are normed by the application of two scoring parameters—the reference word count (RWC) and the penalty scalar (PS). The reference word count gives a frame of reference for the penalty totals, for example, the number of error penalties per 1,000 words. The penalty scalar applies to both the overall normed penalty total and the overall quality score, and it is used to scale their values up or down in order to get them into expected ranges. The minimum and best value for the ONPT is 0.

Quality stakeholders are generally familiar with quality scales in which the best value is at the top of a range, say, from 0 to 100. The second quality measure, the overall quality score, is this type of a quality measure. The maximum and best value for the OQS is itself a variable, the maximum score value (MSV), which defaults to 100. The penalty scalar is used to produce quality scores that users of the evaluation find meaningful, such as a traditional grading scale.

The adjective "overall" is applied to both the overall normed penalty total and the overall quality score. The use of this term highlights the fact that a single ONPT and a single OQS are generated for the quality evaluation as a whole. It contrasts with the descriptor "error type", which can also be coupled with these quality measures. The application of the quality measures to error types is not covered in this document, but an error type normed penalty total (ETNPT) and an error type quality score (ETQS) can be defined for a particular error type dimension (top-level error type) or for an error type at another level in the MQM error typology.

Understanding the MQM Scoring Model requires a clear grasp of the factors that are used to define it. They are as follows:

- indices, to give meaning to the subscripts used in sigma notations and other formula elements,
- error compilation outputs, to refer to the data generated in a particular error annotation project and tallied in the error compilation stage,
- scoring parameters, the quantitative data that define a particular MQM metric, and
- quality measure calculations, the formulas used to calculate ONPT and OQS.

All of these values—inputs and outputs—are numbers. There is one additional set of functions that have string values, namely the labels functions. They relate the indices for error types and error severity levels to the descriptors for error types and error severity levels that are normally used—*Accuracy*, *mistranslation*, *addition*, etc.; *neutral*, *minor*, *major*, and *critical*. Assigning label mappings to the numerical indices links them to familiar terminology, such as standard error type names, or custom terminology, such as localized error type display names. At the same

time, it enforces rigorous use of algebraic notation in defining the elements that contribute to the quality measure calculations.

MQM Scoring Model Architecture

Contents

The MQM Scoring Model is an extension of the scoring model given in MQM 1.0. It combines error counts with scoring parameters to produce usable, defensible quality measures. This measure definition process is elaborated below under Reference Metrics and Quality Scores. The resulting quality measures are easy to use and easy to apply in most contexts.

Indices

Contents

Two indices are used in the quality measure calculations for the MQM Scoring Model:

Error type index i, with upper bound m = 42, or other count of active error types

i = 1, 2, 3, . . . 42, for example, based on the error types in the DQF-MQM error typology

Error severity level index j, with upper bound n = 4 (for default neutral, minor, major, and critical error severities)

j = 1, 2, 3, 4

Labels

Contents

Two labels are used to describe the fundamental components that are at the core of the quality measure calculations in the MQM Scoring Model:

ETN_i = error type name, for error type i (with the names of error type dimensions capitalized)

ETN_1 = “Accuracy”,
ETN_2 = “mistranslation”,
ETN_3 = “addition”, . . .
ETN_{42} = “Internationalization”

$ESLN_j$ = error severity level name, for error severity level j

$ESLN_1$ = “neutral”,
$ESLN_2$ = “minor”,
$ESLN_3$ = “major”,
$ESLN_4$ = “critical”

The ETN_i and $ESLN_j$ labels are not actually used in the MQM Scoring Model. They are used to relate the indices to standard MQM and DQF-MQM error types and severity levels. As a result, quality managers and quality evaluators may refer to a “mistranslation error”, but in the scoring

model, this term is equivalent to ETN_2. Referring to the error type name that corresponds to ETN_2 identifies 2 as the error type index to be used in related formula elements.

Error Compilation Outputs

Contents

This is where the quality measure calculations begin. Two essential data elements are output from the error annotation and error compilation quality evaluation stages:

EC_{ij} = error count, for error type i and error severity level j

Unit: errors (integer)
Range: $0 \leq EC_{ij}$

EWC = evaluation word count, the number of words in the evaluation text, that is, the source text of a full or sample translation product that is the object of a quality evaluation

Unit: words (integer)
Range: $1 \leq EWC$

Scoring Parameters

Contents

Five essential parameters from an MQM-compliant TQE metric factor into the quality measure calculations:

SLP_j = severity level penalty, for error severity level j

SLP_j = 0, 1, 5, 25 or 0, 1, 4, 16 or 0, 1, 3, 9 . . . .

