LREC 2008 Tutorial

Evaluating Machine Translation in Use: from Theory to Practice

May 26, 2008 – Palais des Congrès Mansour Eddahbi – Marrakech, Morocco

Paula Estrella	Tony Hartley	Maghi King	Andrei Popescu-Belis
ETI/TIM/ISSCO University of Geneva Switzerland	Centre for Translation Studies University of Leeds United Kingdom	ETI/TIM/ISSCO University of Geneva Switzerland	IDIAP Research Institute, Martigny Switzerland
paula.estrella@issco.unige.ch	a.hartley@leeds.ac.uk	maghi.king@gmail.com	andrei.popescu-belis@idiap.ch

Contents

Objective of the Tutorial

This tutorial offers an introduction to the field of Machine Translation evaluation, and in particular to FEMTI, the Framework for the Evaluation of Machine Translation in ISLE, which groups together a wide range of evaluation metrics, following a contextual evaluation approach. In a practical application of the framework, participants will be shown how to apply FEMTI to an operational example of MT use, in order to construct a well-motivated quality model. The results from the practical exercise will be compared, and a synthesis will be proposed in the end, explaining how feedback from the community can be input into FEMTI.

Outline and Schedule: 9:00–12:30

9:00 – 9:10	Welcome
9:10 – 9:40	Contextual MT Evaluation: Origins, Problems and Concepts
9:40 – 10:10	The FEMTI Guidelines: Principles and Support Interfaces
10:10 – 10:30	Application of FEMTI to an MT Use Case: Description of GPHIN and of the Exercise
10:30 – 11:45	Specification of Context of Use and Quality Model using FEMTI: Hands-on Exercise (suggested break: 11:00 – 11:20)
11:45 – 12:15	Presentation and Discussion of Results
12:15 – 12:30	Overall Discussion and Conclusion

Introduction

The Framework for the Evaluation of Machine Translation in ISLE (FEMTI) is a tool that helps evaluators of MT systems to define contextualized quality models, by relating the intended context of use of an MT system to the quality model used to evaluate it. First put together by the Evaluation Working Group of the ISLE European/USA project, FEMTI is based on feedback obtained

through several types of workshops, and has evolved in the past years to a web-based open tool for MT evaluation, which has also been used in workshops with hands-on exercises.

The FEMTI web-based tool – publicly available at http://www.issco.unige.ch/femti/ – allows evaluators to specify an intended context of use in terms of tasks, input data and users; the tool then automatically suggests a list of quality characteristics (i.e. a *quality model*) that should be evaluated. This automatic suggestion is based on a *Generic Contextual Quality Model* which is constructed by experts in MT evaluation and enriched through subsequent use. The exercise proposed in this tutorial will also provide input to FEMTI's GCQM.

In addition, although many MT developers and users were involved in the creation and improvement of FEMTI, an effort must still be made to extend its use towards corporate and individual users of MT, who often need guidelines and inspiration for establishing evaluation criteria. To reduce the time needed to setup an evaluation, and to increase the completeness and applicability of FEMTI, it is now useful to define also a list of typical evaluation plans, i.e. typical scenarios of use accompanied by typical quality characteristics that should be evaluated in those cases, and possibly the most frequently used metrics. This tutorial concentrates therefore on a real life use scenario, considering what elements would go into a quality model for the MT systems answering the respective user needs, and on how a particular evaluation designed for this scenario would be represented in FEMTI.

Description of Practical Application (Exercise)

The objective of the exercise is to define a contextualized evaluation plan for an MT system that would answer the user needs described in the scenario below. The plans defined by various groups will be compared and integrated, in order to improve the Generic Contextual Quality Model currently available in FEMTI.

