Finding the System that Suits you Best: Towards the Normalization of MT Evaluation

Paula ESTRELLA, Andrei POPESCU-BELIS and Nancy UNDERWOOD

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

{paula.estrella, andrei.popescu-belis, nancy.underwood}@issco.unige.ch

Abstract

The *Framework for the Evaluation of Machine Translation,* FEMTI, brings together the many disparate metrics and methods which have been devised for MT and helps evaluators to design an evaluation plan based on the context of use intended for the system. FEMTI allows therefore the generation of more standardized and reusable evaluation plans. By evaluators we mean not only developers and programmers, but also end users, managers, and anyone else with a stake in the acquisition or deployment of a system. Thus, the use of FEMTI is not limited to experts in the field of MT.

In this paper we describe FEMTI and the latest enhancements we are making to it, in particular the interfaces which not only allow evaluators to create their own tailor-made evaluation plans, but also to contribute their experience and expertise in constantly improving the resource for the community at large.

1 Introduction

Evaluating machine translation (MT) systems is not only for researchers and developers to test the performance or improvements of the system they are implementing, but also for (potential) users who need to find out which, if any, system will best satisfy their needs. Over the years, very many different techniques and metrics for evaluating have been devised and applied – see for instance the proposals and references in (Hartley and Popescu-Belis 2004, Hovy, King and Popescu-Belis 2002a, Van Slype 1979).

The techniques which have been used for the evaluation of MT systems vary widely, according to the purpose of carrying out the evaluation, the context in which the MT system is expected to be used, the resources available to carry out the evaluation and the language

Proceedings of the 27th International Conference on Translating and the Computer (ASLIB),
London, UK, November 24-25, 2005, p.23-34.

pairs involved. Since it is not always immediately obvious that existing evaluation metrics might be re-used, evaluators have tended to devise new ways of evaluating from scratch.

Another factor accounting for the wide range of evaluation techniques is the intrinsic difficulty of choosing appropriate metrics for MT evaluation. For example since there is rarely, if ever, a single "correct" translation, it is not possible to create a unique "gold standard" translation for a given text, to which the output of the MT system could be automatically compared. This is partly the reason why so many metrics targeting "translation quality" have been proposed – either to be applied by human judges, such as fidelity and fluency (White and O'Connell 1994), or derived automatically through a statistical comparison with one or more human translations (Babych and Hartley 2004, Doddington 2002, Papineni *et al.* 2001).

The importance of the intended context of use of an MT system for its evaluation is not always taken into enough consideration in evaluation design. Some popular evaluation campaigns[1] do not consider the type of user of the system, or other requirements related to speed of translation or integration into existent software solutions. For example, a system with acceptable output quality that runs only under Windows may not be suitable for an organisation where the only operating system used is UNIX.

The implementation of the *Framework for the Evaluation of Machine Translation* (FEMTI) was inspired by two ISO standards: one on software, ISO/IEC 9126 (1991) and the other on the procedures for software evaluation in general, ISO/IEC 14598-1 (1999). These standards were initially used in the EAGLES project to derive evaluation guidelines for language processing software in general (EAGLES Evaluation Working Group 1996). FEMTI is the result of the evaluation working group of the ISLE (International Standards for Language Engineering) project.

2 Context-based Evaluation

As mentioned above, our work is primarily based on two standards. The ISO/IEC 14598 series of six standards defines a process for the evaluation of software by different types of evaluators, such as developers, acquirers and evaluators. It also provides general guidelines to plan, manage and support the evaluation process including, for example, templates for evaluation reports. The ISO/IEC 9126 series is focused on software quality, and defines a general purpose quality model divided into *internal quality, external quality* and *quality in use;* in addition, it dedicates an appendix to recommendations and requirements for software products metrics.

According to ISO/IEC 14598-1 (ISO/IEC 1999: p.12, fig.4), the generic software life-cycle starts with the analysis of the user needs that will be answered by the software, which determine a set of software specifications. From the point of view of quality, these are the external quality requirements. During the design and development phase, software quality

[1] For example the campaigns carried out by the US National Institute for Standards and Technology: http://nist.gov/speech/tests/mt/.

becomes an internal matter related to the characteristics of the software itself. Once a product is obtained, it becomes possible to assess its internal quality, then the external quality, i.e., the extent to which it satisfies the specified requirements. Finally, turning back to the user needs that were at its origins, quality in use is defined as the extent to which the software really helps users to fulfil their tasks.

