MT Summit XI Tutorial, Copenhagen, Denmark

Context-based Evaluation of MT Systems: Principles and Tools

Maghi King, Andrei Popescu-Belis and Paula Estrella
ISSCO/TIM/ETI, University of Geneva

Monday, September 10, 2007, 9:00–12:30, Room 212

This tutorial introduces a principled approach to the evaluation of MT systems in their context of use. A method for defining contextual evaluation plans will be described, along with an interactive tool supporting the definition of such plans. This tutorial provides the audience with knowledge and tools for MT evaluation, which will enable them to better justify their choices of particular MT systems, or to argue about the quality of MT systems in a more principled way.

Framework for Machine Translation Evaluation (FEMTI)

FEMTI is an achievement of the Evaluation Working Group of the ISLE European project. FEMTI relates the quality model used to evaluate an MT system to the context of use of the system. The initial version of FEMTI, which synthesized many of the previous efforts to structure the concepts and metrics used in MT evaluation, was quite complex to use. Therefore, a web-based, user-friendly tool supporting the design of a contextual quality model was implemented, and is now available online at http://www.issco.unige.ch/femti. The new graphical interfaces help evaluators to produce evaluation plans for MT systems intended for a given context, and allow experts to contribute to FEMTI with their knowledge about the qualities and metrics that are relevant in a given context.

As dissemination and feedback are constitutive features of FEMTI, this tutorial intends to stimulate the debate within the MT community by presenting the most recent features of the FEMTI framework and interfaces. FEMTI has greatly benefited from the feedback obtained through hands-on studies and tutorials (UniGe May 2001, NAACL'01, LREC'02), paper-presentation workshops (LREC'00, AMTA'00, MT Summit VIII 2001, MT Summit IX 2003) and expert group meetings (UniGe May 2001, LREC'02, and USC/ISI February 2003). Some of these workshops already included hands-on exercises in MT evaluation, supported by preliminary versions of FEMTI.

References

- Hovy E. H., King M. & Popescu-Belis A. "Principles of Context-Based Machine Translation Evaluation." *Machine Translation*, vol. 17, n° 1, pp. 1-33, 2002.
- Estrella P., Popescu-Belis A. & Underwood N. "Finding the System that Suits you Best: Towards the Normalization of MT Evaluation." *Proc. of the 27th International Conference on Translating and the Computer*, ASLIB, 24-25 November 2005, London.
- Popescu-Belis A., Estrella P., King M. & Underwood N. "A model for context-based evaluation of language processing systems and its application to machine translation evaluation." *Proc. of LREC 2006 (4th International Conference on Language Resources and Evaluation)*, Genoa, pp. 691-696.

Contact information

ISSCO/TIM/ETI
University of Geneva
40, bd. du Pont-d'Arve
1211 Geneva 4, Switzerland
maghi.king@gmail.com, andrei.popescu-belis@issco.unige.ch, paula.estrella@issco.unige.ch
http://www.issco.unige.ch and http://www.issco.unige.ch/femti.

Outline of the tutorial

- Principles and methods for context-based evaluation of machine translation.
- Main concepts and contents of FEMTI: classification of contexts of use, classification of quality characteristics, correspondences or links between them.
- FEMTI interface for evaluators: specification of a mechanism that helps evaluators to generate MT evaluation plans based on the context of use they specify; instructions of use, examples, and hints about the implementation.
- FEMTI interface for MT evaluation experts: formal relation based on 'context vectors', 'quality vectors' and 'generic contextual quality models' (GCQM); instructions of use for creating or modifying a GCQM using the interface; demonstration.
- Practical exercise offering the audience the opportunity to apply the FEMTI guidelines to produce an evaluation plan, using a simplified, paper-based version of the framework.

Outline of the exercise

The objective of the exercise is to define a contextualized evaluation plan for an MT system, and then to compare the plans defined by various groups in order to improve the Generic Contextual Quality Model currently available in FEMTI.