Unit: error penalties (integer or decimal number)
Range: $0 \leq SLP_j$
Default values: 0, 1, 5, 25

ETW_i = error type weight, for error type i

ETW_i = 1.3, 2.0, 1.6, . . . 1.7
These values are illustrative only; the default value for all error type weights is 1.

Unit: error weights (integer or decimal number)
Range: $0 \leq ETW_i$
Default values: 1, for all error types

Note that an error type weight of 0 effectively removes that error type from the quality analysis, notwithstanding the number of error counts of that error type at different severity levels.

RWC = reference word count

100 or 1,000 or 10,000 . . . .

Unit: words (integer)
Range: $1 \leq RWC$
Default value: 1,000

The RWC adjusts the ONPT so that it moves ONPT values up from the fraction range into the integer range. Without the RWC, ONPT values would be between 0 and 1. With an RWC of 1,000, for example, the ONPT would have values in the tens and hundreds, which are easier for most people to work with than units in the hundredths and tenths.

MSV = maximum score value

5 or 20 or 100 . . . .

Unit: quality score (integer)
Range: $0 < MSV$
Default value: 100

The MSV represents a perfect upper score on a scale that is familiar to the quality manager and users of the evaluation. The default MSV is 100, which was a hardcoded value in the MQM 1.0 scoring model and is a widely accepted industry norm.

PS = penalty scalar

0.5 or 1.0 or 2.5 . . . .

Unit: scalar (decimal number)
Range: $0 < PS$
Default value: 1.0, subject to recalibration by an organization for factors such as target language, text type, or quality expectations

Setting the PS to 1.0 is a first approximation. Based on the effect of other scoring parameters, the PS could be adjusted up or down to shift the ONPT and OQS values into quality ranges that match common user expectations, so the default PS may vary in different implementations.

In any event, the PS should be based on data from multiple evaluations and validated against independent measures of quality. This process is described in great detail in the Scoring Parameter Settings section below. The PS also needs to correct for the hidden scaling effects built into different severity penalty sequences and error type weight sequences, so at some point the PS may have an updated, data-driven default value in the MQM Scoring Model as well.

Quality Measures Contents

There are six values calculated below. Four of the values—error type penalty total (ETPT), absolute penalty total (APT), per-word penalty total (PWPT), and overall quality fraction

(OQF)—are intermediate values used in the calculation of a derivative quality measure. The two overall quality measures are overall normed penalty total (ONPT) and overall quality score (OQS). Organizations may base their quality evaluations on ONPT and not make reference to OQS, or they may treat ONPT as an intermediate value used to calculate their main quality measure, OQS. They may also use both ONPT and OQS, but for different purposes. For instance, OQS can be used in high-level quality reports, while ONPT is used in the definition of particular quality ratings, such as pass/fail ratings.

$ETPT_i$ = error type penalty total, for error type i

Unit: weighted error penalties (integer or decimal number)
Range: $0 \leq ETPT_i$, best (min) value = 0

APT = absolute penalty total

Unit: weighted error penalties (integer or decimal number)
Range: $0 \leq APT$, best (min) value = 0

PWPT = per-word penalty total

Unit: normed error penalties (decimal number)
Range: $0 \leq PWPT$, best (min) value = 0

ONPT = overall normed penalty total

Unit: normed error penalties, or error penalties per RWC words (decimal number)
Range: $0 \leq ONPT$, best (min) value = 0

OQF = overall quality fraction

Unit: quality fraction (decimal number)
Range: $OQF \leq 1$, best (max) value = 1

Users should not view 0 as the lower end of the OQF range. OQF may, in fact, be a negative number, with no clear lower bound.

OQS = overall quality score

Unit: quality score (decimal number)
Range: $OQS \leq MSV$, best (max) value = MSV (default 100)

Users should not view 0 as the lower end of the OQS range. OQS may, in fact, be a negative number, with no clear lower bound.

If MSV = 100, the default setting, then some users might view OQS as a quality percentage. It is not; it bears no relation to any definition of percentage. It is a quality score, whose value can only be properly interpreted if training and instructions given to the evaluators and

details of the error typology, the error annotation, the error compilation, and the quality analysis, in particular, the five scoring parameters introduced above, are all known.