The ISO/IEC 9126-1 standard states that software quality results in general from six quality characteristics: *functionality, reliability, usability, efficiency, maintainability, portability*. These characteristics have been refined in the more recent version of the standard, providing a loose 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 always measurable features of the software, that is, attributes. A measurement is "the use of a metric to assign a value (i.e., a measure, be it a number or a category) from a scale to an attribute of an entity" (ISO/IEC 1999: 4.19, p.3).

The ISO/IEC 9126 standard is, as we have emphasized above, domain independent and intended to be applicable to all kinds of software. If it is to be applied to software in a particular application domain, it needs to be specialised through the definition of attributes and metrics which fit that particular domain; such is the case of FEMTI, which organizes metrics proposed for MT evaluation in a hierarchy of quality characteristics and metrics used to evaluate an MT system, which is related to the purpose and intended context of use of the system (Hovy, King and Popescu-Belis 2002b).

3 Components of FEMTI

Given that the ISO/IEC 9126 standard concerns the evaluation of software in general, it is necessary to tailor the quality model in order to apply it to a particular type of application. This involves the definition of attributes and metrics which fit that particular domain. FEMTI is thus designed for the evaluation of MT software. FEMTI is intended to be used by various evaluators and organisations, public or private, and in diverse situations: to compare systems before deploying one in the workflow of an enterprise or to ensure that the chosen system will fit the needs of a class of users. By evaluators we mean not only developers and programmers but end users, managers, and anyone else with a stake in the acquisition or deployment of a system, thus the use of FEMTI is not limited to experts in the field of MT.

FEMTI is made up of three distinct elements: (1) a classification of possible "user requirements" or features of the intended context of use; (2) a classification of the possible quality characteristics that can be evaluated for a MT system, with associated metrics; and (3) a set of links between the first and the second classification. We now describe in more detail these three elements, providing examples at each level.

3.1 Part I: User Requirements

Part I of FEMTI is a classification of the purpose of the evaluation, the object of the evaluation and the main features defining the context of use. This enables evaluators to

define why they need to carry out the evaluation, what is to be evaluated as well as an intended context of use, namely the type of user of the MT system, the type of task, and the nature of the input to the system. Some of the top level characteristics of Part I are listed in Figure 1.

