Towards Automatic Generation of Evaluation Plans for Context-based MT Evaluation

Andrei POPESCU-BELIS
Paula ESTRELLA
Margaret KING
Nancy UNDERWOOD

ISSCO WORKING PAPER 64

August 2005

ISSCO/TIM/ETI
University of Geneva
40, bd. du Pont-d'Arve
1211 Geneva 4, Switzerland

andrei.popescu-belis@issco.unige.ch

Abstract

This report extends the FEMTI guidelines for context-based MT evaluation with new functionalities aimed at evaluators and experts. The proposed interface to FEMTI generates an outline evaluation plan depending on the characteristics of the context in which an MT system will be used, entered by the evaluators. We first summarize the principle of context-based MT evaluation and the initial FEMTI proposal. Then, we introduce a vector-based representation of the context and of the quality characteristics, which underlies the process of evaluation design. We then show how this process is simplified by the proposed interfaces to FEMTI, and how expertise can be input into the system by using more advanced interfaces. A unified account of expert vs. evaluator use of FEMTI is finally proposed.

The proposals outlined here represent work-in-progress accomplished in the project: "Quality models and resources for the evaluation of machine translation" (SNF division II, grant number 103318, 2004-2006).

Contents

1 Introduction

FEMTI, the Framework for the Evaluation of Machine Translation in ISLE, was proposed by the ISLE Evaluation Work Group as a synthesis of the many aspects of MT quality and their associated metrics (Hovy, King and Popescu-Belis 2002). The main point of the FEMTI guidelines is that MT systems are used in a variety of contexts (e.g. various tasks and users), which do not all require the same qualities from a system. The quality of the output itself is only an aspect of overall system quality (along with speed and updatability for instance), and can be in turn decomposed into more elementary features, such as fluency or terminological correction. The abundance of such quality metrics makes it difficult for MT evaluators to build on previous methods or results. FEMTI combines a classification of the possible contexts of use with a generic quality model for MT (Hovy, King and Popescu-Belis 2002) into a synthesis of many approaches to MT evaluation, intended as a valuable practical resource in the field.

We propose in this paper a set of user-friendly interfaces to the initial FEMTI resource, aimed both at novice and expert MT evaluators. The proposal is based on a vector-space representation of the context of use and of the quality model, which are connected by a *generic contextual quality model (GCQM)*. Considering the need for simplicity of use, especially for evaluators using FEMTI for the first time, we first distinguish the consultation by inexperienced users from the consultation by experts. Evaluators can use the resource to define a context of use for the system to be evaluated, and obtain in return a list of documented evaluation metrics. Experts are needed in order to build the GCQM based on analytic knowledge and previous evaluations. The implementation unifies the two consultation modes, thus also allowing evaluators to tune their quality models based on their own experience.

The remainder of the paper is organized as follows. In Section 2, we review the main arguments for context-based MT evaluation, leading to the FEMTI proposal. In Section 3 we describe the formal underpinnings of the interfaces to FEMTI, in particular the GCQM. Finally, section 4 describes the user's and the expert's views of FEMTI, followed by a unified perspective and some implementation issues.

2 Context-based MT evaluation

2.1 Origins of FEMTI and related work

The FEMTI guidelines are rooted in more general considerations of language technology evaluation (King and Maegaard 1998) put forward in the EAGLES project (EAGLES Evaluation Working Group 1996), in relation to ISO/IEC standards on software evaluation (ISO/IEC 2000, ISO/IEC 2001).

Among the direct forerunners, the work of the Japanese Electronic Industry Development Association (JEIDA) argued that user needs are essential in assessing the quality and usefulness of a end-user MT system (Nomura 1992; Nomura and Isahara 1992a, 1992b). Two sets of 14 parameters each were identified: one that characterizes the desired context of use of an MT system, and the other that characterizes the MT system and its output. A mapping between these two sets of parameters allows one to determine the degree of match, and hence to predict a system's ability to respond to the user's needs.

In line with the general philosophy of DARPA evaluations, the DARPA MT evaluation campaigns from the 1990s (Doyon, Taylor and White 1998, White and O'Connell 1994) adopted a very different stance: they concentrated on the functionality of systems to the exclusion of all other quality characteristics, and even within this constrained area limited their investigation to three aspects of output quality, ignoring, for example, questions of efficiency, robustness or usability—more generally, whether the results are suitable for the needs of a specified user. The impossibility of providing a gold standard required the use of human judges to score fidelity and fluency, and to answer multiple choice questions using translated documents.

In the recent DARPA/NIST campaigns, the idea of a gold standard for MT has been equated to a set of reference translations of the same text. The average distance of a candidate translation from this set is measured using the BLEU metric (Papineni 2002, Papineni *et al.* 2001), or its NIST variant (Doddington 2002), which compare sequences of words (n-grams) occurring in a candidate translation to sequences of words occurring in the set of reference translations. Many other techniques for automatic evaluation have been proposed (Melamed, Green and Turian 2003, Niessen *et al.* 2000), often showing an increased correlation with human judgments of quality in terms of fidelity or fluency, or with some global ranking of candidate translations (Babych and Hartley 2004a, 2004b). However, even if such metrics capture some aspects of output quality, many other aspects of system quality are at least equally important with respect to real user needs.

