Improving quality models for MT evaluation based on evaluators' feedback

Paula Estrella[1], Andrei Popescu-Belis[2], Maghi King[1]
[1] ISSCO/TIM/ETI, University of Geneva
40, bd. du Pont-d'Arve
1211 Geneva, Switzerland
[2] IDIAP Research Institute
Av. des Prés-Beudin 20
1920 Martigny, Switzerland

E-mail: paula.estrella@issco.unige.ch, andrei.popescu-belis@idiap.ch, maghi.king@gmail.com

Abstract

The Framework for the Evaluation for Machine Translation (FEMTI) contains guidelines for building a quality model that is used to evaluate MT systems in relation to the purpose and intended context of use of the systems. Contextual quality models can thus be constructed, but entering into FEMTI the knowledge required for this operation is a complex task. An experiment has been set up in order to transfer knowledge from MT evaluation experts into the FEMTI guidelines, by polling experts about the evaluation methods they would use in a particular context, then inferring from the results generic relations between characteristics of the context of use and quality characteristics. The results of this hands-on exercise, carried out as part of a conference tutorial, have served to refine FEMTI's 'generic contextual quality model' and to obtain feedback on the FEMTI guidelines in general.

1. Introduction

The Framework for the Evaluation of Machine Translation (FEMTI) was introduced by the ISLE Evaluation Working Group, as an implementation of the EAGLES evaluation guidelines in the case of MT systems. The guidelines help evaluators to define quality models used to evaluate MT systems, based on the purpose and intended context of use of the systems (Hovy et al. 2002, King et al. 2003). The FEMTI framework is publicly available and benefits from user-friendly interfaces that guide evaluators with the definition of quality models and the selection of metrics (http://www.issco.unige.ch/femti/).

In this paper, we report results from an experiment aimed at transferring knowledge from MT evaluation experts into the FEMTI guidelines. After a brief reminder of FEMTI's main features (Section 2), we demonstrate their relevance by relating them to two recent examples of contextualized evaluation of MT systems, for real-world uses (Section 3). The conditions of the hands-on MT evaluation exercise are outlined in Section 4, while the method for transforming the experts' output into FEMTI knowledge and the results obtained with this approach are described in Section 5. Section 6 shows the integration of results into the FEMTI framework, while some conclusions and future work are outlined in Section 7.

2. Evaluation of MT Systems in Context

2.1. FEMTI Components

FEMTI is made up of three components: the first one ('Part I') is a classification of possible user requirements or, more generally, of *contexts of use* of an MT system. Therefore, FEMTI Part I classifies the characteristics concerning the *task* to be performed by the MT system, the *author* and *text type* of the input to the system, and the *type of user* of the system. The second component of FEMTI ('Part II') is a classification of quality characteristics based on the general ones first defined in the ISO/IEC 9126-1 standard for software evaluation (ISO/IEC 1991, 2001). The six top level characteristics of FEMTI Part II are the ISO ones, namely *functionality*, *reliability*, *usability*, *efficiency*, *maintainability* and *portability*, plus an extra characteristic, namely *cost*. Each top level characteristic is further decomposed into a hierarchy of characteristics which are specific to MT systems, and human and/or automatic metrics are attached to each node of this hierarchy, which can be used to evaluate the corresponding feature of the system, i.e. a quality attribute. The third component of FEMTI is the linking mechanism between Part I and Part II described in the next section.

2.2. Linking Mechanism

The generation of evaluation plans is based on particular quality models, which are subsets of the FEMTI Part II repertoire of quality characteristics and metrics. To extract such a subset, evaluators define first the requirements for the system (Part I), then apply a mechanism that uses the relations between each user requirement and the relevant quality characteristics of the system to be evaluated to construct the quality model.

Therefore, the influence of the intended context of use on the quality model is represented formally as a matrix, called a *generic contextual quality model* or GCQM (Popescu-Belis et al. 2006). Each row of the FEMTI GCQM represents a quality characteristic of Part II, and each column is a context characteristic from Part I. (For efficiency and readability, only those items from Parts I and II – context/quality characteristics – that have a connection or relation with one or more items are

represented in the actually implementation of the GCQM.) The fact that a specific context characteristic requires the evaluation of a specific quality is represented as a relation or *link between one item in Part I and another item in Part II*. In the matrix representation of the GCQM, a link is represented by a *weight* in the corresponding cell indicating the importance of the link.