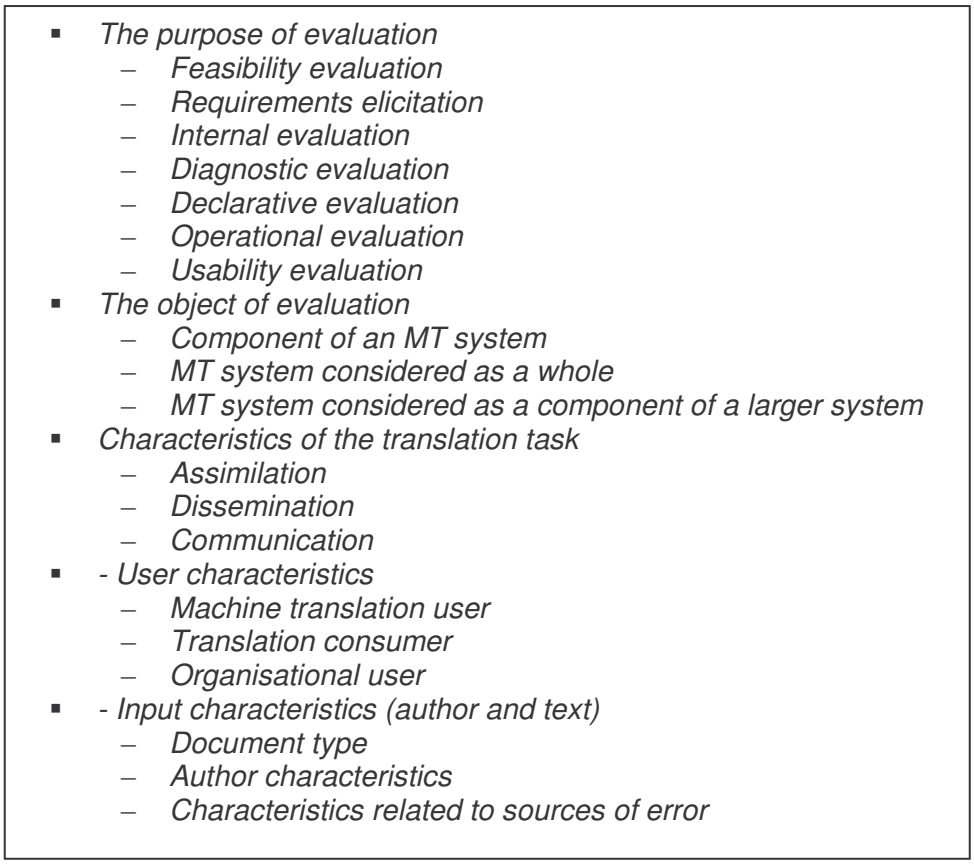

Figure 1. Part I classification

3.2 Part II: Quality Characteristics

Part II is a classification of the MT software quality characteristics into hierarchies of sub-characteristics, with metrics at the bottom level. The upper levels match the ISO/IEC 9126 characteristics, while the lower levels instantiate the specialization of the standard for MT. The quality characteristics bottom out into metrics that are specific to MT systems, and synthesize the state of the art in MT evaluation today. Figure 2 represents a small excerpt of the characteristics of Part II. In the left-hand box some of the top level characteristics of Part II are displayed, while all the sub-qualities that contribute to *Efficiency* are displayed in the box on the right.

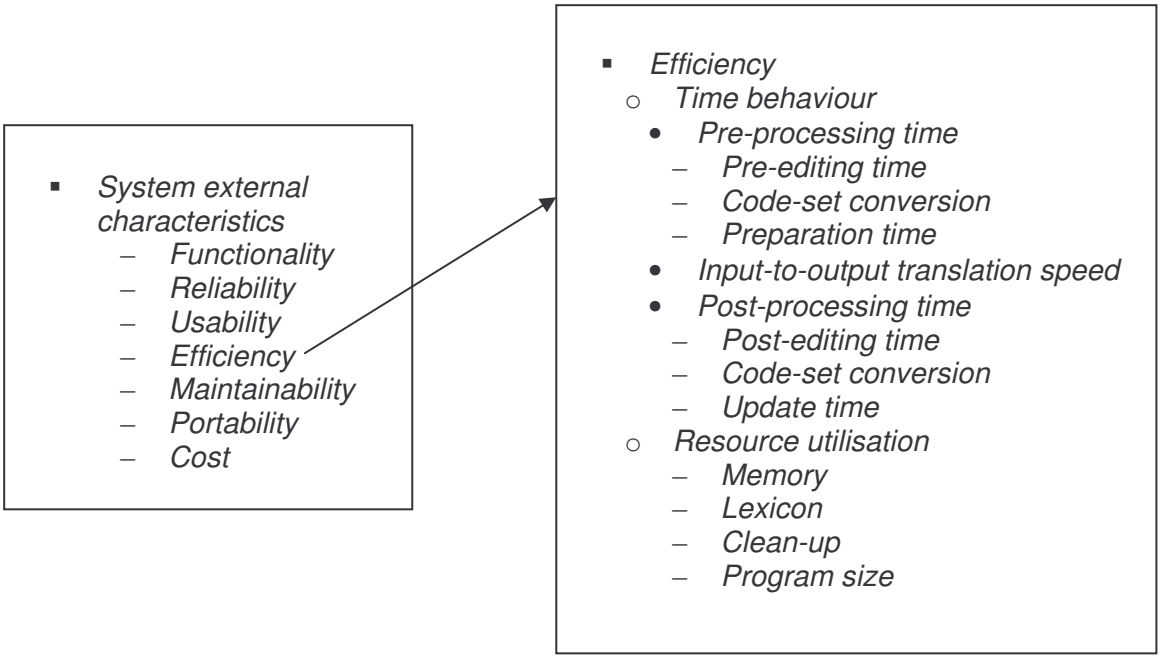

Figure 2. Excerpt form Part II

3.3 Links from Part I to Part II

The most original aspect of FEMTI is the linking mechanism, or mapping, from Part I to Part II. After selecting a context of use, purpose and object of evaluation from Part I, this mechanism is used to propose to the evaluator a list of qualities that should be evaluated. This contextualized quality model consists of characteristics and sub-characteristics from Part II that are relevant to the particular context of use defined through the selected characteristics of Part I. The theoretical bases for the linking mechanism are stated in (Hovy, King and Popescu-Belis 2002b:3.2-3.3).

