The FEMTI guidelines for contextual MT evaluation: principles and resources

Paula Estrella[1,*], Andrei Popescu-Belis[2], Maghi King[3]
[1] FaMAF, National University of Córdoba, Argentina
[2] Idiap Research Institute, Martigny, Switzerland
[3] ISSCO/TIM/ETI, University of Geneva, Switzerland

Summary. A large number of evaluation metrics exist for machine translation (MT) systems, but depending on the intended context of use of such a system, not all metrics are equally relevant. Based on the ISO/IEC 9126 and 14598 standards for software evaluation, the Framework for the Evaluation of Machine Translation in ISLE (FEMTI) provides guidelines for the selection of quality characteristics to be evaluated depending on the expected task, users, and input characteristics of an MT system. This approach to contextual evaluation was implemented as a web-based application which helps its users design evaluation plans. In addition, FEMTI offers experts in evaluation the possibility to enter and share their knowledge using a dedicated web-based tool, which have been tested in several evaluation exercises.

1. Introduction

A variety of approaches have been proposed for the evaluation of machine translation systems, and numerous metrics have been proposed as well. Researchers typically focus on output quality, which is generally the most important aspect of research-oriented systems. Output quality can be measured using human-based as well as automatic metrics designed to capture the quality of machine translation (MT). A system can also be assessed indirectly through its operational use, in a task-based evaluation approach. In either approach, MT systems can be compared against each other during an evaluation campaign. However, end-users of MT tend to include other factors in an evaluation, not only related to output quality. The methodology that takes into account the intended context of use of a system when designing its evaluation has become known as *context-based evaluation*. This paper describes the application of this approach to the evaluation of MT systems, which has resulted in the Framework for the Evaluation of Machine Translation in ISLE (International Standards for Language Engineering), abbreviated FEMTI. This framework aims at standardizing the MT evaluation process and provides support tools that help users define contextual evaluation plans. The goal of FEMTI is to organize the different characteristics of an MT system into a coherent taxonomy and to help evaluators select the right subset of characteristics to

be assessed given the specific purpose of the evaluation and the factors related to the environment where the system will be deployed.

This paper is structured as follows: Section 2 gives an overview of the context-based evaluation paradigm; Section 3 introduces the quality model used by FEMTI, a notion inspired from ISO/IEC standards; Section 4 presents the different components that constitute the FEMTI framework, while Section 5 presents the activities that were carried out to disseminate the framework and collect feedback from experts. Finally, Section 6 presents conclusions and possible extensions of FEMTI.

2. Methods for the evaluation of MT systems

To measure the quality of an MT system by evaluating its output, automatic metrics, task-based ones, and the subjective rating of certain aspects of translation quality have all been used. Some practitioners have also taken into account the intended context of use of an MT system, in what is called *context-based evaluation*. One of the first initiatives considering other factors than simply MT output quality was a report by the Japan Electronic Industries Development Association (JEIDA), which advocated a framework for the evaluation of MT systems from a user's and developer's point of view (Nomura, 1992). Two sets of criteria were proposed: evaluators (users or developers) are required to answer one questionnaire about their present work situation and another one about their specific needs. After that, radar charts are created with the results of both questionnaires and finally, the evaluator chooses the type of system that appears to be the most suitable based on the overlap of the two radar charts.

The Evaluation Working Group of the EAGLES EU project (Expert Advisory Group on Language Engineering Standards) also adopted a user-oriented point of view on the evaluation of human language technology products. The general framework for evaluation proposed by this group was partly inspired by the ISO/IEC 9126 standard on the evaluation of software (ISO/IEC, 1991) which was used to relate potentially important attributes of a product to a class of users. The framework also covered the implied needs of users in what was called the *consumer report paradigm* (EAGLES Evaluation Working Group, 1996), where users identify the class of users that better represents their needs (among a predefined set of user classes) and select the characteristics of the product believed to be relevant for that class of users. Subsequent projects using the EAGLES framework have contributed to its validation and to test its usefulness for evaluation design (Canelli, Grasso, & King, 2000; Rocca, Spampinato, Zarri, & Black, 1994; TEMAA, 1996).

(Hovy, 1999) proposed an intermediate solution between the JEIDA and EAGLES methodologies, consisting of a hierarchy or taxonomy of both user needs and quality characteristics of systems, originally called *user purpose* and *user process*, dealing with the reason for translation and the translation method, respectively. Each level of the hierarchy had a set of associated metrics and was decomposed into finer detail. Although this solution was formally very close to that of EAGLES or JEIDA, Hovy's work was more flexible, as it allowed the evaluator to decide the level of detail and other features to include in the evaluation – as opposed to the other solutions that had a fixed predefined set of features for user types and systems.

The continuation of EAGLES into the (ISLE) EU project focused on the evaluation of MT systems and on how to relate user needs to system quality characteristics. The ISLE Evaluation Working Group applied the ISO/IEC 9126 and 14598 standards to MT software and extended existing methodologies, building up the FEMTI framework (Hovy, King, & Popescu-Belis, 2002). After the ISLE project, work on FEMTI continued with the goal of converting these guidelines into a more interactive tool that would guide the evaluator through the generation of customized evaluation plans (Estrella, Popescu-Belis, & Underwood, 2005). The FEMTI framework is now a web-based application publicly available at http://www.issco.unige.ch/femti and will be presented in detail in Section 4.

3. ISO/IEC standards applied to context-based evaluation

The FEMTI framework took as a starting point the ISO/IEC 9126 (ISO/IEC, 2001) and ISO/IEC 14598 (ISO/IEC, 1999) standards, which are domain independent guidelines for the evaluation of software products and are, therefore, intended to be applicable to all kinds of software.