2.2 Definitions related to MT evaluation

Recent standards in the field of software evaluation (ISO/IEC 2000, ISO/IEC 2001) provide a general definition of quality models and situate these in the software lifecycle. According to ISO/IEC 14598-1 (ISO/IEC 2000: p.12, fig.4), the software life-cycle starts with the analysis of the user requirements or needs that will be answered by the software, which determine a set of software specifications, or, from the point of view of quality, external quality requirements. During the development phase, software quality becomes an internal matter related to the characteristics of the software itself. Once a product is obtained, it becomes possible to assess the internal quality, then the external quality, i.e., the extent to which it satisfies the specified requirements. Finally, turning back to the initial user needs, quality in use is the extent to which the software really helps users to fulfill their tasks.

According to (ISO/IEC 2001), internal and external quality can be decomposed into six quality characteristics: functionality, reliability, usability, efficiency, maintainability, and portability, which can be refined into a hierarchy of sub-characteristics. When particularized for a given software domain and context of use, such a hierarchy is called a quality model. Its terminal nodes are quality attributes, that is, features of the software that can be measured by a metric.

2.3 Aspects of the quality of MT systems

Given that MT systems fall under the scope of the ISO/IEC guidelines for software evaluation, it is natural that the FEMTI guidelines particularize the ISO/IEC ones, following the EAGLES/ISLE approach.

In the realm of quality models for MT software, functionality plays the leading role, especially through two quality attributes generally called *fluency* (the capacity to produce lexically and syntactically well-formed sentences) and *fidelity* (the capacity to preserve the meaning of the source text). System developers and real-world users often add other quality attributes, notably *price*, *system extensibility*, or *coverage*. The aforementioned JEIDA studies, as well as other comparisons of commercial MT systems, make use of a few dozen criteria. As another example, the OVUM report (Mason and Rinsche 1995) includes *usability*, *customizability*, *application to total translation process*, *language coverage*, *terminology building*, and *documentation*. In fact, as discussed by Church and Hovy (1993), for some real-world applications, functionality-related attributes may even take a back seat to these sorts of factors.

Besides, it is not always clear what aspect of quality is really measured by the recent automated metrics quoted above. Most of these metrics correlate somewhat to fidelity and fluency as measured on the DARPA 1994 data (or on more recent samples)—to the extent that adequacy and fluency are themselves correlated. For instance, the weighted N-gram metric proposed by Babych and Hartley (2004a) has two dimensions, namely weighted precision, which approximates human-assessed fluency, and weighted recall, which approximates adequacy (DARPA 1995 data).

It must be noted here that use-oriented evaluations, such as task-based evaluations (White and Taylor 1998) do not make use of the characteristics above, but evaluate an MT system with respect to the scores of a translation-related task. Therefore, such methods do not belong in FEMTI, though their role in properly defining classes of contexts of use is significant. For instance, Tomita (1992) measured how well students answered TOEFL’s comprehension questions using the output of MT systems; the scores to be compared are here the TOEFL, not the FEMTI ones. Similarly, the utility of MT to various types of Japanese learners of English has been assessed by Fuji and Isahara (2001).

2.4 The FEMTI guidelines

The Framework for the Evaluation of Machine Translation in ISLE (Hovy, King and Popescu-Belis 2002) emphasizes the central influence of the context of use of an MT system on the qualities that should be measured in order to evaluate the system. FEMTI is intended to help evaluators to construct a quality model based on the expected context of use of the software. The 2003 version of FEMTI (King, Popescu-Belis and Hovy 2003) is a freely available web-based evaluation resource, which was implemented through a large scale cooperative effort involving a significant part of the MT evaluation community[1]. Originating in Hovy's (1999) hierarchical representation of both context and quality characteristics of MT systems, FEMTI is made of two interrelated classifications or taxonomies, part I and part II.

2.4.1 FEMTI part I: context

The first taxonomy enables evaluators to define the intended context of use of the MT system(s) that must be evaluated, or, in other words, a set of user requirements. The main aspects to be considered here are the type of user of the MT system, the type of task, and the nature of the input to the system. However, the purpose of evaluation and the exact object of the evaluation are also relevant.

2.4.2 FEMTI part II: quality

The second taxonomy lists the MT software quality characteristics as hierarchies of sub-characteristics, with internal and/or external quality attributes at the bottom level. The upper levels match the ISO/IEC 9126 characteristics, while the lower levels are made of MT-specific attributes. For each attribute, definitions and references to the metrics used by the community are also provided.

2.4.3 Context-to-quality relations

The most original aspect of FEMTI is a mapping from the first part to the second part, which states the quality characteristics, sub-characteristics and metrics that are relevant to each feature of the context of use. For instance, 'terminology precision' (an attribute of functionality in part II) is an important quality for MT systems aimed at 'document routing / sorting' (a context of use from part I). When the links for all the context characteristics from part I that apply to a given context are followed, the result is a set of quality attributes from part II, which constitutes a quality model.

Although defined from a theoretical point of view (Hovy, King and Popescu-Belis 2002), these links from part I to part II were not fully worked out in FEMTI 2003. In the present paper, we propose a procedure and a set of tools to perform this operation, along with the more theoretical notion of a general contextual quality model.