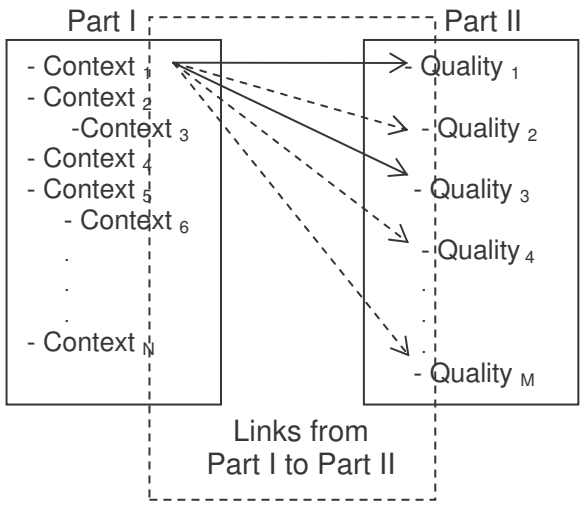

Figure 3. Linking between user-defined contexts and quality characteristics

4 Implementation

The version of FEMTI developed by the Evaluation Group of the ISLE project is available at http://www.issco.unige.ch/projects/isle/femti/ and consists mainly of an interface to navigate through the contents (Part I and Part II) rather than an interactive tool that guides the user towards obtaining parameterised quality models. Our current goal is to design an enhanced tool that helps evaluators to select online the user requirements from Part I, and then returns a quality model that can be modified online as well. The new version of FEMTI presented in this paper is the first step to achieve our goal. In this section we provide more details about the implementation of FEMTI's current version.

4.1 Dynamic Web Publishing

The ISLE interface to FEMTI is content-oriented as it used the eXtensible Markup Language (XML) to store and structure the content (qualities and metrics) gathered throughout the many projects that predated FEMTI. Its content is statically converted to several HTML files that allow the display of the two parts. The content-oriented approach will be pursued because it provides more flexibility and ease of change, by separating the data from the presentation mechanisms.

For the current version, we adopted a dynamic way of generating the HTML pages using the open source Cocoon[2] publishing framework. Cocoon is based on the principle of "*separation of concerns"*, letting the developers focus on different aspects of the application. More precisely, Cocoon considers three independent layers: content, presentation layout and logic. These three aspects are treated using XML files, eXtensible Stylesheet Language Transformations (XSLT) stylesheets and eXtensible Server Pages (XSP)[3], respectively. Their combination into a website is achieved by the Cocoon pipelines, whose operation is shown in Figure 4. This innovative architecture delivers content in several formats by simply changing the components of the pipeline to indicate the desired output format; content can be rendered as PDF, RTF, HTML, raw text or XML, among others.

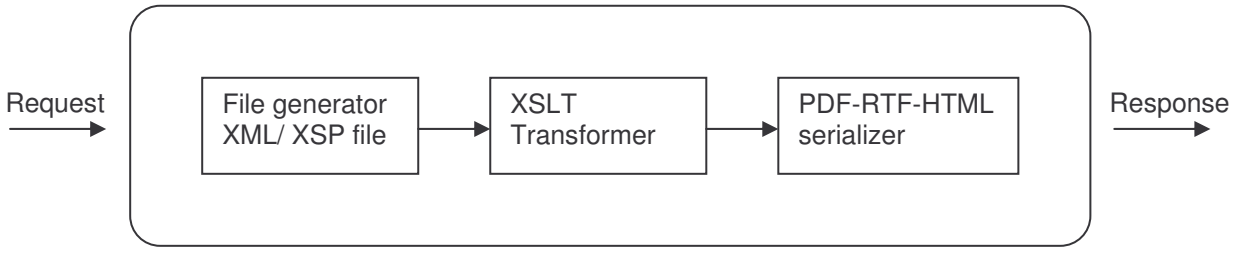

Figure 4. Cocoon pipeline

[2] More information at http://cocoon.apache.org/
[3] More details about this technology at http://cocoon.apache.org/2.1/userdocs/xsp/

Another advantage of using Cocoon instead of static web pages is that we can continue working on the content of FEMTI without interfering with the implementation of the interfaces. This is especially important when new results emerge from research in MT evaluation and must be added to FEMTI; for example, we might decide to add or delete an MT metric or characteristic but this event should not imply reworking the interfaces.

4.2 Workflow and Interface

FEMTI's web interface offers a simple way to generate evaluation plans in a few steps, as summarized by Figure 5.