Resources: FEMTI's *context characteristics* ("Part I") are outlined on page 6 of this document, and *quality characteristics* ("Part II") are outlined on pages 7 and 8. *Full versions* of Part I and II are available at the tutorial: (a) in print; (b) on the presenters' laptop; (c) via Internet (*if available at LREC*) at http://www.issco.unige.ch/femti. Please use the attached *form* to write down the answers to the exercise, using index numbers of the characteristics to refer to them on the form (rather than their full names). The exercise can be done individually or in small groups.

1. Study the GPHIN scenario of use for MT systems (also explained by the organizers).
2. Describe the intended context of use of an MT system for the GPHIN network by selecting several context characteristics from the list in FEMTI Part I. Additionally, you can also rank these characteristics from the most to the least important.
3. Based on the context characteristics selected in the previous step and using your own experience, select from FEMTI Part II a list of relevant quality characteristics that the MT system under evaluation should possess. This list constitutes the contextualized quality model to be used for evaluation.
4. For each context characteristic, indicate the important quality characteristics that should be evaluated – in other words, answer the question "what quality characteristics correspond to each of the selected context characteristics?" In addition, quantify their importance on a 3-point scale (3: very important; 2: important, 1: nice to have). Write down the results on the attached form, following the example provided in the first line.

5. To complete the quality model, you can also choose metrics (from FEMTI or not) to measure the quality characteristics that you selected.
6. When you have finished defining your contextualized quality model (context and quality characteristics), please hand your form to the presenters, who will integrate all results in preparation for the general discussion.

Questions for the general discussion

1. What constraints and demands does the use scenario place on the MT systems to be incorporated into it?

2. Do these constraints and demands relate, directly or indirectly, to the quality characteristics set out in the FEMTI framework?

3. What metrics might be appropriate to assessing a system's suitability, taking into account the constraints and demands sketched out as an answer to the first two questions?

Scenario of Use of an MT System: GPHIN[1]

The Global Public Health Intelligence Network (GPHIN) is an Internet-based early warning system, which permanently monitors newswires and web sites for information on disease outbreaks and other public health events, and disseminates the information that was selected as relevant nearly in real time.

GPHIN was established by the Public Health Agency of Canada to provide timely and accurate information to the World Health Organization, the European Centre for Disease Control, the US Center for Disease Control, international governments and others whose task it is to react to and manage public health incidents. The information gathered and disseminated by GPHIN supports rapid assessment and response to emerging health risks around the world.

GPHIN monitors information sources in nine languages, using *machine translation* to translate non-English articles into English, and English articles into the other languages of the system. The information is filtered for relevance by an automatic process which is then complemented by human analysis. The output is categorized and made accessible to users. If any item seems potentially to warrant urgent attention, it is immediately forwarded as an alert to users of the network.

Detailed Workflow

1. GPHIN software pulls relevant articles every fifteen minutes from 10'000+ sources of information such as news feed aggregators (e.g. Al Bawaba or Factiva) and other web sites.

2. Selected articles are filtered and categorized into one or more of GPHIN's taxonomy categories: animal diseases; human diseases; plant diseases; other biologics; natural disasters; chemical incidents; radioactive incidents; and unsafe products.

[1] The organisers are deeply grateful to Michael Blench, who provided material for the use scenario and gave permission for its use (Blench, 2007; Mawudeku and Blench, 2005). Any inaccuracies in representing the work of GPHIN are of course the sole responsibility of the organisers.

3. Machine translation is used to translate non-English articles into English and English articles into the other languages.

4. The MT output (called a "gist") is given to an appropriate human analyst. Her task is to ensure that the essence of the article can be understood, not to produce a good translation.

5. Each article is assigned a relevancy score by an automatic procedure. This is supplemented in the middle range of scores by manual analysis and triage. Relevancy scoring may lead to one of three outcomes:
   - An alert is sent by email to the GPHIN end-users including the article, and the article is published to the GPHIN database.
   - The article is published to the GPHIN database, but no alert is given.
   - The article is trashed as irrelevant.