[1] The website created by the ISLE Evaluation Work Group acknowledges the contributions to FEMTI: http://www.issco.unige.ch/projects/isle/femti/ (see the "About FEMTI" section).

3 A vector-space representation of context and quality characteristics in FEMTI

In order to provide an operational resource, the relations between context of use and qualities must be made subject to computational methods. We have argued previously (Hovy, King and Popescu-Belis 2002, p.55) that a quality model could be represented by a linear averaging function applied to the scores provided by quality metrics—with null coefficients for metrics that are irrelevant to the given context (*ibid*., p.55-56).

The construction of the linear selection function based on the applicable context characteristics was previously based on an algorithm that summed and normalized the weighted links from part I to part II. Here, we propose a more abstract model, based on vectors and matrixes, offering a clearer representation of expert knowledge and user input. This model is compatible with the previous considerations, and remains of course hidden to the FEMTI end-users, who access the model only through the user-friendly interfaces described in Section 4.

3.1 Context and quality vectors

An ideal view of the part I and part II taxonomies is that they can be represented as vectors, by considering a pre-order traversal of the nodes of each taxonomy.

3.1.1 Part I: the context vector

A context of use or a set of user requirements can be described as a list of features from part I, such as “task = document routing” or “user = not proficient in source language”. In the *context vector*, the positions corresponding to the features that apply bear ‘1’ or ‘true’ and the others ‘0’ or ‘false’, using a Boolean representation. A numeric representation would allow the coding of the importance of each context feature, using for instance a normalized vector.

However, one can see by looking at FEMTI that some context characteristics are not subject to the “applicable / not applicable” choice, for instance “input: doc type: genre” or “user: organization: quantity of translations”. Such characteristics are however easy to render discrete, by specifying for each of them the possible cases. Listing the broad text genres, or a logarithmic ranking of the quantity of translations, allows these context features to be represented into a Boolean context vector.

Parts I.3, I.4 and I.5, which deal respectively with the task, the user, and the type of input of the MT system, are typical sub-hierarchies that are decomposed into features, each pointing towards qualities from part II. Part I.1 deals with the purpose of evaluation, and has a similar, albeit shallower, structure. More difficult is the status of part I.2 dealing with the object of evaluation, which does not contain links to part II, and should probably be moved into a preamble to FEMTI.

3.1.2 Part II: the quality vector

In a similar way, a set of qualities from part II, i.e. a quality model, can be represented as a vector, obtained for instance by the traversal of the hierarchy's leaves. The quality model can be Boolean if it consists of a list of qualities, or numeric, if the list is weighted. In the latter case, the weights are those used in the linear assessment or averaging function that generates the final score, i.e. they encode the contribution of the score obtained by the system for each measured quality attribute to the overall score of the system (if such a score is desired).

Depending on whether the quality vector is Boolean or numeric, there are two ways to derive the averaging function from the quality vector. In the Boolean case, the quality attributes/metrics used for evaluation are those marked '1' or 'true' in the vector, and all receive the same weight (it is probably more informative to provide a score vector as a result rather than the average). In the numeric case, the weights provided by the quality vector are used in the assessment function, or weighted average of the scores.

In what follows, for clarity reasons we will only consider Boolean vectors—for context and quality. That is, the context features are simply 'applicable' or 'not applicable' to a given context of use, and quality attributes are simply 'relevant' or 'irrelevant' within a quality model. A more complex option, which we foresee in a future implementation, is the use of a Boolean context vector and a numeric quality vector, that is, with weighted quality attributes/metrics—or at least an "important vs. not-so-important" distinction.

3.2 Generic contextual quality model (GCQM)

We now describe how a given context of use (context vector) determines the quality model (quality vector) to be used for the evaluation of MT systems aimed at that context. The goal is to embody into FEMTI a procedure that associates to any context vector a quality vector, based on the previous experience of experts in MT evaluation regarding the qualities that are relevant to a given context. In FEMTI 2003, this experience is embodied into the sets of relevant qualities assigned to various features of the context of use (part I).

Our proposal is that the correspondence from FEMTI part I to part II can be computed linearly, using a fixed matrix which we name the *generic contextual quality model (GCQM)*. If this matrix is noted $\boldsymbol{M}$, then the quality vector Q corresponding to the context vector C is simply the matrix product of $\boldsymbol{M}$ and C:

$$Q = \boldsymbol{M} \cdot C$$

To clarify, suppose that C is an m-dimensional vector (part I has a total of m leaves) and that Q is an n-dimensional vector (part II has a total of n leaves). Then, M is a matrix of m columns and n rows, which can also be Boolean or numeric. If both vectors and the matrix are Boolean, then the matrix product must also be computed in a Boolean way (multiplication means 'and' and sum means 'or'); if the matrix is not Boolean, then the product can be computed numerically and the resulting quality vector is a numeric one.

A simple interpretation of the GCQM matrix is that row *i* indicates which quality attributes are relevant to the context characteristic *i*—namely, if coefficient *j* in this row is non-zero, then quality attribute *j* is relevant to the context-characteristic *i*. Therefore, a GCQM is not *per se* a quality model, but a generic correspondence between contexts and qualities.