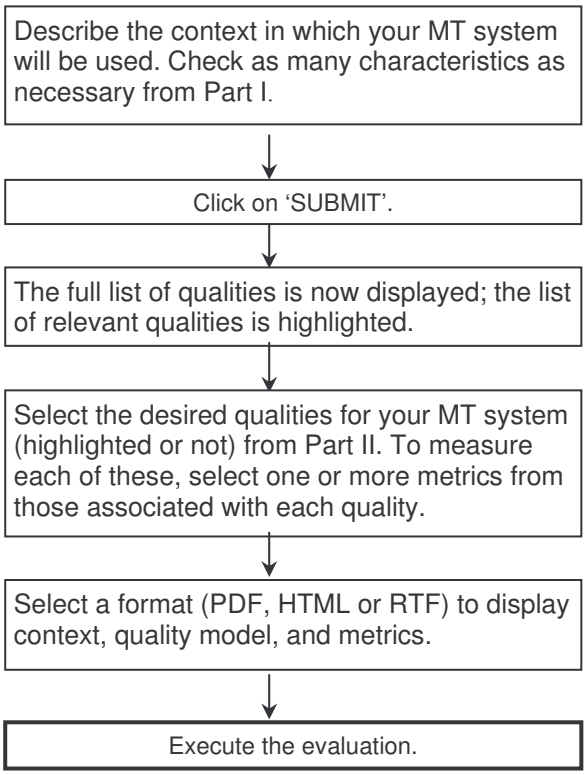

Figure 5. FEMTI workflow for creating an evaluation plan

Figure 6 shows the interface that enables evaluators to follow these steps. The left-hand frame shows the contents of Part I, and the right-hand frame shows the contents of Part II. At the bottom of this frame the evaluators will find buttons to submit a context of use once they have selected all characteristics. After submission, the interface will display the resulting quality model, which can then be saved in one of the above-mentioned formats. For better navigation both parts can be collapsed or expanded by clicking on the minus and plus signs beside each node.

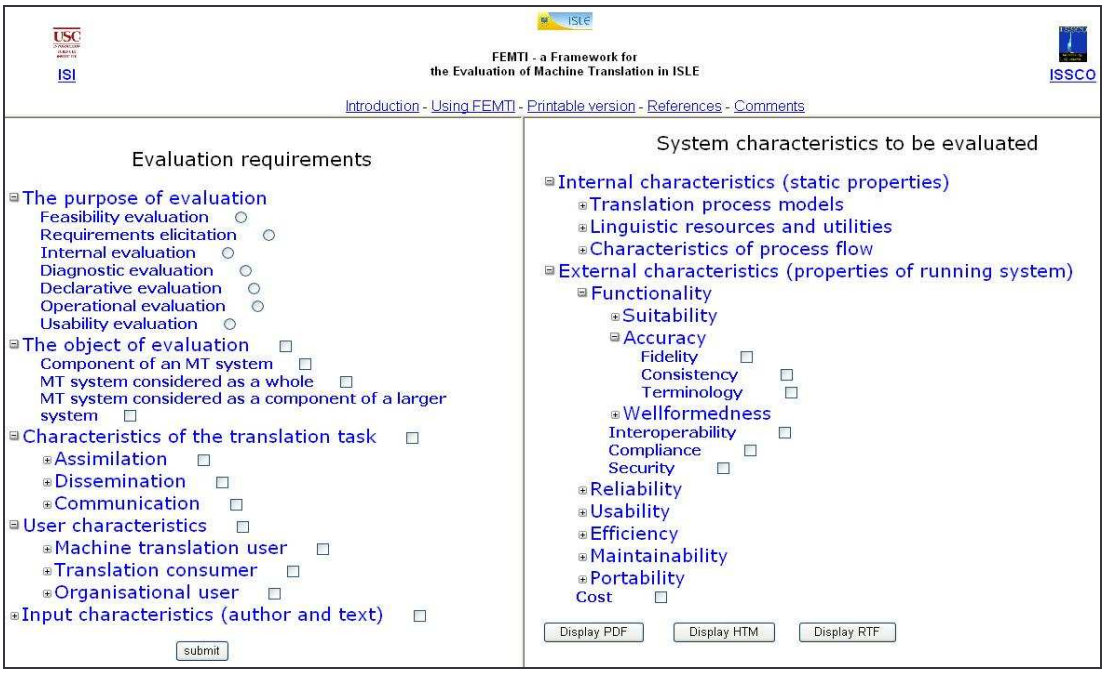

Figure 6. FEMTI interface

The following additional links in the top frame are always accessible to help evaluators while using FEMTI as well as to let them provide feedback:

- *Introduction*: leads to a summary of FEMTI's background and components.
- *Using FEMTI*: explains step-wise how to generate a tailor-made evaluation plan using this framework.