If these scoring model parameters are not stated or implied, the quality measures, ONPT and OQS, are meaningless. And if two evaluations do not share these parameters, a comparison of their quality measures is not directly possible. That is one of the motivations for using a standardized implementation of MQM, such as the DQF-MQM metric, namely, that quality measures are comparable, and reliable benchmarks across organizations, translators, projects, languages, and time are possible.

Quality Formulas Contents

The formulas used to compute ETPT, APT, PWPT, ONPT, OQF, and OQS are as follows:

$$ETPT_i = \left(\sum_{j=1}^{n} (EC_{ij} \times SLP_j)\right) \times ETW_i$$

$$APT = \sum_{i=1}^{m} ETPT_i$$

$$PWPT = APT / EWC$$

$$ONPT = PWPT \times PS \times RWC$$

$$OQF = 1 - (ONPT / RWC)$$

$$OQS = OQF \times MSV$$

The OQS can also be calculated directly from the PWPT and the scaling parameters:

$$OQS = (1 - (PWPT \times PS)) \times MSV$$

The quality measures and the intermediate calculations can be mapped to the error compilation outputs and scoring parameters upon which they are based (Table 1). The error compilation outputs, EC and EWC, and the scoring parameters, SLP, ETW, RWC, PS, and MSV, enter the formulas at different points in calculating the intermediate values, ETPT, APT, PWPT, and OQF, and the final quality measures, ONPT and OQS. . In addition to the error compilation outputs and the scoring parameters, each quality value is derived from the quality values that precede it.

Table 1: Quality Value Dependencies Contents

Quality Value	Error Compilation Output	Scoring Parameter
Error type penalty total	Error counts	Severity level penalty Error type weight
Absolute penalty total	—	—
Per-word penalty total	Evaluation word count	—
Overall normed penalty total	—	Reference word count Penalty scalar
Overall quality fraction	—	—
Overall quality score	—	Maximum score value Penalty scalar

Quality Score Applications and Comparisons Contents

Quality measures do not have value in the abstract. They have value if they can be compared to stakeholder expectations of what good and bad scores would be, and they are informative to the degree that they can be compared to quality measures from previous organizational quality evaluations and to broader industry quality norms.

Two or more quality measures may not be directly comparable. Nevertheless, it is possible in many cases to derive missing quality measures and to convert one set of quality measures to another form that is comparable to a reference set. The methodology for accomplishing these practical goals is described in the sections that follow.

Metrics and Parameters Contents

Before examining the procedures for completing and comparing quality measures and for fine-tuning the penalty scalar, it is useful to consider the structural elements of the MQM Scoring Model upon which they depend. The next section reviews how scoring parameters are applied at different stages of the quality evaluation process to define different metrics. That is followed by a section that subdivides the quality analysis stage into two processes. Using a spreadsheet metaphor, these are referred to as *top-half* and *bottom-half* processes. Everything that follows at that point takes place in the bottom-half calculations and derives from the per-word penalty total and the scaling parameters.

Defining MQM-Compliant Metrics Contents

MQM is not a translation quality evaluation metric; it is a framework for defining MQM-compliant metrics. Translation quality evaluation, in the MQM model, comprises four stages, described above. Each of the stages affords choices and references parameters that define the translation quality metric as developed in that stage.

This approach has two distinct advantages. First, elements that were implicit or static in MQM 1.0 and other analytic quality standards are explicit and open to change in MQM 2.0. For example, in MQM 1.0, the quality score was measured against a hardcoded 100-point scale. In MQM 2.0, the quality score is scaled by a parameter, the maximum score value, that defaults to 100.

Second, by separating out stages, functions, and parameters, it is possible to reuse existing work up to a particular stage and specify or change a default parameter to yield a new metric and derive custom quality measures from that point going forward. To borrow a term from software development, the MQM model supports *late binding*, meaning that parameters are specified and values are calculated at the last possible point. The advantage of this approach is that it strikes a balance between conservation of previous effort and retention of future metric options.

For example, after defining a set of active error types, severity levels, and root causes and applying these in a quality evaluation project, a compliant error annotation tool outputs an error list for that evaluation project. The error compilation stage uses the error list and applies root cause filtering and duplicate error conflation to create an error summary. One error summary could include all annotated errors, regardless of their root cause. Another could just include errors that had a root cause of *translator error* or, alternatively, errors with any root cause other than *translator error*. These divergent outcomes—and follow-on quality measures—hinge on the root cause filters applied in the error compilation stage.