As has always been the case with FEMTI, it is the knowledge of evaluation experts that must be embodied into a GCQM, i.e. into the links between contexts of use and quality models. The vector based approach provides a simple solution to this requirement, by offering the possibility to each expert to define a GCQM, as a matrix, and then to compute the average of the pool of individual GCQMs into an aggregate matrix. If Boolean GCQMs are considered, they can be aggregated either using a logical 'and' (only qualities selected by all experts are kept), or using a logical 'or' (all qualities selected are kept).

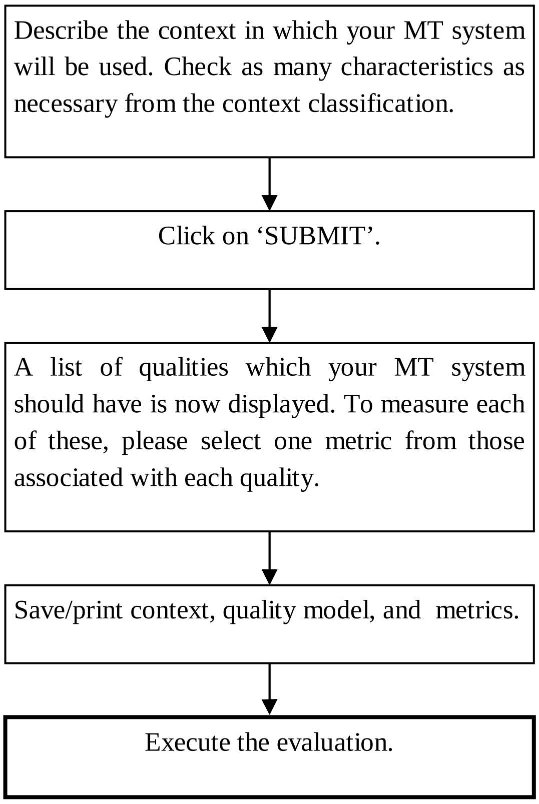
The GCQM is therefore the main data structure that embodies the knowledge of the relation between FEMTI part I and part II. This proposal elaborates on the algorithm proposed previously, which it reformulates in a more elegant and easier to implement fashion. However, the status of taxonomy nodes with respect to leaves in the context and quality vectors is still subject to analysis. According to one view, only the leaves of the taxonomies could appear in the context/quality vectors, while the node could simply be considered to indicate the sets of leaves underneath. This abstraction conflicts somehow with the possibility that part II nodes contain particular metrics, which do not appear in the leaves.

4 Workflow and interfaces for generating quality models

FEMTI users are not supposed to understand the mechanism outlined above in order to start using the taxonomy. Even evaluation experts should not bother about matrices and coefficients: they should only enter relevant qualities for relevant context features. The workflow and interfaces below are designed with these goals in mind, and are finally combined into a unifying perspective for these two types of users.

4.1 The user's view

The FEMTI novice user is an MT evaluator in search of a quality model. For such users, FEMTI provides help in properly defining the context of use, then proposes a quality model depending on this context, using the expertise it incorporates.



Describe the context in which your MT system will be used. Check as many characteristics as necessary from the context classification.

Click on ‘SUBMIT’.

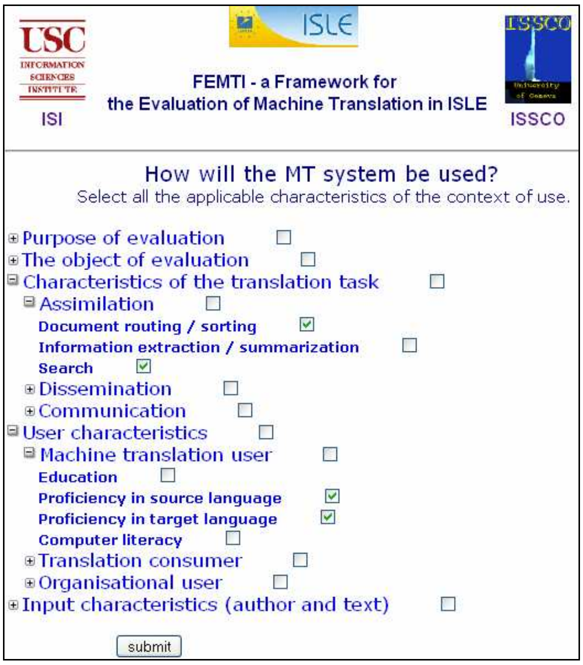
A list of qualities which your MT system should have is now displayed. To measure each of these, please select one metric from those associated with each quality.

Save/print context, quality model, and metrics.

Execute the evaluation.

Figure 1. FEMTI workflow for evaluators

The workflow of the FEMTI user (Figure 1) follows the natural stages in context-based evaluation. The user is presented with a list of characteristics of the context among which she has to select the ones that describe the intended context of use—see Figure 2 for a screenshot of the corresponding interface (note the checkboxes and the +/– signs to hide/display sub-hierarchies).



FEMTI - a Framework for
the Evaluation of Machine Translation in ISLE

How will the MT system be used?
Select all the applicable characteristics of the context of use.