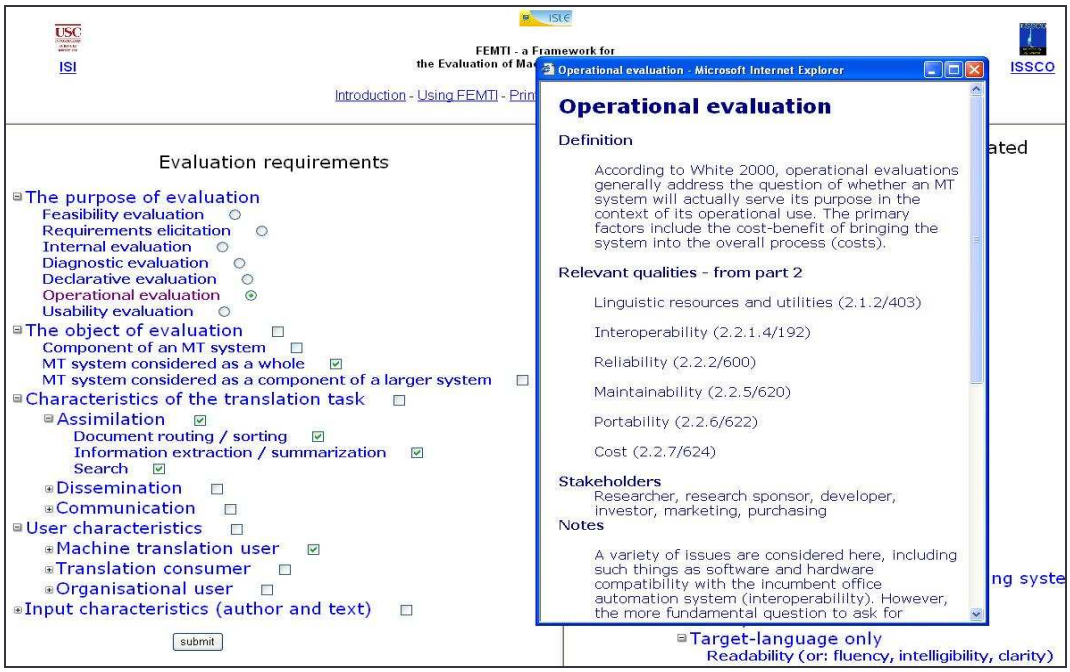

Figure 7. Defining a context of use

- *Printable version*: a handy version of both Part I and II, useful to work offline or prepare the evaluation beforehand.
- *References*: general bibliography of machine translation, including the list of references used in FEMTI.
- *Comments*: besides the comment link for each particular characteristic, FEMTI users can post general comments using the form displayed here.

As an example of using FEMTI, suppose we want to buy an MT system and define our context of use shown in Figure 7 by clicking the boxes as to indicate the following choices:

- Type of evaluation is to be performed: operational evaluation
- Object to be evaluated: MT system considered as a whole
- Type of task to be performed by the MT system: assimilation, i.e. to monitor a large volume of texts produced outside the organization
- Type of user of the system: machine translation user