1. Select one of the two scenarios of use outlined below for the MT system under evaluation. In agreement with the other participants, it will be possible to: (a) focus the entire group on only one scenario; (b) enrich the selected scenario with additional specifications of the intended use; (c) propose your own scenarios of use.

2. What context characteristics are relevant? Which are the most *vs*. least important? — Select from the list of characteristics of the context of use (FEMTI Part I) the ones that best describe the intended context of use of the MT system under evaluation. An outline of the list is provided in this document, and full versions will be available: (a) in print; (b) on the presenters' laptop; (c) via Internet at http://www.issco.unige.ch/femti.

3. What quality characteristics correspond to each of the system characteristics you have picked out? What is their relative importance? — Based on the context characteristics, on your own experience of MT systems, and on the indications available in FEMTI for these characteristics, proceed to select (from FEMTI Part II) a list of relevant quality characteristics that the MT system under evaluation should possess.

   Using the form provided below, indicate for each characteristic of the context, which qualities from the list (FEMTI Part II) are important for an MT system that will be used in that context; you can also quantify the importance on a 3-point scale (3: very important; 2:

important, 1: nice to have). Use the numbers of the characteristics (rather than names) to refer to them on the form.

An outline of the list of quality characteristics is provided below, and full versions will be available: (a) in print; (b) on the presenters' laptop; (c) via Internet at http://www.issco.unige.ch/femti. The final list of quality characteristics constitutes the contextualized quality model to be used for evaluation (metrics must be chosen for each quality).

4. When you have finished defining your contextualized quality model, please hand your form to the presenters, who will synthesize the results in preparation for a general discussion.

Scenarios of use for MT systems

1. 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 a machine translation system to select relevant articles, which you will then get translated by humans.

2. Each competitor in the Olympic Games will receive an information pack containing information on public transport, where to find shops, where to go in a medical emergency. The information will be prepared in Chinese by the local tourist office. Although it is not realistic to produce versions in all of the languages of the competitors, it is hoped to produce versions in as many languages as possible. You are hoping to use machine translation to increase the number of languages that can be produced.

Characteristics of the intended context of use

FEMTI Part I, or Evaluation Requirements

1.1 Purpose of evaluation

- *1.1.1 Internal evaluation*
- *1.1.2 Diagnostic evaluation*
- *1.1.3 Declarative evaluation*
- *1.1.4 Operational evaluation*
- *1.1.5 Usability evaluation*
- *1.1.6 Feasibility evaluation*
- *1.1.7 Requirements elicitation*

1.2 Characteristics of the translation task

- *1.2.1 Assimilation*
  - 1.2.1.1 Document routing or sorting
  - 1.2.1.2 Information extraction or summarization
  - 1.2.1.3 Search
- *1.2.2 Dissemination*
  - 1.2.2.1 Internal or in-house dissemination
    - 1.2.2.1.1 Routine internal dissemination

1.2.2.1.2 Experimental internal dissemination
1.2.2.2 External dissemination - publication
1.2.2.2.1 Single client external dissemination
1.2.2.2.2 Multi-client external dissemination
1.2.3 Communication
1.2.3.1 Synchronous communication
1.2.3.2 Asynchronous communication

1.3 Input characteristics (author and text)

1.3.1 Document type
1.3.1.1 Genre
1.3.1.2 Domain or field of application
1.3.2 Author characteristics
1.3.2.1 Proficiency in source language
1.3.2.1.1 Novice
1.3.2.1.2 Intermediate
1.3.2.1.3 Advanced
1.3.2.1.4 Superior
1.3.2.2 Professional training
1.3.3 Characteristics related to sources of error
1.3.3.1 Intentional error sources
1.3.3.2 Medium-related error sources
1.3.3.3 Performance -related error sources

1.4 User characteristics

1.4.1 Machine translation user
1.4.1.1 Linguistic education
1.4.1.2 Proficiency in source language
1.4.1.2.1 Novice
1.4.1.2.2 Intermediate
1.4.1.2.3 Advanced
1.4.1.2.4 Superior
1.4.1.2.5 Distinguished
1.4.1.3 Proficiency in target language
1.4.1.3.1 Novice
1.4.1.3.2 Intermediate
1.4.1.3.3 Advanced
1.4.1.3.4 Superior
1.4.1.3.5 Distinguished
1.4.1.4 Computer literacy
1.4.2 Organisational user
1.4.2.1 Quantity of translation
1.4.2.2 Number of personnel
1.4.2.3 Time allowed for translation