	C_{10}	C_{12}	C_{35}	C_{40}
Q_1		**n/a**		
Q_{12}		**1**		**n/a**
Q_{13}			**2**	
Q_{24}	**3**			

Figure 1. Excerpt of a GCQM showing four context characteristics and four quality characteristics: context C_{10} has the highest connection to quality Q_{24}

The scale used for the weights is: **1** (low importance), **2** (medium importance), **3** (high importance) or **n/a**, which indicates the presence of a link with no specific weight (this allows backwards compatibility with the first version of FEMTI, which did not use weighted links). For example, context C_{10} is linked to quality Q_{24} indicated by a non-zero value and the weight of the link is **3**, as shown in Figure 1. When an evaluator specifies a context of use with FEMTI, the FEMTI software represents the selected characteristics as a *context vector*, which is used internally to retrieve relevant quality characteristics from Part II that should be evaluated for that context of use. This process thus generates a *quality vector* by computing the matrix product of the *context vector* and of the GCQM matrix.

2.3. Automating the Generation of Quality Models with FEMTI

Recent work on FEMTI was devoted to develop new web-based interfaces as well as to define and implement an automatic tool that provides suggestions for relating contexts of use to quality characteristics, as described formally in the previous section (Estrella et al. 2005). There are two interfaces to the framework, one for evaluators and one for experts. The evaluators' interface helps to generate MT evaluation plans based on a given context of use, offering significant improvement in terms of automation over the first version of FEMTI. The interface for experts is intended to help experts define the links in the GCQM between items in Part I and items in Part II, based on their own implicit knowledge of relevant quality characteristics, or on experience from previous evaluations. Experts can assign weights on the links as mentioned above, including the 'n/a' mention if a relation is hypothesized but its precise strength is not apparent to them (many links imported from the initial version of FEMTI are currently marked with 'n/a').

The two interfaces to FEMTI are tightly related, since it is through the experts' interface that GCQMs are created, but it is in the evaluators interface that *an averaged* GCQM combining those created by experts is used. In the evaluators' interface, once an evaluator defines an intended context of use (represented internally as a vector), the scripts implementing the linking mechanism described in the previous section retrieve relevant links from the *averaged* GCQM and compute the relevant quality characteristics from Part II, which are highlighted by the interface. The last step in building a customized quality model consists of selecting the actual quality characteristics that will be evaluated among those proposed by FEMTI, and finally the corresponding metrics to measure them.

3. Two Recent Examples of Contextual MT Evaluation

Evaluation of MT systems is a complex task to normalize, even with a user-friendly interface that automatically suggests relevant quality for an intended context of use. For many MT developers, evaluation remains a matter of measuring the "quality" of output text using reference-based metrics such as BLEU or human-based metrics such as accuracy/fluency. However, when evaluating software dedicated to real-life commercial applications, context-based evaluation cannot be avoided if optimal choices have to be made for large organizations, as the two following case studies show.

GPHIN is a multilingual Internet-based early warning system that gathers preliminary reports with potential public-health significance on a nearly real-time basis (Blench 2007). GPHIN continuously gathers and disseminates multilingual information from newswire and web sites, which requires several MT sub-systems or 'engines', which were carefully selected to ensure the success of the system as a whole. The requirements for these engines are very strict, given the mode of operation of GPHIN. In consequence, a list of important qualities that determine whether or not to incorporate a new MT engine into the system was defined, which includes robustness, high interoperability and ease of dictionary update. Robustness is a critical aspect, and no actual engine passed the stress tests, so a special module was developed to fix this issue. The GPHIN system proved to be very productive in monitoring and efficient in early detection of outbreaks, the most relevant case being the detection of a SARS outbreak almost three months in advance.