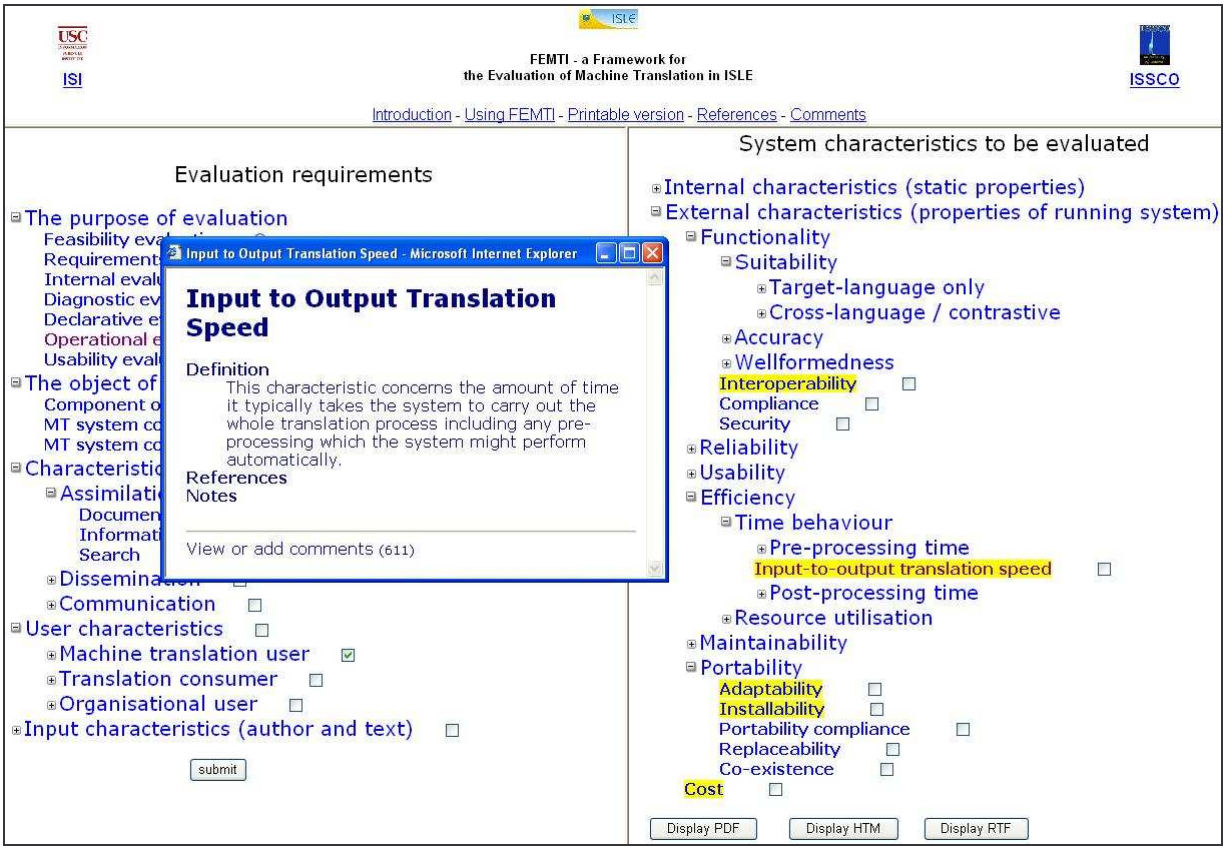

Figure 8. Qualities proposed by FEMTI

Clicking on an item in the Part I hierarchy activates a pop-up window which displays the definition and some additional information about it; at the bottom of this pop-up window a link offers the possibility to add or view comments on the selected item. Figure 7 shows the result of clicking on the "*operational evaluation"* node.

After submitting our context (Part I), Part II is displayed on the right-hand side where the qualities suggested by FEMTI are amongst other: *input-to-output translation speed*, *improvability*, *installability*, *adaptability* and *cost;* this is shown in Figure 8. As with Part I, clicking on a characteristic of Part II will activate a pop-up window containing a definition of the characteristic. Figure 8 shows the results of clicking on "*input-to-output translation speed*" characteristic.

The next step is to choose which the qualities and metrics desired for the MT system under evaluation. Figure 9 shows what happens when a quality is selected: in this case, the metrics available for *input-to-output translation speed* are displayed.