Input Texts

The input texts cover a broad scope: the system tracks disease outbreaks, infectious diseases, contaminated food or water, bio-terrorism and exposure to chemicals, natural disasters and issues relating to the safety of products, drugs and medical devices. Texts are in one of nine languages: Arabic, Chinese (simplified and traditional), English, Farsi, French, Russian, Portuguese, and Spanish.

The texts are harvested from news wires and web sites. They may differ from standard conventional prose in the same language in one or more of the following ways: deliberate or accidental misuse of terms; misspellings; use of local vocabulary rather than standard vocabulary; style of prose; presence or absence of diacritics; use of poetic licence or polysemic terms; use of abbreviations; spacing of the letters of a word; use of capitalization; and Chinese grammar rules.

Users of GPHIN

Several classes of users of the MT output may be distinguished. The human analysts involved in the GPHIN workflow described above are multi-ethnic, multi-cultural, multi-lingual, and multi-disciplined; they work together in close synergy. The end users are those in the organizations served by GPHIN, but no general observations can be made about them. The people involved in the development and maintenance of the network and of the MT components form a third class of users; they also extend and to refine taxonomies and lexica, including those used for MT.

Evaluation of the GPHIN Network

GPHIN has its own set of criteria for the *evaluation of the system as a whole*:

- Usefulness (specificity): the value of the reports selected and disseminated
- Timeliness: the speed with which reports are made available
- Sensitivity: the relevancy of the reports
- Flexibility: ease of making modifications
- Stability (robustness): downtime, staffing
- Cost: sustainability

Evaluation of the MT Component

The MT component is based on the use of the "best of the breed": GPHIN constantly monitors MT developments to ensure that the best systems for individual language pairs are incorporated into the system. At the time of writing, six separate MT engines are being used. The use of MT engines produced by different manufacturers raises some integration issues, such as:

- Instability
- Crashes
- Unpredictable performance
- Poor documentation
- Awkward APIs
- Lack of standards across products
- Bugs
- Clashes between different manufacturers
- Memory leaks

Some of these issues are resolved by the use of an in-house piece of software called nTranslator™, which normalizes APIs; detects engine crashes, freezes and re-boots components; overcomes incompatibilities; solves display problems; and converts file formats.

This tutorial is concerned with the design of an evaluation *for the MT components of the GPHIN system*, using the FEMTI framework as a guide for designing the evaluation, by capturing the particularities of an evaluation in this specific context.

References

Blench, M. "Global Public Health Intelligence Network (GPHIN)". *MTSummit XI*, Copenhagen, Denmark. Available at http://www.mt-archive.info/MTS-2007-Blench.pdf

Mawudeku, A. & Blench, M.: Global Public Health Intelligence Network (GPHIN). *MT Summit X*, Phuket, Thailand, September 2005, invited paper; pp.i-7-11. Available at http://www.mt-archive.info/MTS-2005-Mawudeku.pdf. Presentation available at http://www.mt-archive.info/MTS-2005-Blench.pdf

Heymann, D.L, Rodier, G.R, et al: Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases. *The Lancet Infectious Diseases*, Volume 1, Number 5, 1 Dec. 2001, pp. 345-353.

Estrella P., Popescu-Belis A. & Underwood N. "Finding the System that Suits you Best: Towards the Normalization of MT Evaluation". *27th International Conference on Translating and the Computer*, ASLIB, 24-25 November 2005, London.

Hovy E. H., King M. and Popescu-Belis A. "Principles of Context-Based Machine Translation Evaluation". *Machine Translation*, vol. 17, n. 1, pp. 1-33, 2002.

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". *LREC 2006 (Fourth International Conference on Language Resources and Evaluation)*, Genoa, Italy, p.691-696.

Characteristics of the intended context of use: FEMTI Part I

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
    - 1.4.1.2.1 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
  - 1.4.1.4 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

Output Form for the Application Session: **Context vs. Qualities**

Group members: _____

Notes, comments:

Characteristics of your context of use	Relevant quality characteristics with their importance (1, 2 or 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.**
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	
____	_____	