From an even more systematic perspective close to FEMTI, Stadler and Peter-Spörndli (2007) present a quality model developed for MT systems to be used by a translation service provider, which was based on the ISO definition for quality and on the premise that different types of customers have different needs. The main quality attributes in this model were extracted from user feedback and include translation quality, usability, actual translation directions and supported file formats. This work shows many similarities to the FEMTI context-based approach to evaluation; in particular, as the selected quality attributes are present in FEMTI, this model is fully compatible with FEMTI provided the same terminology for quality

characteristics is used. Ideally, the same quality model could be obtained using FEMTI, by entering the respective context characteristics, if the current GCQM was completely populated with links.

4. Manually Relating Contexts of Use and Quality Models

The topics and tools related to context-based evaluation constituted the basis of a tutorial offered as part of the Machine Translation Summit XI. The objective was not only to introduce the tools to potential users, but also to obtain input from them, based on their previous evaluation experience, in order to improve FEMTI by enriching its GCQM with more links between Parts I and II.

4.1. Significance of the Experiment

In the current state, FEMTI is a valuable source of knowledge about MT evaluation, and the evaluation plans generated with it have the potential to be used as a starting point for a true evaluation. However, users of FEMTI with less experience in MT evaluation might need some guidance for the generation of evaluation plans, for example about the quality characteristics of the system he/she would like to evaluate. Part of such guidance could be provided by the links in FEMTI's GCQM.

Populating links in the GCQM is a hard and demanding effort, given the large size of the GCQM and the fact that these links should be validated by several MT experts. Therefore, the purpose of the experiment presented in this paper is to enrich the current GCQM by generating more links based on experts' knowledge and experience.

4.2. Description of the Elicitation Exercise

The goal was to create one GCQM for a specific scenario, based on the participants' experience with MT system. Part of the tutorial was indeed devoted to the practical activity, which consisted of the following steps:

1. Identify the context characteristics from FEMTI Part I that would best formalize the scenario of use of an MT system described verbally in Figure 2 below.
2. Indicate the quality characteristics from FEMTI Part II, which are related to each of the selected context characteristic.
3. If possible, indicate the importance of each quality characteristic, for each context characteristic, on a 3-point scale.

For practical reasons, the participants to the tutorial could not use the web-based experts' interface, but a printed compilation of FEMTI's content was prepared to their intent. Participants were arranged in four groups of about four persons and focused only on a small subset of context and quality characteristics, due to time constraints. The results of each group were summarized by the organizers during a break, and then discussed at the end of the tutorial; these results are presented in the following section.

> You have a contract with the International Olympic Committee to track what is said in the Chinese press about the preparations for the Olympic Games in China. You do not read Chinese, but you do have a limited budget for translation.
> You think you may be able to use an MT system to select relevant articles, which you will then get translated by humans.
> - What system characteristics are relevant? Which are the most important? And the least important?
> - What quality characteristics correspond to each of the system characteristics you have picked out?
> - What is their relative importance? (Rate importance as: 3 = very important, 2 = important, 1 = nice to have)

Figure 2. Scenario of use given to participants.

5. Results of the Experiment

The scenario of use (Figure 2) was defined as a compromise between precision and generality: indeed, participants needed a reasonably clear scenario of MT use to define evaluation methods, but this scenario had to be general enough to avoid biasing the participants towards any specific context/quality characteristics. We expected to see some overlap across groups in the answers, i.e. that most or some of the groups would use the same elementary context characteristics to define the scenario, and that they would relate them to the same quality characteristics. To illustrate the outcome of the exercise, the result of one of the groups – context characteristics and related quality characteristics with their weights – is shown in Table 1.

Context characteristics	Quality characteristics
Document routing	terminology (3) fidelity (3) well-formedness (1) comprehensibility (2) dictionary updating (3)
Superior (author's proficiency in source language)	dictionaries (3)
Novice (user's proficiency in source language)	fidelity (3) well-formedness (2) translation speed (3) introduction cost (3)

Table 1. Results generated by group 4.

However, it appeared that each group interpreted the scenario in a slightly different manner, choosing different elementary context characteristics to define the translation task of the MT system. The groups agreed on the top level context characteristic, 'assimilation', but when further specifying the task, each group interestingly chose a different sub-task: 'search' vs. 'information extraction' vs. 'document routing/sorting' (the fourth group did not specify the task any further). The diverging choice of the exact specification of the translation task then made groups focus also on different relevant qualities, though

quite close to each other, leading in the end to the creation of different GCQMs for each group. (It is, of course, reasonable to design specific evaluation plans for systems intended to perform different tasks.) In addition, it is worth noting that the contexts of use that were defined not only focused on the translation task, but also included other important items such as the user of the MT system, thus suggesting that FEMTI offers indeed a richer methodology for evaluation.