Quality characteristics: FEMTI Part II

2.1 Functionality

2.1.1 Accuracy

2.1.1.1 Terminology

2.1.1.2 Fidelity - precision

2.1.1.3 Well-formedness

2.1.1.3.1 Morphology

2.1.1.3.2 Punctuation errors

2.1.1.3.3 Lexis - Lexical choice

2.1.1.3.4 Grammar - Syntax

2.1.1.4 Consistency

2.1.2 Suitability

2.1.2.1 Target-language suitability

2.1.2.1.1 Readability

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

2.1.2.2.1 Style

2.1.2.2.2 Coverage of corpus-specific phenomena

2.1.2.3 Translation process models

2.1.2.3.1 Methodology

2.1.2.3.1.1 Rule-based models

2.1.2.3.1.2 Statistically-based models

2.1.2.3.1.3 Example-based models

2.1.2.3.1.4 Translation memory incorporated

2.1.2.3.2 MT Models

2.1.2.3.2.1 Direct MT

2.1.2.3.2.2 Transfer-based MT

2.1.2.3.2.3 Interlingua-based MT

2.1.2.4 Linguistic resources and utilities

2.1.2.4.1 Languages

2.1.2.4.2 Dictionaries

2.1.2.4.3 Word lists or glossaries

2.1.2.4.4 Corpora

2.1.2.4.5 Grammars

2.1.2.5 Characteristics of process flow

2.1.2.5.1 Translation preparation activities

2.1.2.5.2 Post-translation activities

2.1.2.5.3 Interactive translation activities

2.1.2.5.4 Dictionary updating

2.1.3 Interoperability

2.1.4 Functionality compliance

2.1.5 Security

2.2 Reliability

2.2.1 Maturity

2.2.2 Fault tolerance

2.2.3 Crashing frequency

2.2.4 Recoverability

2.2.5 Reliability compliance

2.3 Usability

2.3.1 Understandability

2.3.2 Learnability

2.3.3 Operability

2.3.3.1 Process management

2.3.4 Documentation

2.3.5 Attractiveness

2.3.6 Usability compliance

2.4 Efficiency

2.4.1 Time behaviour

2.4.1.1 Overall Production Time

2.4.1.2 Pre-processing time

2.4.1.3 Input to Output Translation Speed

2.4.1.4 Post-processing time

2.4.1.4.1 Post-editing time

2.4.1.4.2 Code set conversion (post-processing)

2.4.1.4.3 Update time

2.4.2 Resource utilisation

2.4.2.1 Memory usage

2.4.2.2 Lexicon size

2.4.2.3 Intermediate file clean-up

2.4.2.4 Program size

2.5 Maintainability

2.5.1 Analysability

2.5.2 Changeability

2.5.2.1 Ease of upgrading multilingual aspects

2.5.2.2 Improvability

2.5.2.3 Ease of dictionary update

2.5.2.4 Ease of modifying grammar rules

2.5.2.5 Ease of importing data

2.5.3 Stability

2.5.4 Testability

2.5.5 Maintainability compliance

2.6 Portability

2.6.1 Adaptability

2.6.2 Installability

2.6.3 Portability compliance

2.6.4 Replaceability

2.6.5 Co-existence

2.7 Cost

2.7.1 Introduction cost

2.7.2 Maintenance cost

2.7.3 Other costs

Justification: correspondence between context and quality characteristics

Group members:

Selected scenario (1 or 2):
(write here supplementary specifications, if any)

Characteristics of your context of use	Relevant quality characteristics with their importance (1-3)	Notes
1.2.1.3	*2.4.1.3 (3), 2.1.1.1 (3), 2.1.1.3.3 (2)*	*This is only an example.*
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	