- ⊞ Purpose of evaluation □
- ⊞ The object of evaluation □
- ⊟ Characteristics of the translation task □
  - ⊟ Assimilation □
    - Document routing / sorting ☑
    - Information extraction / summarization □
    - Search ☑
  - ⊞ Dissemination □
  - ⊞ Communication □
- ⊟ User characteristics □
  - ⊟ Machine translation user □
    - Education □
    - Proficiency in source language ☑
    - Proficiency in target language ☑
    - Computer literacy □
  - ⊞ Translation consumer □
  - ⊞ Organisational user □
- ⊞ Input characteristics (author and text) □

submit

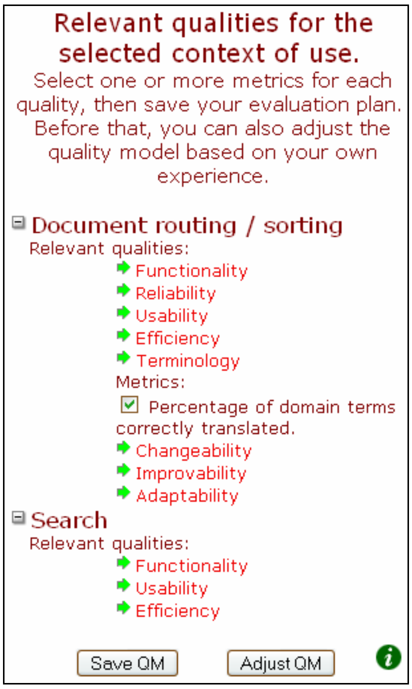
Figure 2. Interface for specifying the context of use of the MT system to be evaluated

Once the context is defined, clicking on "submit" displays the relevant quality attributes (based on internal computation of $Q = \boldsymbol{M} \cdot C$), which are highlighted among all the other qualities (FEMTI part 2) as shown in Figure 3 for a screenshot. At this point, the user has to choose one metric per attribute, if several are proposed, a stage that seems difficult to automate. However, the selection can be done based on the descriptions and comments attached to the metrics—a point on which more work is needed in FEMTI—and depends on the resources available to the evaluator.

Once metrics are selected, the result of the FEMTI consultation can be saved as a PDF file, which includes the context model, the quality model, and the metrics. These constitute the basis for the evaluation plan, automatically generated from the description of the context.

4.2 The expert's view

Experts are required to define links between FEMTI part I and part II, that is, relevant qualities for each characteristic of the context. Although these links are embodied in a GCQM, it is clearly not the experts' task to figure out how the matrix has to be filled. Therefore we propose an interface-based procedure that generates the formal result.



Relevant qualities for the selected context of use.

Select one or more metrics for each quality, then save your evaluation plan. Before that, you can also adjust the quality model based on your own experience.

⊟ Document routing / sorting

Relevant qualities:

- ➡ Functionality
- ➡ Reliability
- ➡ Usability
- ➡ Efficiency
- ➡ Terminology

Metrics:

☑ Percentage of domain terms correctly translated.

- ➡ Changeability
- ➡ Improvability
- ➡ Adaptability

⊟ Search

Relevant qualities:

- ➡ Functionality
- ➡ Usability
- ➡ Efficiency

Figure 3. Retrieval of the quality model (arrows indicate selected qualities, with their metrics)

The expert's workflow, outlined in Figure 4, requires the expert to work on each context characteristic separately: the first stage is thus the choice of a characteristic from FEMTI part I, which is done using an interface similar to the one in Figure 2 (except that only one box can be checked). Then, the full part II taxonomy is displayed, allowing the expert to select the relevant qualities, as in the interface in Figure 5, without the bottom-right pop-up window in this case.

The expert can then save these links, and proceed to another context characteristic, immediately or in a later session. The identity of the expert is preserved via the personal GCQM file that stores all the links previously entered. When starting a new session, the expert only needs to reload her GCQM file into the interface.