Quality characteristic	Quality sub-characteristics
Functionality	Suitability, Accuracy, Interoperability, Security, Functionality compliance
Reliability	Maturity, Fault tolerance, Recoverability, Reliability compliance
Usability	Understandability, Learnability, Operability, Attractiveness, Usability compliance
Efficiency	Time behavior, Resource utilization, Efficiency compliance
Maintainability	Analysability, Changeability, Stability, Testability, Maintainability compliance
Portability	Adaptability, Installability, Co-existence, Replaceability, Portability compliance

Figure 1. Generic quality model proposed by ISO/IEC 9126.

The 14598 series provides guidelines and examples to support different stakeholders during the evaluation process, while the 9126 series defines the components of a generic quality model. These series complement each other, since the specification of a quality model is part of the evaluation process and this process could be different, depending on the stakeholders involved (evaluators, developers, acquirers, etc).

In the ISO/IEC 9126 view, quality is defined as "the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs" (ISO/IEC, 2003a). The ISO/IEC quality model aims at representing the different aspects of a product that together will make its overall quality, resulting from the six top-level quality characteristics: *functionality*, *reliability*, *usability*, *efficiency*, *maintainability*, *portability*. These quality characteristics are decomposed as shown in Figure 1, and the attributes of the quality model (i.e. the terminal nodes in such a hierarchy) are measurable features of the software product. In all cases, metrics are required to measure these attributes and, therefore, a set of metrics should be associated to each attribute of a quality model. The ISO/IEC 9126 series offers specific parts devoted to external metrics (ISO/IEC, 2003a) and internal metrics (ISO/IEC, 2003b).

2.1 Functionality
- 2.1.1 Accuracy
  - 2.1.1.1 Terminology
  - 2.1.1.2 Fidelity – precision
  - 2.1.1.3 Consistency
- 2.1.2 Suitability
  - 2.1.2.1 Target-language suitability
    - 2.1.2.1.1 Readability
      - ⇒ Metric 1: Cloze Tests
      - ⇒ Metric 2: Subjective rating of intelligibility
      - ⇒ Metric 3: Reading time
    - 2.1.2.1.2 Comprehensibility
    - 2.1.2.1.3 Coherence
    - 2.1.2.1.4 Cohesion
  - 2.1.2.2 Cross-language - Contrastive suitability
  - 2.1.2.3 Translation process models
  - 2.1.2.4 Linguistic resources and utilities
- 2.1.3 Well-formedness
  - 2.1.3.1 Morphology
  - 2.1.3.2 Punctuation errors
  - 2.1.3.3 Lexis - Lexical choice
  - 2.1.3.4 Grammar – Syntax
    - ⇒ Metric 1: Percentage of phenomena correctly treated
    - ⇒ Metric 2: List of error types

Figure 2. Partial decomposition of the *Functionality* quality characteristic in the FEMTI quality model for MT software. Metrics are exemplified for two quality attributes, *Readability* and *Grammar/syntax*.

If the generic model proposed in ISO/IEC 9126 is to be applied to software in a particular domain, it needs to be specialized through the definition of attributes and metrics which fit that particular domain. In FEMTI, the ISO/IEC generic quality model was tailored to the MT domain, maintaining its top-level structure and extending it with an additional top-level quality characteristic, namely *Cost*, and with sub-characteristics specific to MT systems. An example of instantiating the model for the MT domain is shown in Figure 2, which illustrates the resulting decomposition of *Functionality*. From the figure, it appears that some characteristics were added at the same level as the ISO/IEC ones (e.g. *Well-formedness*), while others were further decomposed (e.g. *Suitability* → *Target-language suitability* → *Readability*). This figure also shows the place of metrics in the quality model, for example under *2.1.2.1.1 Readability*, and *2.1.3.4 Grammar – syntax*. Numbering of the taxons was added to facilitate cross-referencing in all subsequent work using FEMTI. Besides offering a broader view of a system's overall quality, this ISO-inspired quality model for MT systems allows evaluators to integrate many other aspects of quality beyond the generic characteristic *output quality*, usually assessed with the popular *adequacy* and *fluency* metrics.

4. Making the FEMTI guidelines operational

The first version of the FEMTI framework was developed until 2003 with support from the ISLE EU project. This version focused on the integration of the existing quality and context characteristics for MT into classifications that organize them hierarchically. The main limitation of the initial interface that was designed to access FEMTI's content was that it demanded a significant effort from the users who wanted to build an entire evaluation plan using it: they had to manually construct the plan by keeping track of their selection (context and quality characteristics plus metrics) while navigating back-and-forth the hierarchies. Another limitation was that its web pages had to be re-generated each time a change was made to the contents of FEMTI, due to its implementation as a set of separate, static web pages. Therefore, the goal of the new version of FEMTI was to increase its usability by creating a set of complementary tools that help users browse the framework when creating quality models and to reduce the maintenance needed by using a dynamic document server for the implementation.

This section outlines the support tools developed as part of FEMTI; Section 4.1 describes the tool for evaluators, then Section 4.2 describes the mechanism in FEMTI that implements the context-based approach to evaluation and Section 4.3 describes the mechanism that allows knowledge from the MT community, to be entered into FEMTI.

4.1. Generating customized evaluation plans

The target audience of FEMTI is the evaluators (end-users, developers, acquirers, etc.) who want to specify an evaluation plan for one or more MT systems intended to be used in a particular environment. This can be achieved using the *evaluators' interface* of FEMTI, which contains the following parts:

- A classification of possible contexts of use (*Part I*): a hierarchy of features describing the intended environment of use for the MT system.
- A classification of quality characteristics (*Part II*): a hierarchy of desirable system characteristics, whose top level nodes match the generic quality model proposed by the ISO/IEC 9126-1 standard, and a set of metrics associated to most quality characteristics.
- A *context-to-quality relation*: an automatic mechanism that retrieves the relevant quality characteristics according to the specified context of use.

Figure 3 shows the workflow that evaluators must follow in order to generate a quality model using FEMTI. Evaluators start by defining the intended environment of use of the MT system by selecting characteristics related to the *translation task* to be performed by the system, the *author* and *text* characteristics and the *type of user* of the system (as well as a preliminary reflection on the *purpose* of the evaluation). When this is done, evaluators work with Part II, where they select the quality characteristics and metrics of interest, starting with a blueprint that is automatically suggested by FEMTI based on the selected environment of use.