This same principle of defining alternative metrics based on scoring parameters applies in the quality analysis stage and in the calculations that are part of the MQM Scoring Model. There are two inputs to the quality analysis stage (the outputs from the error annotation and error compilation stages), namely, the matrix of error counts for different error types and severity levels and the evaluation word count.

There are also five scoring parameters. The two determinative parameters are the severity level penalties and error type weights, and the three scaling parameters are the reference word count, the maximum score value, and the penalty scalar. These values are combined, using the formulas stated above, to calculate two quality measures—the overall normed penalty total and the overall quality score.

Top-Half and Bottom-Half Quality Analysis

Contents

This system of data points, parameters, and output values lends itself to being broken down into two discrete systems internal to the quality analysis stage. As mentioned above, these systems can be described as *top-half* and *bottom-half* calculations. The top half takes as input the error compilation outputs, applies the determinative parameters, and outputs two intermediate calculations—the absolute penalty total (APT) and the per-word penalty total (PWPT). In the

bottom half, various formulas apply scaling parameters to the PWPT to produce the two quality measures.

In the top-half calculations, the only relevant parameters are the severity level penalties and the error type weights, the latter of which are all set to 1 by default. Different severity penalty sequences have been applied in different MQM implementations over time. To apply a metric using the now-default 0-1-5-25 to a legacy metric using another sequence, such as 0-1-5-10 (used in LISA QA and DQF-MQM) or 0-1-10-100 (used in MQM 1.0), it would be necessary to have access to the original error counts and evaluation word count. Then, in order to create quality measures that are analogous to those produced using a legacy metric, the legacy determinative parameters would have to be substituted for the current default parameter values. Of course, the same sort of top-half adjustment could be made going in the other direction, to convert a legacy evaluation to an MQM 2.0 default evaluation.

Frequently, however, efforts to compare results for the same quality evaluation across different metrics do not have access to the complete set of outputs from the original error annotation and error compilation stages. Nevertheless, given one set of final quality measures and the scaling parameters used to create them, a number of useful comparisons and conversions can be made using different sets of scaling parameters.

These kinds of comparisons are effected in the bottom half of the MQM Scoring Model. These comparisons require a complete set of scaling parameters and any one of the following values: the per-word penalty total (PWPT), the overall normed penalty total (ONPT), or the overall quality score (OQS). All of the discussion that follows applies to this bottom-half analysis, and it requires this set of data points.

Many apples-to-apples comparisons can be made in the calculations that follow from the PWPT. For example, if two evaluations share the reference word count and the maximum score value but have different penalty scalars, a simple formula can be introduced to convert one set of quality measures to the other based on the differences between the two penalty scalars, as shown below.

Incomplete Quality Measures

Contents

For a particular quality evaluation, normally both the overall normed penalty total and the overall quality score are known, but if one of them is missing and the other is known, it is possible to derive the missing quality measure after the fact.

Given the overall normed penalty total and the scaling parameters, it is straightforward to calculate the associated overall quality score:

$$OQS = (1 - (ONPT / RWC)) \times MSV$$

Given the overall quality score and the scaling parameters, it is possible to work backward to calculate the overall normed penalty total:

ONPT = (1 – (OQS / MSV)) × RWC

Given the overall normed penalty total and the scaling parameters, it is possible to calculate the per-word penalty total, but no other top-half values:

PWPT = ONPT / (RWC × PS)

Likewise, given the overall quality score and the scaling parameters, it is possible to calculate the per-word penalty total:

PWPT = (1 – (OQS / MSV)) / PS

Alternative Scoring Parameters Contents

As stated above, the MQM translation quality evaluation model applies late binding of parameters to previous data values wherever possible. In particular, given a particular quality evaluation and the resulting error counts and evaluation word count and a particular set of determinative scoring parameters, a unique per-word penalty total can be calculated. This per-word penalty total can then be used to calculate a wide range of values for the overall normed penalty total and the overall quality score, depending on the values of the scaling parameters. This means that it is possible, in turn, to apply different scaling parameters to existing quality scores and generate new quality scores based on the new scaling parameters.