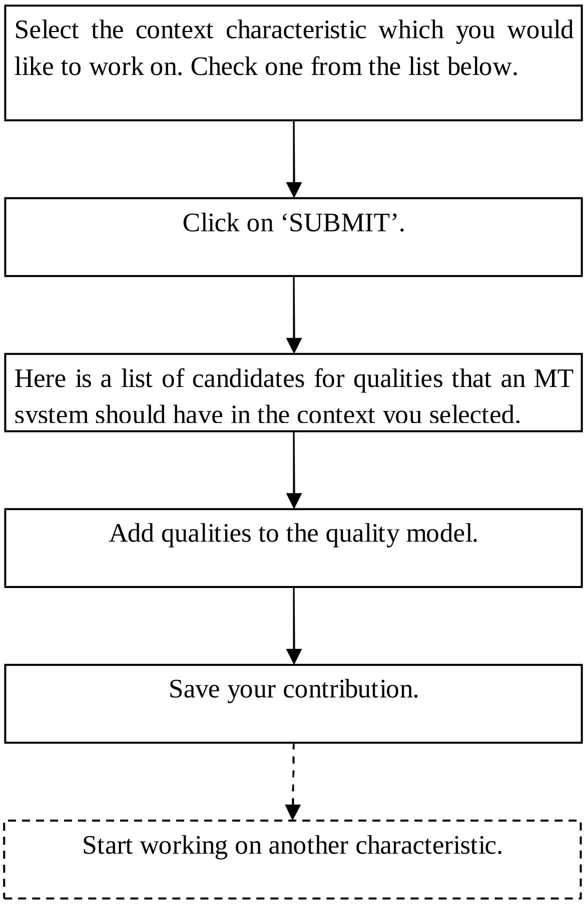
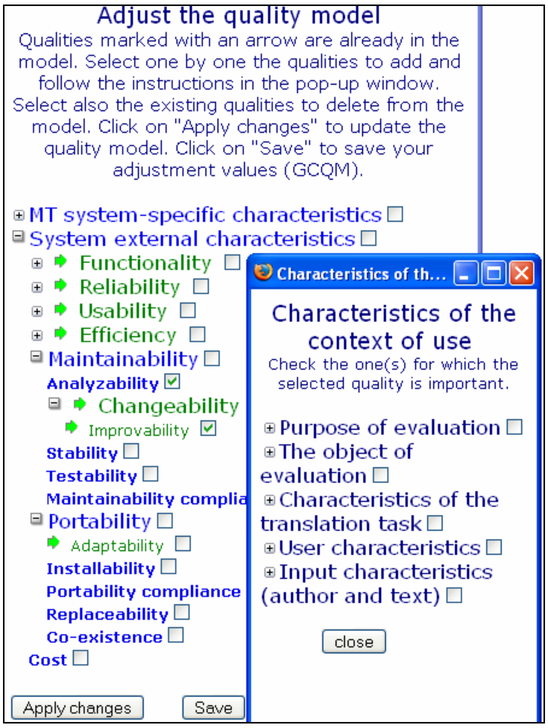

Select the context characteristic which you would like to work on. Check one from the list below.
Click on 'SUBMIT'.
Here is a list of candidates for qualities that an MT system should have in the context you selected.
Add qualities to the quality model.
Save your contribution.
Start working on another characteristic.

Figure 4. Expert's workflow

4.3 Integrating user and expert interfaces

It appears from the two previous sections that the user (MT evaluator) and the expert share some of the new FEMTI interfaces. The main difference is that experts are asked to define a quality model (working on context characteristics one by one), while users are presented with a quality model that they cannot change—though they can choose their metrics, or choose not to measure certain attributes, or even go back offline to the full FEMTI part II to include new qualities. However, we believe that the possibility for an informed user to *tune or adjust a quality model* should also be taken into account, which makes the distinction between users and experts less obvious.

We propose therefore to allow both user and experts to adjust quality models, following the common workflow shown in Figure 6. Of course, being able to adjust and to save a quality model does not necessarily mean that the result will be validated as expertise and stored into FEMTI (in technical terms, the GCQM saved by a user is not necessarily stored in the global repository).



Adjust the quality model

Qualities marked with an arrow are already in the model. Select one by one the qualities to add and follow the instructions in the pop-up window. Select also the existing qualities to delete from the model. Click on "Apply changes" to update the quality model. Click on "Save" to save your adjustment values (GCQM).

- ⊞ MT system-specific characteristics □
- ⊟ System external characteristics □
  - ⊞ ➧ Functionality □
  - ⊞ ➧ Reliability □
  - ⊞ ➧ Usability □
  - ⊞ ➧ Efficiency □
  - ⊟ Maintainability □
    - Analyzability ☑
    - ⊟ ➧ Changeability
      - ➧ Improvability ☑
    - Stability □
    - Testability □
    - Maintainability complia
  - ⊟ Portability □
    - ➧ Adaptability □
    - Installability □
    - Portability compliance
    - Replaceability □
    - Co-existence □
  - Cost □

Characteristics of th...

Characteristics of the context of use

Check the one(s) for which the selected quality is important.

- ⊞ Purpose of evaluation □
- ⊞ The object of evaluation □
- ⊞ Characteristics of the translation task □
- ⊞ User characteristics □
- ⊞ Input characteristics (author and text) □

Figure 5. Adjusting the quality model

The first step in using FEMTI (Figure 6) becomes then the selection of the GCQM to be used: non-expert users select the current FEMTI one (the average of the individual GCQMs entered by experts), while an expert can start with a blank GCQM or continue working on a previously saved one. Users/experts are then prompted to select contextual features from part I: as many as needed for the user, and in principle only one at the time for the expert. When proceeding to part II, users are presented with the quality model corresponding to their context, while experts are shown a blank quality model to which they must add qualities.

If they find it useful, users can also add/remove qualities from the model they received, as shown in Figure 5. However, in this case a pop-up window will prompt them to specify the context characteristic to which the added quality is relevant. The window does not appear if only one characteristic was selected prior to the adjustment, as is the case with an expert dealing with one context characteristic at the time.

If the quality model is modified, users/experts can save it, and, if allowed, add it to the FEMTI pool of GCQMs. Users can also save/print the result: context, quality model, and metrics.

This unified view enables the input into FEMTI of expertise from previous MT evaluations. Such evaluations must first be analyzed in terms of context characteristics and qualities.