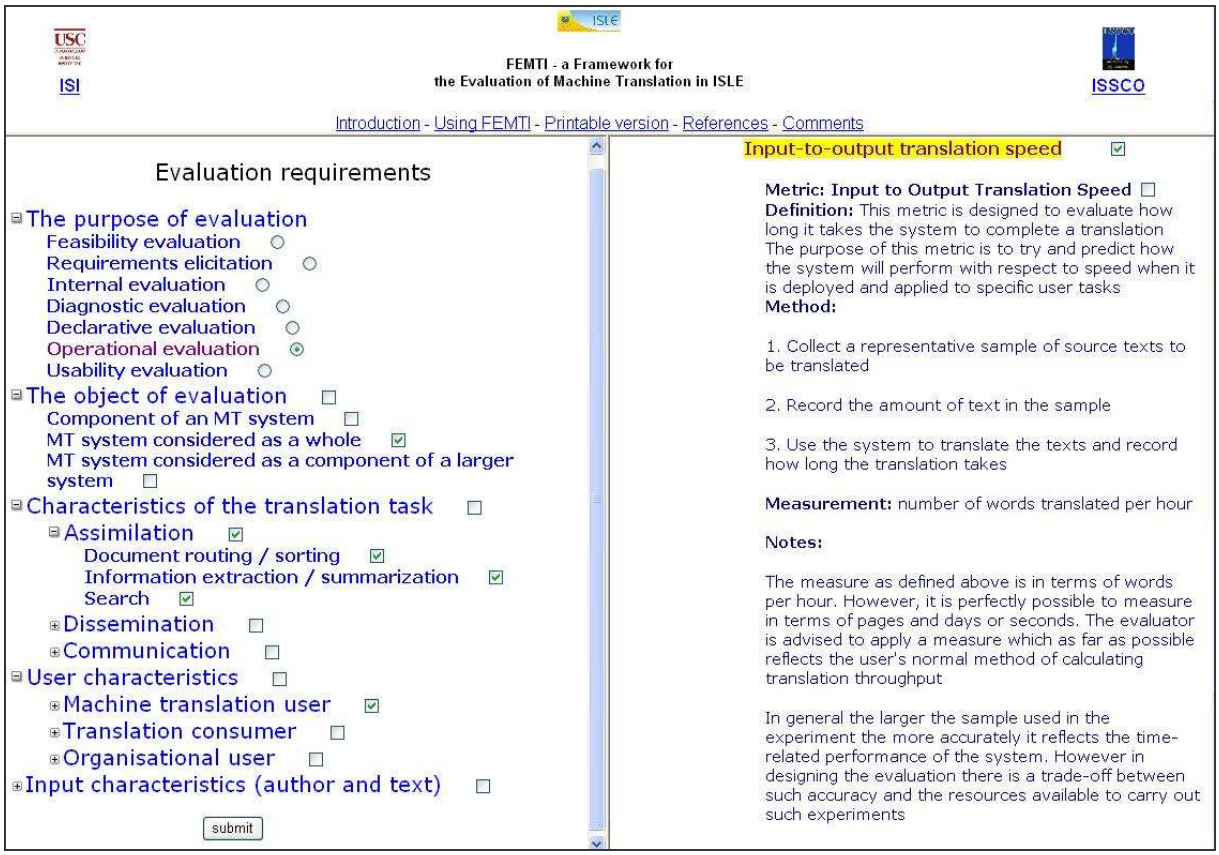

Figure 9. Metrics associated with one of the qualities suggested by FEMTI

The final step is to choose a format to display the evaluation plan by pressing the corresponding button; we offer three standard formats (PDF, RTF and HTML) to cover different user working environments. The evaluation plan includes the context of use, qualities and metrics with the corresponding definitions and notes and can, of course, be saved or used to generate another document, for example by reformatting a plan saved in RTF.

5 Future Work

Since FEMTI is founded on the wealth of experience of the MT community at large (both users and developers), we are constantly on the look-out for feedback and new knowledge which will enhance the performance of the tool in providing as comprehensive a resource as possible to support evaluators in creating the most suitable evaluation plan for their needs. Therefore, our future work includes an expert's interface in addition to the one for creating a specific evaluation plan, which will enable expert users to propose links from Part I to Part II of FEMTI as well as to modify the knowledge in FEMTI.

For the moment, a context or quality characteristic is either "applicable" or "not applicable", thus limiting the evaluator in his choice. A further improvement will involve changing this way of selecting characteristics to allow evaluators to rank them, for example as *indispensable*, *important* and *not important*. This change will as well imply further work to study how it affects the linking mechanism and the quality model itself.

The basic skeleton is at present well-developed, but we still need to put more flesh on the bones, in particular by incorporating the advice and experience of MT users and of MT evaluators. The future work outlined here aims at collecting and consolidating user reactions and at transforming FEMTI into a more powerful tool of real practical utility to the whole of the MT community.

6 References

Babych B. and Hartley T. (2004): "Extending the BLEU MT Evaluation Method with Frequency Weightings", *Proceedings of ACL 2004 (42nd Annual Meeting of the Association for Computational Linguistics)*, Barcelona, Spain, pp. 621-628.

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

EAGLES Evaluation Working Group (1996): *EAGLES Evaluation of Natural Language Processing Systems*, Final Report Center for Sprogteknologi, EAG-EWG-PR.2 (ISBN 87-90708-00-8).

Hartley A. and Popescu-Belis A. (2004): "Évaluation des systèmes de traduction automatique", in S. Chaudiron (ed.), *Évaluation des systèmes de traitement de l'information*, Paris, Hermès, pp. 311-335.

Hovy E. H., King M. and Popescu-Belis A. (2002a): "An Introduction to MT Evaluation", *Handbook of the LREC 2002 Workshop "Machine Translation Evaluation: Human Evaluators Meet Automated Metrics"*, Las Palmas de Gran Canaria, Spain, pp. 1-7.

Hovy E. H., King M. and Popescu-Belis A. (2002b): "Principles of Context-Based Machine Translation Evaluation", *Machine Translation*, vol. 17, n° 1, pp. 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: Information Technology -- Software Product Evaluation -- Part 1: General Overview*, Geneva, International Organization for Standardization / International Electrotechnical Commission.

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

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

Van Slype G. (1979): *Critical Study of Methods for Evaluating the Quality of Machine Translation*, European Commission / Directorate for General Scientific and Technical Information Management (DG XIII), BR 19142.

White J. S. and O'Connell T. A. (1994): "The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches", *Proceedings of AMTA Conference, 5-8 October 1994*, Columbia, MD, USA.