The key to calculating these new values is to calculate the per-word penalty total and then apply the new scaling parameters, using the original quality formulas in the MQM Scoring Model Architecture section. The formulas that accomplish this are given below.

If the acronyms RWC, PS, and MSV are reserved for the original scaling parameters and the acronyms ONPT and OQS are reserved for the original quality measures, then NRWC, NPS, and NMSV can be used for the new scaling parameters, with NONPT and NOQS used for the new quality measures. PWPT remains the same with either set of scaling parameters or quality measures.

NONPT = PWPT × (NRWC × NPS)

NOQS = (1 – (PWPT × NPS)) × NMSV

Given a particular set of error counts, evaluation word count, and determinative parameters, there is a unique per-word penalty total (PWPT), which can be calculated from the original quality measures and original scaling parameters. This calculation can be substituted in the above formulas to yield formulas that depend on the original quality measures and scaling parameters, in addition to the new scaling parameters, but which make no direct reference to the PWPT:

$$NONPT = (ONPT / (RWC \times PS)) \times (NRWC \times NPS)$$

$$NOQS = (1 - (NONPT / NRWC)) \times NMSV$$

Scoring Parameter Settings

Contents

The scoring parameters are the elements in the quality analysis stage that determine the translation quality evaluation metric, so it is critical that the values assigned to these parameters are well chosen, that is, based on a good understanding of what the parameters are designed to do and how best to set good default values for them. The following section describes ways to do set good parameter values.

Scoring Parameter Derivation

Contents

A classification can be made of the scoring parameters based on how their values are specified and how they are interpreted. The parameters fall into three groups: *conventional*, *subjective*, and *empirical*. The reference word count and the maximum score value are conventional; the severity level penalties and error type weights are subjective; and the penalty scalar is empirical.

Conventional Parameters

Contents

Conventional parameters (RWC and MSV) are not calculated or fine-tuned; they are simply chosen to suit organizational conventions, locale norms, or industry best practices. They do not have to be defended for any other reason than that they have a useful value that is commonly accepted by a target audience of quality managers and quality stakeholders. Consequently, a reference word count defaults to 1,000 because that is the reference word count against which overall normed penalty totals are most often measured, and the maximum score value defaults to 100 because most implementers, by convention, measure their overall quality score on a 100-point scale.

An implementer can certainly choose to use nondefault RWC and MSV values, but then it is incumbent on them to indicate this when they share their quality measures. RWC and MSV are both scoring parameters, so it is easy to convert the resulting quality measures to the values that would have been attained using the default parameters, applying the formulas given in the previous section.

Subjective Parameters

Contents

Subjective parameters (SLP and ETW) are essentially educated guesses. They are not wholly conventional, and they are not likely to become conventional and universal, nor can they be determined objectively and mathematically like an empirical parameter. Nevertheless, there are better and worse ways to set them.

Two things can be said about subjective parameters. First, while it is not demonstrably feasible to offer best values for them, it is possible to provide best practices for setting them. Second, standard practice suggests that the default values for severity level penalties and error type weights suggested here are good, reasonable values. They can be used confidently—and consistently—until someone can demonstrably offer a better set of values.

There are a few best practices for setting severity level penalties. First, it should be obvious that a severity penalty sequence of 0-1-5-25 with a penalty scalar of 2 will produce the exact same quality measures as a sequence of 0-2-10-50 with a penalty scalar of 1. For ease of comparison, then, it is a best practice to always set the minor severity penalty to 1 (scaling the major and critical penalties accordingly) and relegate overall scaling factors to the penalty scalar.

Although, as stated, the ratio between successive severity penalties is subjective, this difference can logically represent the number of, for instance, minor errors that are deemed to have the same impact on quality as one major error. This ratio can then be used in setting the major and critical severity penalties.

The other subjective parameter, the error type weight, is used to handicap the contributions of different error types and moderate the impact of different severity levels across the error types. However, an error type weight greater than 1 overstates the impact of minor severities of major error types and may understate the impact of major severities of minor error types. This is problematic if severity levels are intended to have the same meaning and significance across all error types.

The primary function of the error type weights is to give greater weight to the error types that are seen as more significant or impactful or to give less weight to the error types that are seen as less significant or impactful than other error types in their contribution to the two quality measures. There is a secondary role that the error type weights are useful for: If a particular error type is more relevant in a particular text type or in a particular target language, then that difference can be corrected by raising the ETW of that error type.