Then, the context characteristics can be entered, and a default quality model (using the current GCQM) can be computed. Based on the comparison with the qualities/metrics used in the past evaluation, the quality model could be tuned, and the result can be saved into a new GCQM, which can be added to the FEMTI pool.

4.4 Implementation

The new implementation of FEMTI described above is currently under work. The screenshots provided in Figures 2, 3, and 5 offer insights on the final result. The main goal in the implementation is FEMTI's usability, with respect to both users and experts: interfaces must as intuitive as possible.

One of the important changes since FEMTI 2003 is the use of a dynamic document server named Cocoon (a piece of software that is freely available at http://xml.apache.org/cocoon), which generates web pages on-the-fly from XML content. The combination of stylesheets and Javascript allows the generation of expandable hierarchies with checkboxes, and the behind-the-scenes processing of the forms using matrix products. The management of expert identities and rights, based on individual GCQM data structures, is currently under study, as is the averaging of GCQMs into a unique FEMTI matrix.

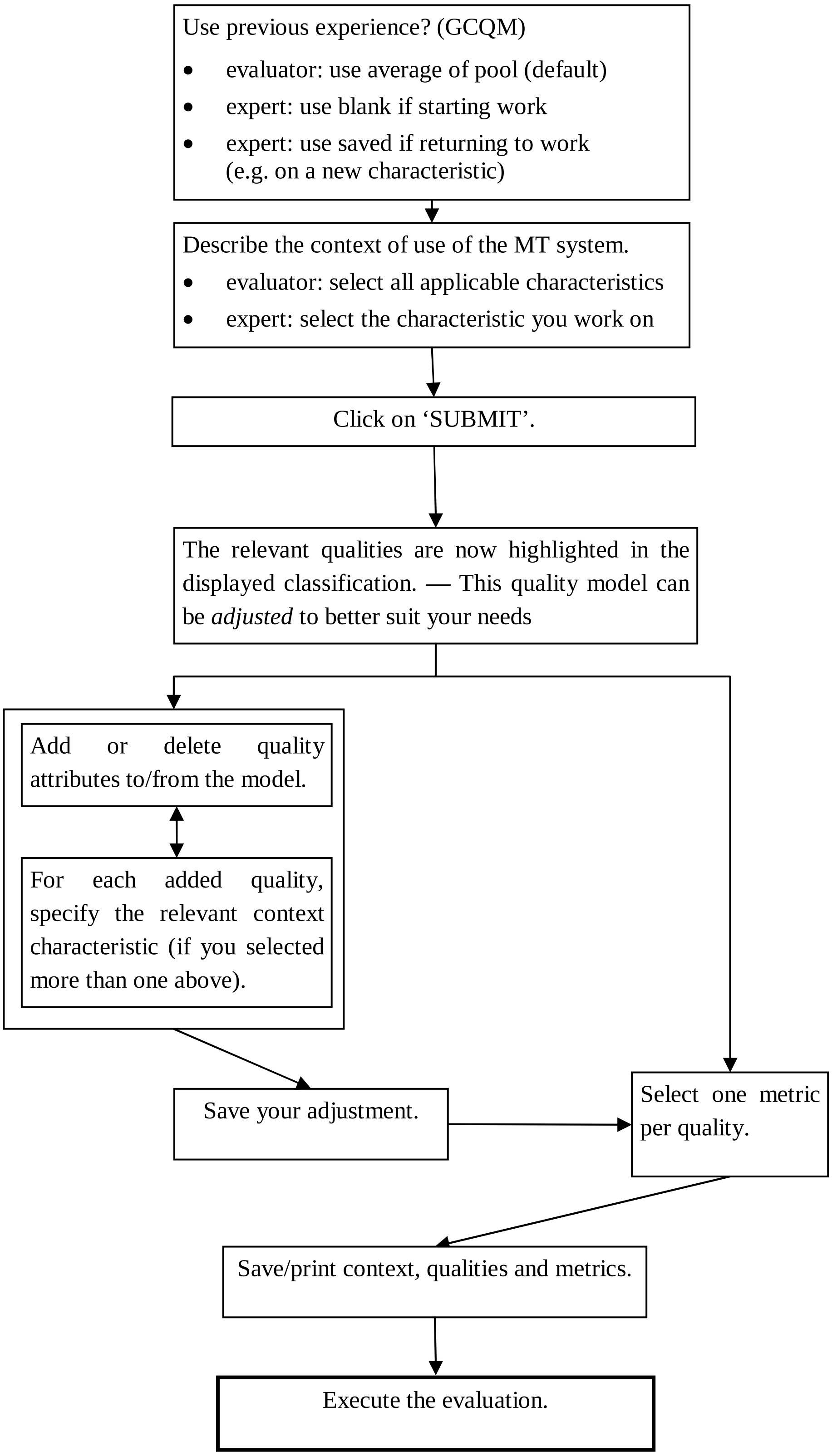

Use previous experience? (GCQM)
• evaluator: use average of pool (default)
• expert: use blank if starting work
• expert: use saved if returning to work (e.g. on a new characteristic)
Describe the context of use of the MT system.
• evaluator: select all applicable characteristics
• expert: select the characteristic you work on
Click on 'SUBMIT'.
The relevant qualities are now highlighted in the displayed classification. — This quality model can be *adjusted* to better suit your needs
Add or delete quality attributes to/from the model.
For each added quality, specify the relevant context characteristic (if you selected more than one above).
Save your adjustment.
Select one metric per quality.
Save/print context, qualities and metrics.
Execute the evaluation.