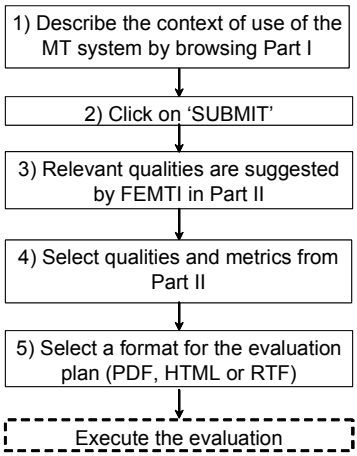

Figure 3. Workflow for the evaluators' interface of FEMTI.

Quality characteristics can be aspects directly related to translation quality (such as adequacy, readability, style, etc) or related to the desired features of the MT system (such as file formats handled, portability to different operating systems, user-friendliness of the interfaces, etc). Consequently, the metrics used to measure the selected quality characteristics include human-based or automatic metrics for translation quality, such as, for adequacy, the rating of sentences on a 5-point scale by humans (White & O'Connell, 1994), or the BLEU metric for fluency (Papineni, Roukos, Ward, & Zhu, 2001). Checklists could be used to measure other features, for instance to make a list of the operating systems, languages and formats supported.

The result of using FEMTI is a document containing the context and quality characteristics chosen by the user plus the metrics. The set of items contained in this report is thus called a *customized quality model.* Users indicate to the FEMTI interface the actual format in which the document can be saved, currently HTML, RTF or PDF.

The execution of the evaluation requires further steps that are outside the scope of FEMTI and focus on the practical details of the evaluation, for example to prepare the necessary test material, to state acceptance levels for each metric, to interpret the results the result of applying the metrics and so on. Therefore, the report generated with FEMTI serves as a basis during the preparation and execution of an evaluation, for example, to choose a the test set representative of the text domain and genre specified with characteristics from Part I or to gather relevant toolkits to apply the metrics selected in Part II.

4.1.1 Using the evaluators' interface

The following screen captures illustrate the use of the evaluators' interface. Figure 4 shows the initial state of the tool, where Part I is displayed on the left frame of the screen and Part II is displayed on the right frame. The labels for each characteristic in Part I and II are hyperlinked to the relevant content, which is displayed in a separate window when clicked on.

In the first example displayed here, suppose that an evaluator has to buy an MT system in order to monitor a large volume of texts produced outside the evaluator's organization. Initially, the evaluator defines a context of use by selecting a type of evaluation, in this case *Operational evaluation* (node 1.1.4) is suitable as he wants to address the question of whether the MT system he will buy will actually serve its purpose; he further specifies the context selecting the type of task the system is supposed to perform (*Assimilation* (node 1.2.1)) and the type of users of the system (*Machine translation user* (node 1.4.1)). These steps of the workflow are illustrated in Figure 5.

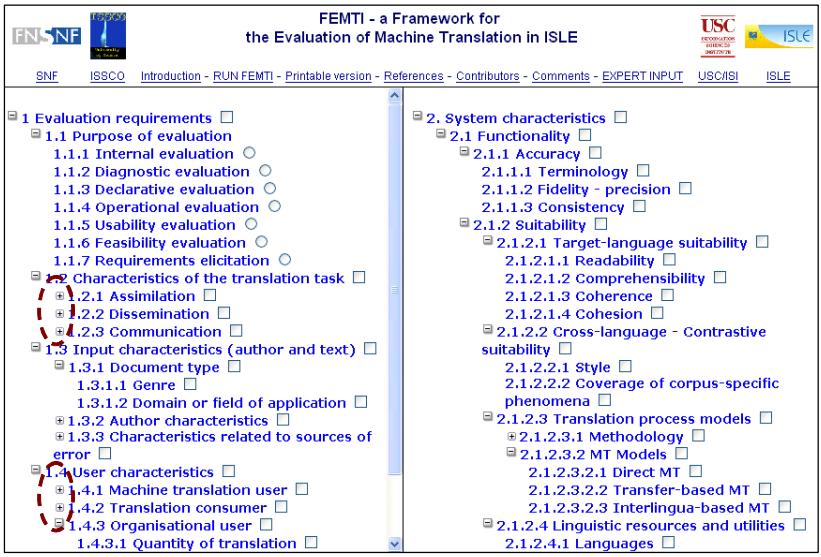

Figure 4. Home page for the evaluators' interface; classifications can be expanded or collapsed using the +/– buttons exemplified with dashed circles.

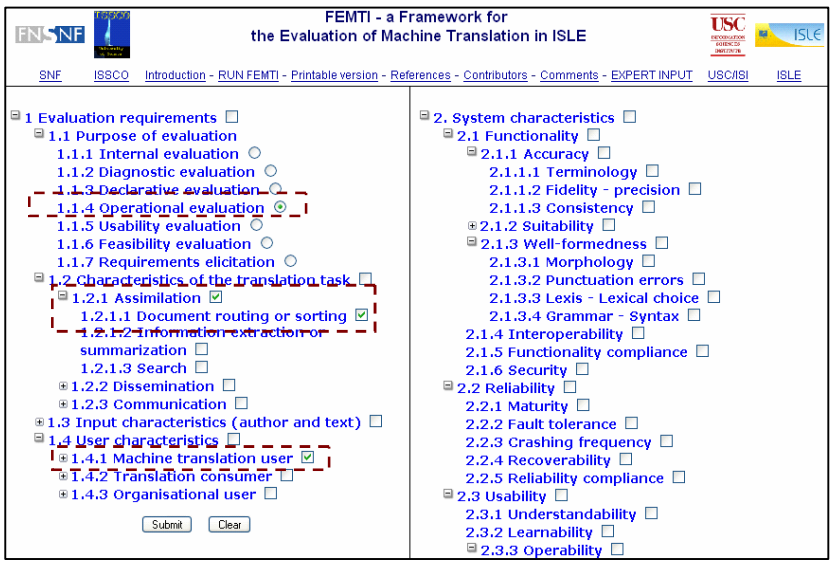

Figure 5. Part I: sample definition of the context of use for an MT system intended to monitor a large volume of texts.