In the answers obtained from the four expert groups participating in these experiments, several context characteristics were shared by more than one group, independently of the exact translation task chosen by the groups, suggesting that for scenarios similar to the one proposed in Figure 2, the following context characteristics are of significant interest for an evaluation:

- domain or field of application (text/document type – 1.3.1.2 in FEMTI)
- superior proficiency in source language (author – 1.3.2.1.4)
- novice proficiency in source language (user – 1.4.1.2.1)
- superior / distinguished proficiency in target language (user – 1.4.1.3.4/5)

The numbers after a context or quality characteristic indicate their position in the FEMTI framework, available at http://www.issco.unige.ch/femti/.

We also expected links (relations between context and quality characteristics) to reach a significant number of terminal nodes in Part II, i.e. the measurable attributes. Indeed, if most of the context characteristics were linked to only one quality characteristic, this would mean that the context has no longer an impact on evaluation and assessing that single quality characteristic would always be enough, which is not the case for real-world systems, as illustrated above. In our experiment, we observed a diversity of answers, due to the different contexts defined in the first place. However, a series of quality characteristics appear to be important for the given scenario, such as *fidelity*, *terminology*, *dictionaries*, *input to output translation speed* and *cost*.

The results of the exercise – i.e. the context characteristics, quality characteristics, and weighted links – were summarized at the end of the tutorial in a table representing the GCQM jointly generated by the experts, as shown in Figure 3. The groups were numbered from 1 to 4 (first figure in each table cell) and the numbers 1–3 in parenthesis denote the relevance of each relation assigned by each group; no number between parenthesis means that there is a link but no weight was assigned, as explained in Section 2.2. An example is shown in Figure 3, where some of the results from Table 1 transferred to the GCQM are marked with dashed boxes, e.g. the translation task is *document routing (1.2.1.1)* and it is related to *terminology (2.1.1.1)*, assigning a score of 3 for the importance of this relation.

FEMTI: GCQM	1.2.1.1 Document routing	1.2.1.2 Information extracti	1.2.1.3 Search	1.3.1.1 Genre	1.3.1.2 Domain or field of a	1.3.2 Author characteristic	1.3.2.1.4 Superior	1.3.3.1 Intentional error so	1.4.1.2.1 Novice	1.4.1.2.2 Intermediate	1.4.1.3.4 Superior	1.4.1.3.5 Distinguished	1.4.2 Organisational user	1.4.2.1 Quantity of translat	1.4.2.2 Number of personn	1.4.2.3 Time allowed for tra
2.1.1.1 Terminology	4(3)	2(2)			3(3) 2(2)											
2.1.1.2 Fidelity	4(3)	2(1)	1		2(1)				4(3)							
2.1.1.3 Well-formedness	4(1)								4(2)							
2.1.2.1.2 Comprehensibility	4(2)	2(3)			2(3)											
2.1.2.2.2 Coverage of corpus-specific phenomena				1												
2.1.2.4.1 Languages							2(3)				2(3)					
2.1.2.4.2 Dictionaries		2(2)			2(2)		4(3)									
2.1.2.4.4 Corpora			1	1				1								
2.1.2.5.2 Post-translation activities									1 (1)			1(1)				
2.1.2.5.4 Dictionary updating	4(3)				3(3)											
2.4.1.1 Overall Production Time														2(3)		2(3)
2.4.1.2 Pre-processing time																
2.4.1.3 Input to Output Translation Speed			1						4(3)				3	2(3)		2(3)
2.4.2.1 Memory usage														2(1)		
2.5.2 Changeability					3(3)											
2.5.2.3 Ease of dictionary update							2(1)				2(1)					
2.5.3 Stability																2(1)
2.7 Cost																2(3)
2.7.1 Introduction cost									4(3)							
2.7.3 Other costs													3			

Figure 3. Table summarizing the answers (the first number in each cell refers to the group that indicated the link, and the number in parenthesis is the weight).