A best practice is to keep the error type weights close to 1 so that they accentuate more than they distort. A recommended range for ETWs is between 0.5 and 2.

Empirical Parameter Contents

As noted above, there is just one empirical parameter, the penalty scalar. Being empirical means that it can be derived mathematically, given a valid set of evaluation data. The default PS value of 1 is offered as a neutral, *a priori* value, and it is subject to change as organizations implement the MQM Scoring Model and begin gathering data points from quality evaluations and the resulting quality measures. It can be anticipated that these quality results will, over time and given enough data, generate validated penalty scores that can be used to set new organizational PS defaults and contribute to industry defaults.

The three following sections show how to derive defensible penalty scalars from real-world quality evaluations. The first of these sections discusses how to find a reference quality score from an alternative, serviceable quality metric against which to compare an MQM-based overall quality score. The second section shows how to establish an equivalence between the OQS introduced in this standard and an independent reference quality score, such as a holistic quality score, and then calculate a target penalty scalar that would align these two quality scores. The third section shows how a series of these derivative target penalty scalars can be averaged to yield a new default PS for organization or industry use.

As stated above, the scoring parameters can be classified along two dimensions (Table 2). One dimension is their role in calculating the quality measures, and the two parameter roles are the determinative and scaling parameters. The other dimension reflects the manner in which their values are set, and the three parameter derivations are the conventional, subjective, and empirical parameters.

Table 2: Scoring Parameter Types Contents

Scoring Parameter	Parameter Role	Parameter Derivation
Severity level penalty	Determinative	Subjective
Error type weight	Determinative	Subjective
Reference word count	Scaling	Conventional
Maximum score value	Scaling	Conventional
Penalty scalar	Scaling	Empirical

Reference Metrics and Quality Scores Contents

There are at least two good reasons to compare an MQM-compliant metric to another TQE metric, whether it is another analytic metric, a holistic metric, or some other quantitative form of evaluation. One reason is to compare the strengths and weaknesses of the metrics to see where each is better suited to a given context and to find ways to improve the metrics in comparison to one another.

The second is to align the quality scores produced by the metrics and, where they do not align, to adjust the scoring parameters in the MQM-compliant metric or parameters that drive the reference TQE metric so that the two metrics will align better. This approach is precisely how the single MQM empirical parameter, the penalty scalar, is fine-tuned. The final two sections describe this process in detail.

Reference Metric Selection Contents

Before aligning an MQM-compliant metric with a good reference metric, it is necessary to find a reference metric that is both valid and reliable. A metric is valid if it does a good job of

measuring what it purports to measure, in this case, the quality of a translation. It is reliable if it produces stable, reproducible, and consistent results. A contrastive analysis of an MQM-compliant metric and a reference metric is a perfect example of one measure of reliability, i.e., inter-method reliability, which measures a work product using different assessment methods or instruments. So being valid and reliable is the first requirement of a good reference metric.

A second requirement is that the two metrics must be used to assess the same object, in this case, the same translation product or set of translation products. They cannot just evaluate similar translation products or different samples of the same translation product but the exact same texts.

Finally, the quality scores that they produce must be comparable, not just in the scales that they use, but also in the interpretation of points along those scales. For example, if two metrics both use 100-point scales, but one fails any translation that receives a score below 60, while the other fails any translation with a score below 80, then they are not comparable.

If the default form of a non-MQM quality score is not directly comparable to the MQM overall quality score but it can be converted mathematically to a comparable score—same range of scores, same top score, and same range between good and bad scores—then the comparison should be made using the MQM-compatible version of the score.

Finally, MQM has two quality measures—the overall normed penalty total and the overall quality score. In theory, either could be aligned with a comparable quality score from another metric, but for the purpose of defining a usable PS, it is assumed that the OQS will be the normal point of comparison for most TQE metrics.

Normative Reference Standard Contents

Even if two quality metrics are nominally aligned, if their quality measures for translations of similar quality do not align, then they are not truly comparable. This problem can be resolved to a large degree using an independent third standard with which both can be aligned, even if aligning with this third standard entails a fairly wide margin of error.

The normative reference standard proposed here derives from grading systems and general quality assessments. Based on their experience with these systems, most people "feel" that a score in the 90s is very good, a score in the 80s is pretty good, and a score in the 70s is average but generally acceptable. If translation quality scores match those expectations, stakeholders feel that the results are valid. If scores are perceived as significantly higher or lower then general expectations, then stakeholders are not likely to trust the quality measures for the translations or the metrics used to generate them.