The linking mechanism that implements the context-to-quality relation is activated when the evaluator confirms his selection from Part I by pressing the 'Submit' button at the bottom of the left frame. The result of its operation (fully transparent to the evaluator) is shown in Figure 6: the

quality characteristics relevant to the context defined previously are highlighted in Part II, so that the evaluator selects one or more quality characteristics and metrics.

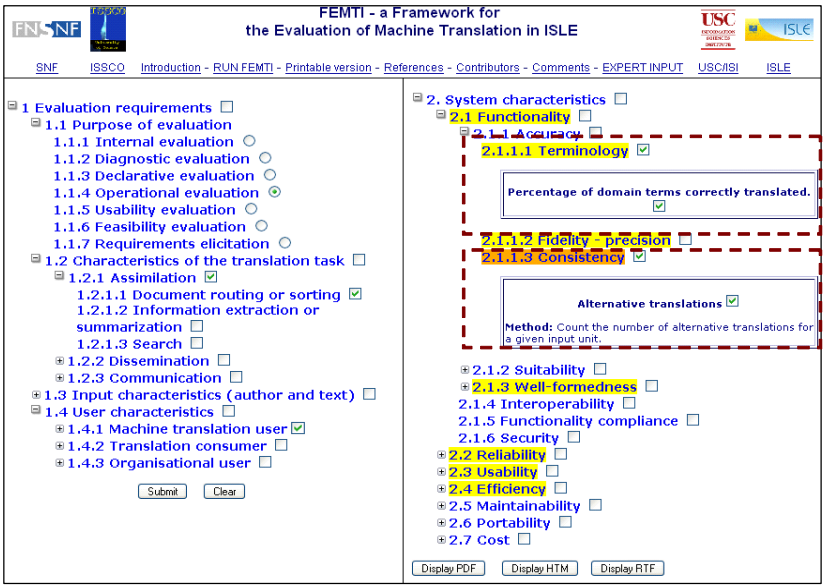

Figure 6. Part II: sample selection of quality characteristics and metrics for an MT system intended to monitor a large volume of texts.

Evaluation type

- Operational evaluation:
Operational evaluations generally address the question of whether an MT system will actually serve its purpose in the context of its operational use. The primary factors include the cost-benefit of bringing the system into the overall process (costs).

Context Characteristics

- Machine translation user:
This refers to the person who interacts with the machine translation system and with the output produced by it.
- Operational evaluation:
Operational evaluations generally address the question of whether an MT system will actually serve its purpose in the context of its operational use. The primary factors include the cost-benefit of bringing the system into the overall process (costs).
- Assimilation:
The ultimate purpose of the assimilation task (of which translation forms a part) is to monitor a (relatively) large volume of texts produced by people outside the organization, in (usually) several languages.
- Characteristics of the translation task:
Characteristics of the translation task refers to the information flow intended for the output, from the point of view of the agent (human or otherwise) who receives the translation.

Additional Quality Characteristics (not suggested by FEMTI)

- Consistency:
Capability of the system to produce from a given input, and at a given point in time, the same output
Metrics:
- *Alternative translations*
- Method: Count the number of alternative translations for a given input unit.

- Terminology:
Correct translation of technical (domain-specific) terms.

Figure 7. Excerpt of evaluation plan generated with FEMTI.

Corresponding to step 4 of the workflow, Figure 6 shows the state of the interface when the evaluator selects one quality characteristic proposed by the linking mechanism (node 2.1.1.1 *Terminology*) and one additional characteristic (node 2.1.1.3 *Consistency*), along with the metric available under each characteristic. Regardless of the automatic result of relating a particular context to a set of quality characteristics, evaluators are free to add or remove any other quality (sub-)characteristics and metrics.

When the selection of the quality characteristics and metrics is complete, the evaluator saves the plan by clicking on 'Display', as illustrated in Figure 7. This document displays the selected context characteristics first, followed by the quality characteristics separated in two sections: a section for the characteristics *suggested by FEMTI* (i.e. resulting from the operations performed by the linking mechanism), which are ranked according to their importance assigned by the linking mechanism, and a section for characteristics *not suggested by FEMTI*, ordered by their index number in Part II.

In this example, the evaluator could have selected other quality characteristics related to the portability of the system (e.g. node 2.6.2 *Installability*), to the efficiency of the system if there is a large volume of texts to translate (e.g. node 2.4.1.3 *Input to Output Translation Speed*) and related to the cost (node 2.7) given that he is supposed to buy an MT system. However, in a different context, some of these aspects might be less important.

Suppose now that the same person must evaluate an MT system that is already available in his organization, and is used daily to translate manuals of a product to be sent to potential customers. In this case, the context of use could be minimally described with the following items from Part I: *Usability evaluation* (node 1.1.5), *External dissemination* (node 1.2.2.2), *Advanced proficiency in source language* (node 1.3.2.1.3 about the author's characteristics) and *Computer literate* (node 1.4.1.4 about the person interacting with the MT system). Given that the chosen task demands high quality translations, many of the characteristics from Part II that are chosen by the evaluator will be related to this aspect of the system, for example *Fidelity* (node 2.1.1.2), *Consistency* (node 2.1.1.3), *Readability* (node 2.1.2.1.1) and *Punctuation errors* (node 2.1.3.2). Other quality characteristics could be related to general features of the system, for example to the language pairs handled (*Languages*, node 2.1.2.4.1) and the *Reliability* of the system (node 2.2), which should have a high tolerance to faults so that it is online most of the time.