Figure 6. Integrated workflow for context-based MT evaluation: users and experts

5 Perspectives

Apart from completing the implementation of the new FEMTI 2005, and making provisional decisions on the points discussed above (especially related to the usability of the resource), three main perspective goals appear at this point.

First, the GCQM must be instantiated with information from experts and from previous evaluations, i.e. (weighted) connections should be defined between context characteristics and quality models. A trial round is planned once the implementation is functional, but a joint expert session, as organized within the ISLE project in the past, would provide more reliable results. Information derived from previous evaluations must also be entered gradually. Second, the two FEMTI taxonomies still require maintenance and update, especially regarding recently proposed quality metrics. A more thorough analysis of metrics in terms of correlations and cost is also needed, but represents a long-term goal. Finally, we plan to propose a consensual, pre-normative document on the standardization of MT evaluation, and on the standardization of context-based HLT evaluation in general.

References

Babych B. and Hartley T. (2004a): "Extending the BLEU MT Evaluation Method with Frequency Weightings", *ACL '04*, Barcelona, pp. 621-628.

Babych B. and Hartley T. (2004b): "Modelling Legitimate Translation Variation for Automatic Evaluation of MT Quality", *LREC 2004*, Lisbon, vol. III/VI, pp. 833-836.

Church K. W. and Hovy E. H. (1993): "Good Applications for Crummy MT", *Machine Translation*, vol. 8, pp. 239-258.

Doddington G. (2002): "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics", *HLT 2002*, San Diego, CA, pp. 128-132.

Doyon J., Taylor K. B. and White J. S. (1998): "The DARPA MT Evaluation Methodology: Past and Present", *AMTA '98 Conference*, Philadelphia, PA.

EAGLES Evaluation Working Group (1996): EAGLES Evaluation of Natural Language Processing Systems, *Final Report*, Center for Sprogteknologi, EAG-EWG-PR.2.

Fuji M. and Isahara H. (2001): "Evaluation Method for Determining Groups of Users Who Find MT «Useful»", *MT Summit VIII*, Santiago de Compostela, pp. 103-108.

Hovy E. H. (1999): "Toward Finely Differentiated Evaluation Metrics for Machine Translation", *EAGLES Workshop on Standards and Evaluation*, Pisa.

Hovy E. H., King M. and Popescu-Belis A. (2002): "Principles of Context-Based Machine Translation Evaluation", *Machine Translation*, vol. 17, n. 1, pp. 45-78.

ISO/IEC (2000): *ISO/IEC 14598-1: Information Technology – Software Product Evaluation – Part 1: General Overview*, Geneva, International Organization for Standardization.

ISO/IEC (2001): *ISO/IEC 9126-1: Software Engineering – Product Quality – Part 1: Quality Model*, Geneva, International Organization for Standardization.

King M. and Maegaard B. (1998): "Issues in Natural Language Systems Evaluation", *LREC'98*, Granada, vol. 1/2, pp. 225-230.

King M., Popescu-Belis A. and Hovy E. (2003): "FEMTI: creating and using a framework for MT evaluation", *MT Summit IX*, New Orleans, LA, pp. 224-231.

Mason J. and Rinsche A. (1995): *Translation Technology Products*, Report OVUM Ltd.

Melamed I. D., Green R. and Turian J. P. (2003): "Precision and Recall of Machine Translation", *HLT-NAACL 2003*, Edmonton, AB, pp. 61-63.

Niessen S., Och F. J., Leusch G. and Ney H. (2000): "An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research", *LREC 2000*, Athens, pp. 39-45.

Nomura H. (1992): *JEIDA Methodology and Criteria on Machine Translation Evaluation*, Japan Electronic Industry Development Association (JEIDA).

Nomura H. and Isahara H. (1992a): "The JEIDA Report on Machine Translation", *AMTA Workshop on MT Evaluation: Basis for Future Directions*, San Diego, CA.

Nomura H. and Isahara H. (1992b): "JEIDA's Criteria on Machine Translation Evaluation", *IPSJ SIGNotes Natural Language*, Tokyo, Information Processing Society of Japan, pp. 107-114.

Papineni K. (2002): "Machine Translation Evaluation: N-grams to the Rescue", *LREC 2002*, Las Palmas, Spain.

Papineni K., Roukos S., Ward T. and Zhu W.-J. (2001): *BLEU: a Method for Automatic Evaluation of Machine Translation*, Research Report, IBM T.J. Watson Research Center, RC22176 (W0109-022).

Tomita M. (1992): "Application of the TOEFL Test to the Evaluation of Japanese-English MT", *AMTA Workshop on MT Evaluation: Basis for Future Directions*, San Diego, CA.

White J. S. and O'Connell T. A. (1994): "The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches", *AMTA Conference*, Columbia, MD.

White J. S. and Taylor K. B. (1998): "A Task-Oriented Evaluation Metric for Machine Translation", *LREC'98*, Granada, vol. 1/2, pp. 21-25.