One grading metric than can be adapted as a basis for calibrating translation quality metrics is the European Credit Transfer and Accumulation System (ECTS). This is a framework developed by the European Commission to establish a normative grading system to unite the disparate grading

systems used by schools in different European and international countries. The unified grades are represented by letter grades and descriptions of the degree of subject matter mastery those grades represent. Those grades are often correlated to percentages of correct answers on one or more exams, yielding numeric grade ranges in a 100-point scale.

The ECTS grade descriptions can be changed from descriptions of student performance or subject mastery to descriptions of translation quality in a straightforward fashion, and these adapted descriptions can be aligned in a table with appropriate ranges of quality scores. If the quality score ranges are interpreted as OQS scores, then it is possible, using the formulas in the previous sections, to calculate the corresponding ONPT values, assuming default values for RWC (1,000) and MSV (100). Combined, these data constitute a solid, if course-grained, definition of what the ONPT and OQS scores—and comparable scores from other metrics—should represent. This comparison (Table 3) provides a yardstick against which all 100-point translation quality scores can be measured.

Table 3: Translation Quality Range Interpretations*

Contents

Letter	ONPT	OQS	Description
A	0 – 100	90 – 100	Excellent, outstanding quality without significant errors
B	101 – 200	80 – 89	Very good, above-average quality but with minor errors
C	201 – 300	70 – 79	Good, generally sound quality but with some errors
D	301 – 400	60 – 69	Satisfactory, fair quality but with significant shortcomings
E	401 – 500	50 – 59	Sufficient, quality barely meets the minimum level
F	> 500	< 50	Failing, improvements required to meet minimum level

* Correlating OQS ranges to ECTS requires that MSV = 100. The ONPT ranges given in the table assume that RWC = 1,000. The PS is a factor in calculating both the ONPT and the OQS, but it is not a factor in correlating their ranges.

Penalty Scalar Derivation

Contents

As noted above, the penalty scalar is the only empirical scoring parameter, which means that its proper value can be calculated, given an appropriate set of data. Data set elements have two components: (1) the scaling parameters and quality measures from an initial MQM quality evaluation and (2) a comparable quality score from a reference metric applied to the same translation product. These two data set components are used to calculate a penalty scalar that can be used with the MQM evaluation to yield an OQS that is equal to the reference quality score.

When enough of these adjusted penalty scalars have been collected, a new default PS can be calculated that can be used consistently to generate OQS scores that correspond generally with the reference quality scores. This exercise can be applied equally well in either direction—to tune an MQM metric to the reference metric or to tune the reference metric to the MQM metric.

These critical calculations all operate in the bottom half of the scoring model. They all begin with the per-word penalty total and apply alternative scoring parameters to calculate alternative quality measures. The key to tuning the penalty scalar is to calculate backward to the PWPT using the initial quality measures and scoring parameters and then calculate forward to arrive at the target OQS, solving along the way for the target PS that will lead to the target OQS. The target OQS, in this case, is set equal to the reference quality score. The formulas that accomplish this are given below.

If the acronyms RWC, PS, and MSV are reserved for the original scaling parameters and the acronyms ONPT and OQS are reserved for the original quality measures, then TOQS can be used for the target OQS, TONPT can be used for the target ONPT, and TPS can be used for the target penalty scalar. PWPT, RWC, and MSV will be the same in calculating both sets of quality measures.

The following two formulas can be used to calculate the target penalty scalar that will make it possible for the two quality scores, OQS and TOQS, to align. The two formulas below highlight the pivotal role of the PWPT in calculating the TPS.