It can be noted from these examples that the quality models generated in each case are quite different even if they are created by the same evaluator and for the same organization. To summarize the examples discussed in this

section, Table 1 compares the contexts of use and quality models corresponding to the two previous examples. As these examples suggest, the most original aspect of FEMTI's new version is the linking mechanism from Part I to Part II storing knowledge about MT evaluation, which is used to formalize the context-to-quality relation and which is explained in more detail in the next section.

Context of use	Example 1	Example 2
Evaluation type	Operational	Usability
Translation task	Assimilation	Dissemination
MT user	Computer literate	Computer literate
Author's linguistic proficiency	Advanced in SL	Advanced in SL and TL
Quality model	Example 1	Example 2
Quality characteristics	Consistency Terminology Installability Translation speed Cost	Fidelity Consistency Readability Punctuation errors Reliability Languages

Table 1. Sample quality models created with FEMTI for two different contexts of use.

4.2. Relating context to quality characteristics

In order to convert FEMTI into a context-based evaluation tool, it is necessary to account for the influence of the context of use on the desired features of the system. Once this relation is identified, it is possible to link each context characteristic to a set of quality characteristics indicating the importance of the connection as weighted links. In FEMTI this relation is now implemented through a core structure called a *Generic Contextual Quality Model* (GCQM), which embodies the knowledge necessary to create customized quality models.

In the GCQM an item in Part I is related to a given item in Part II only if the weight connecting them is not null; in this case, the weight indicates the strength of this connection. The weights on the links to the same quality characteristic are added during the operation of the linking mechanism (step 3 of the workflow shown in Figure 3), so that the higher the number of context characteristics related to one quality characteristic, the higher that quality characteristic's final weight in the resulting quality model. Intuitively, this means that quality characteristics with higher weights are more important with respect to other characteristics in the model. This result of the linking mechanism is used when the quality model is generated (step 5 of the workflow shown in Figure 3) and serves to rank the quality characteristics by decreasing order of importance: the most important ones

according to this mechanism appear first. These weights are included in the resulting quality model, in case evaluators are willing to use them, for example, to compute final scores.

Assuming a quality model for a given domain is a hierarchy of characteristics, sub-characteristics and attributes, as in the case of ISO-based models, it can be flattened (e.g. by traversing it depth-first or breadth-first) to be transformed into a list of items (or equivalently into vectors), which are needed to interact with the GCQM. Once a hierarchy is flattened, its *vector representation* is straightforward: each node becomes a component of the vector. Thus, FEMTI's linking mechanism is general enough to be ported to any other domain where a taxonomy of contexts of use and a taxonomy of quality characteristics exist: the hierarchies are flattened as vectors and the corresponding GCQM is a table, where the rows represent context features and columns represent quality features.

The procedure proposed here to suggest to evaluators a list of relevant quality characteristics starts by converting Part I into a *context vector*, where non-zero components indicate the context characteristics selected by the evaluator. Then, the matrix product of this vector with the GCQM is computed, 'filtering' thus only the relevant quality characteristics, and resulting in a customized *quality vector*, i.e. a set of quality characteristics. This procedure to create quality vectors captures the contribution of every component of the context vector to each component of the quality vector. Therefore, the higher the number of non-zero terms in the computation of a quality vector's component the higher its importance in the specific quality model. Conversely, the higher the number of zero terms, the lower the importance of the component.

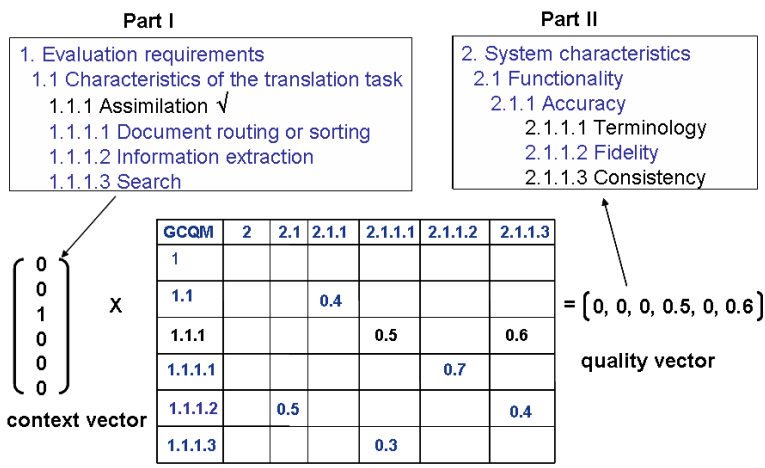

Figure 8. Illustration of the algorithm to obtain a customized quality model (represented as a vector) from a context vector.

The use of the GCQM by the linking mechanism is illustrated in Figure 8. Parts I and II were simplified to consist only of the characteristics depicted in the figure and the relation between them is represented with the weights in the GCQM of the same figure. In this example the user has selected *Assimilation* as the translation task, and the result of filtering the GCQM with that particular context vector is a quality vector with two non-zero components corresponding to *Terminology* and *Consistency*. In practice, when using the evaluators' interface, this would result in *Terminology* and *Consistency* being highlighted in Part II and included in the final evaluation plan if the user selects them.

4.3. Input of expertise into FEMTI's GCQM

A major challenge of the model proposed here to relate context and quality characteristics is to fill in the values of the GCQM. FEMTI's GCQM was initially filled in with the information that was already present in the previous version – more specifically in the section on *Relevant qualities from Part II* in some of the descriptions of context characteristics – but many links are still missing. Additionally, to validate the links created, the GCQM should be populated by several experts. This implies that experts willing to create links for FEMTI, would have to work on a GCQM whose size is currently around 100 by 100, which is particularly unpractical. Therefore, to collect feedback from the MT community, a support tool called the *experts' interface* was developed as part of the FEMTI framework, aiming at simplifying this task.

The goal of the experts' interface is to help experts create and populate as many individual GCQMs as needed, which could be merged to create one 'averaged GCQM' representing the consensus of experts about the relation between Parts I and II of FEMTI. Such an averaged GCQM can be used by the linking mechanism, thus contributing to improve the evaluators' interface as well, by increasing the number of relevant quality characteristics that are suggested automatically.