6. Contribution to FEMTI

From this hands-on exercise, around 40 new links were added to FEMTI's GCQM. Although these changes and the whole mechanism behind FEMTI's interfaces are transparent to the user, once the links are sufficiently populated, the resulting behavior of the FEMTI interface is that many more quality characteristics get suggested once context characteristics are specified. Another interesting aspect of introducing more links is that weights on links are considered for the final quality model to rank the quality characteristics according to their importance, that is, if many of the selected context characteristics point to the same quality characteristic, this one becomes more important to the evaluation and will appear higher in the list of quality characteristics to assess. Figure 4 illustrates the integration of these results into the current version of FEMTI; the qualities highlighted in Part II (right-hand side) come from links previously present in the GCQM plus the new links when selecting the context of Table 1.

7. Conclusion and Future Work

Through the tutorial offered as part of the Machine Translation Summit XI, we were able to gather valuable feedback from MT users and evaluators, as well as to continue the integration of expertise into FEMTI. We believe that continuing such hands-on exercises is an important part of FEMTI development, and will organize a new tutorial at LREC 2008. The results obtained with the 2007 practical exercise were incorporated to the existing GCQM in FEMTI, as they are part of the contribution to FEMTI by the MT community and should be available to the general public.

We intend to carry out larger scale experiments in the future, to populate relations between Parts I and II of FEMTI, as well as to identify areas of improvement for each of the two parts, possibly reorganizing the classifications or including new qualities or metrics. Another intended activity is to generate templates or typical use cases of FEMTI to facilitate its application to real-world scenarios, again based on input from MT experts.

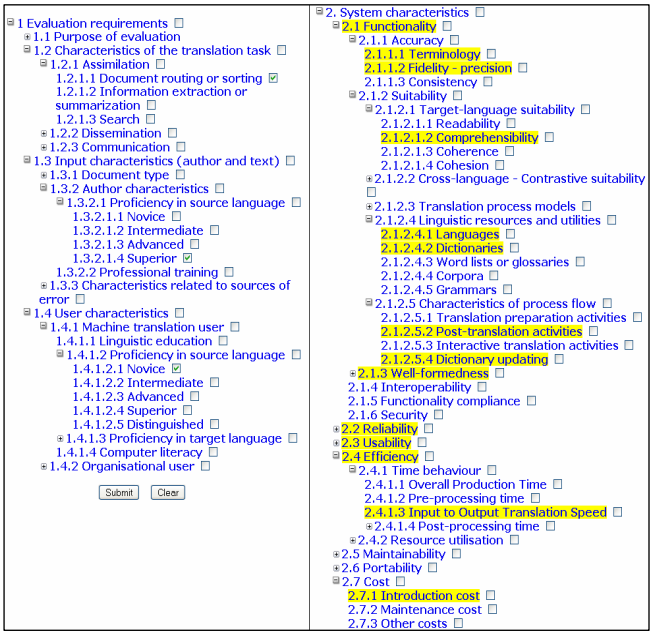

Figure 4. Links retrieved from FEMTI's GCQM when selecting context of use defined in Table 1.

8. References

Blench, M. 2007. Global Public Health Intelligence Network (GPHIN). In *MT Summit XI*, 45 - 49. Copenhagen, Denmark.

Estrella, P., A. Popescu-Belis, and N. Underwood 2005. Finding the System that Suits you Best: Towards the Normalization of MT Evaluation. In *27th ASLIB International Conference on Translating and the Computer*, 23-34. London, UK.

Hovy, E., M. King and A. Popescu-Belis 2002. Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, 17(1), p.43-75.

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

King, M., A. Popescu-Belis, and E. Hovy 2003. FEMTI: creating and using a framework for MT evaluation. In *Machine Translation Summit IX*, 224-231. New Orleans, LA, USA.

Popescu-Belis, A., P. Estrella, M. King, and N. Underwood 2006. A model for context-based evaluation of language processing systems and its application to machine translation evaluation. In *LREC 2006 (5th International Conference on Language Resources and Evaluation)*, 691-696. Genova, Italy.

Stadler, H.-U., and U. Peter-Spörndli 2007. The quest for machine translation quality at CLS Communication. In *MT Summit XI*, 435 - 442. Copenhagen, Denmark.