Note that the TPS formula is not well defined if a quality evaluation yields no errors (PWPT = 0, ONPT = 0, OQS = MSV). The TPS formula is designed to generate a particular less-than-perfect quality score for a less-than-perfect evaluation. It is not possible to turn a perfect quality score (OQS = MSV) into an imperfect quality score (TOQS < MSV), or vice versa. The formulas reinforce this point, in that trying to do so would require dividing by 0. For similar reasons, these formulas and others are not valid if RWC = 0, MSV = 0, or PS = 0.

$$PWPT = ONPT / (RWC \times PS)$$

$$TPS = (1 - (TOQS / MSV)) / PWPT$$

The following formula can be used to derive the TPS in a single calculation that subsumes the calculation of the PWPT:

$$TPS = ((1 - (TOQS / MSV)) \times (RWC \times PS)) / ONPT$$

The TPS is derived from the TOQS, but the target ONPT is dependent, in turn, on the TPS. As a result, the resulting TONPT can be expressed as a function of the TOQS:

$$TONPT = (1 - (TOQS / MSV)) \times RWC$$

Penalty Scalar Averaging Contents

The TPS is defined for one translation product, one MQM-compliant quality evaluation, one reference quality evaluation, and one reference quality score. The evaluation word count of the evaluation text can be used as a weighting factor, and pairs of EWCs and TPSs can be used to

calculate a weighted average of TPSs. This procedure comprises the final step in deriving an improved default PS, as described in this section.

Default Penalty Scalar Adjustments

Contents

The preceding section describes how go from a translation product to two quality evaluations—one MQM-based and one using a reliable, independent quality metric—how to align the quality scores produced by the two metrics, and how to use this alignment to calculate backward to a targeted penalty scalar in the MQM metric that would produce the same quality score as the independent TQE metric. This approach yields an empirical, bottom-up penalty scalar rooted in real-world translation and evaluation data. It clearly provides a better approach than using a hunch or a neutral numerical value for the PS.

Initially, a single calculated penalty scalar is used as a preliminary value, and it can be rough or significantly skewed. Enough similarly derived data points are needed to establish a new, statistically valid PS benchmark. Furthermore, just as different sets of active error types or assignments of different determinative parameters may be appropriate for evaluating different text types and translation requirements, different default penalty scalars may be appropriate in different contexts as well. For example, for a gist MT product, a smaller PS may be appropriate to raise the situationally defined OQS to expected quality levels for this translation type.

In a particular organization, it may be appropriate to define different default PSs for different text types used in the organization or for use in different divisions of the organization. Likewise, given the linguistic idiosyncrasies of different languages and the special translation requirements of different target markets, evaluations of translation products into different target languages and locales will almost certainly require different default PSs. For the same reasons, industry associations like TAUS may choose to tailor their PS guidance using similar criteria.

Default Penalty Scalar Calculation

Contents

The formula for deriving a representative PS that can be used as an approximation of a new default PS is a weighted average. Given a set of translation products with paired MQM-based and independent quality scores and the target PSs that are defined using these quality score pairings, the target PSs can be used as the data points being averaged, and the associated evaluation word counts can be used as the weighting factors.

The averages are not based on the MQM overall quality scores from the series of quality evaluations, which will correspond to translation products with a range of quality levels. In principle, the same good PS will turn a poor PWPT into a poor OQS and turn a good PWPT into a good OQS. The weighted averages are based on the validated PSs that do this. The relevant variables and the weighted average formula are defined as follows:

Evaluation index i, with upper bound m = count of quality evaluations included in the weighted average penalty scalar calculation

EWC_i = evaluation word count, for the i^{th} quality evaluation

TPS_i = target penalty score, calculated for the i^{th} quality evaluation

WAPS = weighted average penalty scalar, approximation of an adjusted default penalty scalar

$$WAPS = \left(\sum_{i=1}^{m} (EWC_i \times TPS_i)\right) / \sum_{i=1}^{m} EWC_i$$

This formula can be made more nuanced by adding a secondary weight, SW_i, which quality managers apply to each of the m quality evaluations. The secondary weight provides the means to apply a subjective weight to quality evaluations in order to recognize, for example, that one translation product is more representative than others, that one pair of evaluations is more reliable than others, or that one quality score better reflects extreme values, such as an unusually poor translation. This additional variable complicates the weighted average formula only slightly:

SW_i = secondary weight, for the i^{th} quality evaluation, subjective weight applied to penalty scalars to adjust their impact on the weighted average up or down

$$WAPS = \left(\sum_{i=1}^{m} ((SW_i \times EWC_i) \times TPS_i)\right) / \sum_{i=1}^{m} (SW_i \times EWC_i)$$

The penalty scalar plays a crucial role in defining usable quality scores. The initial default value of 1 is a placeholder. As the MQM Scoring Model is rolled out and as organizations gain experience using it, it can be anticipated that the level of expertise in applying it will grow, and defensible, data-driven values will emerge and form the basis for future default PS values.