To construct a GCQM for a given domain, in this case MT, experts proceed as shown in Figure 9. Once logged in, experts select one context characteristic from which the links to quality characteristics will be created (step 1) and make this selection effective by pressing a 'Select' button (step 2). Then experts browse Part II to find the quality characteristics that, according to their experience and knowledge of the domain, are relevant to the selected context characteristic (step 3). The links are created by selecting one or more quality characteristics with a weight and saving them to one's own GCQM (step 4). After one cycle of work, experts can log out (step 5) or continue working on a different context characteristic (step 6).

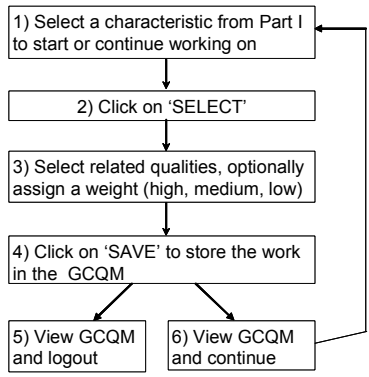

Figure 9. Workflow for the experts' interface of FEMTI.

The use of this tool will be explained using the first example discussed in Section 4.1.1. In order for an evaluator selecting *Assimilation* to get suggestions for the quality characteristics *Terminology* and *Consistency* as above, an expert must have created those relations first. In that case, the expert proceeds as follows: after accessing the framework, he selects the context characteristic *Assimilation* to work on, as shown in Figure 10.

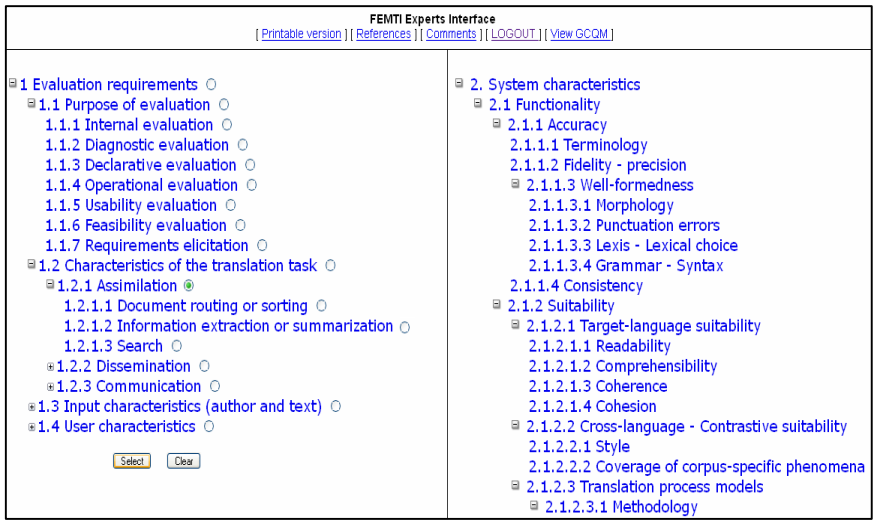

Figure 10. Example of using the experts' interface, where an expert will create links from the context characteristic *Assimilation*.

At this point Part II is expanded with a set of labels that indicate the possible weights for the links to be created, coded for the time being as *high*, *medium*, *low* and *n/a*, the latter indicating that the link exists but the weight is unspecified (numbers are avoided as they would make this task overly complex). Figure 11 shows that the expert has selected two quality characteristics that will be important to the translation task *Assimilation*; in

this case, the expert chooses to assign different weights to these characteristics, namely *medium* for *Terminology* and *low* for C*onsistency.* Figure 12 shows the result of the expert saving the work and viewing the resulting GCQM.

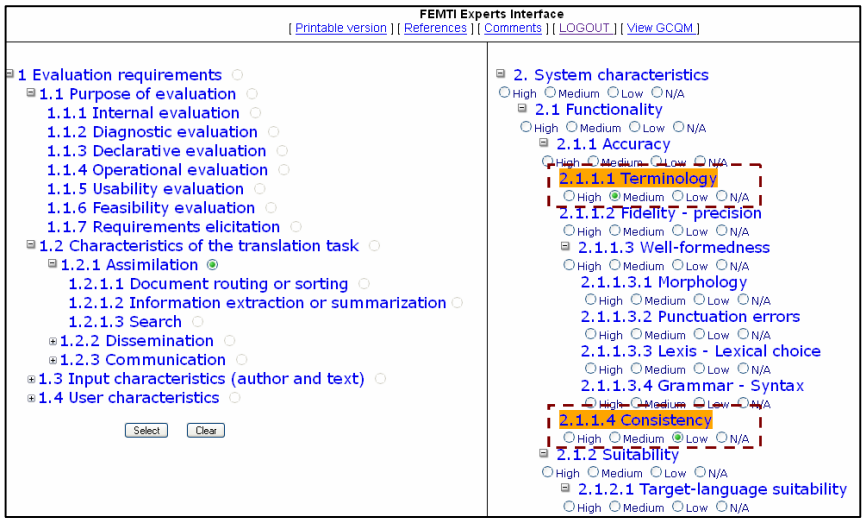

Figure 11. Example of expert selecting quality characteristics *Terminology* and *Consistency* to be linked to *Assimilation.*

FEMTI GCQM	**Assimilation**	Search	Dissemination	Internal dissemination
System characteristics	Low			
Functionality	Low			
Accuracy			High	
Terminology	Medium			
Fidelity - precision			High	
Well-formedness				
Morphology				
Punctuation errors			Low	
Lexis - Lexical choice				
Grammar - Syntax				
Consistency	Low		N/A	
Suitability				
Target-language suitability				
Readability			Medium	
Comprehensibility				

Figure 12. Excerpt of an expert's GCQM showing the relations created from *Assimilation* to *Terminology* and *Consistency*.

As already mentioned, the primary goal of this support tool is to collect knowledge from experts, which will be integrated into the evaluators' interface of FEMTI to improve the suggestion of quality characteristics. A possible way to achieve this is to merge several GCQMs by averaging or accumulating the weights for the same links of the different GCQMs. However, this will be of practical interest only once the experts' interface has been used extensively and enough valid and rich GCQMs are available.

5. Refinement and assessment of FEMTI

This section describes and discusses two activities carried out to collect feedback from the MT community and bring input to FEMTI's GCQM. Two tutorials were set up in 2007 and 2008, using the new version of FEMTI, in order to introduce the framework to potential users and to explain how it can be applied. In addition, the goal was also to encourage the use of the evaluators' interface and to transfer knowledge from the MT community into FEMTI. Following the EAGLES and ISLE series of workshops, these tutorials have been organized in conjunction with major international conferences: the MT Summit in 2007 and the Language Resources and Evaluation Conference in 2008.

The structure of the tutorials was similar in both cases: after introducing the tools, a practical session led participants to specify a quality model for a given scenario of MT use; the quality models were then summarized and discussed during the final slot. Most of the participants used a printed compilation of FEMTI's content while a few accessed the online version. The scenarios proposed to participants were defined as a compromise between specificity and generality: participants needed a reasonably clear scenario to be able to describe it in terms of the context characteristics in FEMTI, but it had to be general enough to avoid biasing the participants too directly towards any specific characteristic. For the exercise participants were arranged in groups of about four persons and were asked to perform the following tasks:

- Identify the context characteristics from FEMTI Part I that would best characterize the given scenario of MT use.
- Indicate the quality characteristics from FEMTI Part II that are believed relevant to each of the selected context characteristic.
- If possible, indicate the importance of each quality characteristic for each context characteristic on a 3-point scale.

For the first tutorial, the proposed scenario featured an MT system that would help select articles from the Chinese press about the preparation for the Beijing 2008 Olympic Games, before handing the articles for proper translation into English by humans. All the four groups of participants

agreed on the top-level context characteristic that defined the translation task (namely, *Assimilation*), but when further specifying it in terms of sub-characteristics, the groups chose different sub-tasks: *Search* vs. *Information extraction* vs. *Document routing*. Other context characteristics that were considered as describing the scenario were: the *Domain or field of application* of the input text, the author's *Superior proficiency in source language*, and the user's *Novice proficiency in source language* and their S*uperior or Distinguished proficiency in target language*. Similarly, a common set of quality characteristics appeared to be important for the given scenario: *Fidelity*, *Terminology*, *Dictionaries*, *Input to output translation speed* and *Cost* – exact answers varied from group to group. From this hands-on exercise, around 40 new links between characteristics from Part I and Part II were created and then added to FEMTI's GCQM by the organizers. Most of them concerned context characteristics that were recently added and had no connections yet to Part II, such as nodes under *Author's proficiency in source* or *target language*.

At the second tutorial, a scenario inspired from a real world use case was proposed. The scenario featured an MT system used for the Global Public Health Intelligence Network (GPHIN) – a web-based early warning system, permanently monitoring several sources of information, in several languages, for disease outbreaks and other public health events, and disseminating the information selected as relevant nearly in real time (Blench, 2007). With the authorization of the GPHIN's contributors, the requirements for the MT system used in their network were presented in detail to the participants, including information about the workflow, type of users, type of texts handled and the evaluation of the overall system and of each MT component.

In the second tutorial the answers of the groups were more detailed than for the first one and showed more overlap across groups, most likely due to the more detailed specification of the scenario. Several relations between Part I and Part II were shared among several groups, thus validating both the description of the scenario and the links themselves; the shared links are:

- *Information extraction/summarization* → *Fidelity; Comprehensibility*
- *Domain or field of application* → *Terminology; Word lists or glossaries*
- *Number of personnel* → *Cost*
- *Time allowed for translation* → *Overall production time; Input to output translation speed*
- *Quantity of translation* → *Input to output translation speed*
- *Multi-client external dissemination* → *Readability*

The particularities of the given scenario are reflected in some context characteristics chosen by several groups, namely *Characteristics related to*

the sources of error, *Document type*, *Genre, Domain or field of application* and *Communication*. In this second tutorial, 115 distinct links were produced, from which 87 were new to FEMTI and were added to the GCQM; the rest of the links will be first validated and then integrated to FEMTI in the near future.

In addition to dissemination of FEMTI and knowledge collection, these tutorials served as an additional validation of the framework, given that participants helped the developers identify areas of Parts I and II to be improved. For instance, the context characteristics regarding *Genre* and *Domain or field of application* are important aspects of the environment of use and should be further decomposed into sub-characteristics to increase their specificity and make them a selectable item in Part I. Similarly, some quality characteristics, such as *Cost* or those related to *Resource utilization,* should be augmented with relevant metrics.

Furthermore, the feedback obtained indicates that, in the current state, FEMTI still requires prior knowledge or experience about MT evaluation in order to be effectively used. As FEMTI users would benefit from more guidance, it is planned to integrate the results of these tutorials into templates or use cases for FEMTI that will be available to the general public. Similarly to EAGLES, increasingly extensive use of the FEMTI framework will help to asses and to validate it, both by experts and evaluators. Ongoing work includes using FEMTI to design the evaluation of speech-to-speech translation systems and one of the expected results of this work is a new list of possible updates to FEMTI.

6. Conclusions and future work

This paper argued that the methodologies taking into account the context of use of a system, for example the JEIDA criteria or the EAGLES consumer report paradigm or FEMTI, are very useful in practice to design informative evaluations that help users get a clear picture of a system's qualities with respect to its intended use. However, context-based evaluation might also seem limited to specific cases, thus reducing the evaluation's reusability, and it also demands more effort from an evaluator to design and execute a contextual evaluation plan. This paper presented an interactive version of the FEMTI guidelines, whose primary goal is to overcome some of the drawbacks of context-based evaluation, especially by offering a set of user-friendly web-based tools to help evaluators generate their plans and to help experts contribute to the field with their knowledge by creating relations between contexts and quality characteristics. In addition to these new functionalities, the current FEMTI provides a simple way to browse through the content, which is an important aspect given the large amount of information available. The most innovative component of FEMTI is the

implementation of an automatic linking mechanism, which uses a GCQM to suggest relevant quality characteristics given a particular context of use. These improvements greatly simplify the evaluators' task when designing an evaluation. FEMTI is thus the first context-based evaluation tool available for MT, and its principles and software infrastructure can be extended to other domains. Combined to particularized ISO/IEC 9126 quality models, the FEMTI tool can contribute to the standardization of evaluation in other domains, as illustrated by (Miller, 2008).

Given that new metrics for MT evaluation appear very often, the contributors and developers of FEMTI are well aware that their work might never be completed. Therefore, future work should keep focusing on FEMTI's content and on providing more practical details about how to design an evaluation with FEMTI. As part of this work, it would be useful to attach an additional section with practical guidelines about the resources that might be needed to execute an evaluation plan, as well as with additional information about the use of automatic and human-based MT metrics for non-experts in the field.

Although the first steps were done to disseminate the framework, to obtain feedback from the MT community and to identify directions for improvement, a more thorough assessment of FEMTI should be performed. For example, this could be done by organizing workshops or expert meetings where the interfaces would be used intensively or, alternatively, these actions could be performed remotely if the organization of such meetings is not logistically possible. Moreover, during such meetings, participants could work on any context characteristic instead of being constrained to a given scenario or they could provide their own context of use, for which a quality model could be created.

Several extensions of FEMTI should also be explored. The current version does not allow evaluators to set the weights in the context or quality vectors, given that the interface only allows them to select or unselect characteristics. In the future, this constraint could be suppressed to let evaluators enter the importance of each selected context characteristic, using a nominal or ordinal scale that provides the weights for both context and quality vectors. Another way of allowing evaluators to tune the weights in their quality models could be to let them load into the evaluators' interface their own GCQM previously created with the experts' interface or to merge the two interfaces into a more sophisticated one, where there is no radical difference between evaluators and experts.

Acknowledgments

The authors would like to acknowledge the steady support of the Swiss National Science Foundation (SNSF), through grants n. 200021-103318 and

200020-113604 for the first author, and through the IM2 National Center of Competence in Research for the second author.

Bibliography

Blench, Michael. (2007). *Global Public Health Intelligence Network (GPHIN).* Paper presented at the MT Summit XI, Copenhagen, Denmark.

Canelli, Maria, Grasso, Daniele , & King, Maghi (2000). *Methods and Metrics for the Evaluation of Dictation Systems: A Case Study.* Paper presented at the Proceedings of the 2nd LREC, Athens Greece.

EAGLES Evaluation Working Group. (1996). *EAGLES Evaluation of Natural Language Processing Systems* (Final Report No. EAG-EWG-PR.2 (ISBN 87-90708-00-8)). Copenhagen, Danmark: Center for Sprogteknologi.

Estrella, Paula, Popescu-Belis, Andrei, & Underwood, Nancy. (2005, 24-25 November 2005). *Finding the System that Suits you Best: Towards the Normalization of MT Evaluation.* Paper presented at the 27th ASLIB International Conference on Translating and the Computer, London, UK.

Hovy, Eduard H. (1999). *Toward Finely Differentiated Evaluation Metrics for Machine Translation.* Paper presented at the EAGLES Workshop on Standards and Evaluation, Pisa, Italy.

Hovy, Eduard H., King, Margaret, & Popescu-Belis, Andrei. (2002). Principles of Context-Based Machine Translation Evaluation. *Machine Translation, 17*(1), 1-33.

ISO/IEC. (1991). *ISO/IEC 9126: Information Technology -- Software Product Evaluation / Quality Characteristics and Guidelines for Their Use*. Geneva: International Organization for Standardization / International Electrotechnical Commission.

ISO/IEC. (1999). *ISO/IEC 14598-1:1999 (E) -- Information Technology -- Software Product Evaluation -- Part 1: General Overview*. Geneva: International Organization for Standardization / International Electrotechnical Commission.

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

ISO/IEC. (2003a). *ISO/IEC TR 9126-2:2003 (E) -- Software Engineering -- Product Quality -- Part 2:External Metrics*. Geneva: International Organization for Standardization / International Electrotechnical Commission.

ISO/IEC. (2003b). *ISO/IEC TR 9126-3:2003 (E) -- Software Engineering -- Product Quality -- Part 3:Internal Metrics*. Geneva: International Organization for Standardization / International Electrotechnical Commission.

Miller, Keith. (2008). *FEIRI: Extending ISLE's FEMTI for the Evaluation of a Specialized Application in Information Retrieval.* Paper presented at the ELRA Workshop on Evaluation "Looking into the Future of Evaluation" at LREC, Marrakech, Morroco.

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

Papineni, Kishore, Roukos, Salim, Ward, Todd, & Zhu, Wei-Jing. (2001). *BLEU: a Method for Automatic Evaluation of Machine Translation* (Research Report, Computer Science No. RC22176 (W0109-022)). Yorktown Heights, NY: IBM Research Division, T.J.Watson Research Center.

Rocca, G, Spampinato, L, Zarri, Gian Piero, & Black, William. (1994). *COBALT: Construction, Augmentation and Use of Knowledge bases from Natural Language Documents.* Paper presented at the Proceedings of the Artificial Intelligence Conference.

TEMAA. (1996). *TEMAA Final Report* (No. LRE-62-070 (March 1996)): Center fo Sprogteknologi, Copenhagen, Danemark.

White, John S., & O'Connell, Theresa A. (1994). *The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches.* Paper presented at the AMTA Conference, 5-8 October 1994, Columbia, MD, USA.

[*] Work performed while at ISSCO, University of Geneva.