Evaluating Machine Translation in Use: from theory to practice

Tutorial programme

1. Introduction to FEMTI
2. Practical exercise
    - 2.1. Description of the scenario of use of an MT system
    - 2.2. Specification of a related quality model in FEMTI
3. Break
4. Synthesis of results and plenary discussion

Tutorial Organisers

Paula Estrella
ETI/TIM/ISSCO - University of Geneva
40, bd. du Pont-d'Arve
1211 Geneva 4, Switzerland
Tel/Fax: +41 22 379 8680 / +41 22 379 8689
Email: paula.estrella@issco.unige.ch

Anthony Hartley
Centre for Translation Studies
School of Modern Languages and Cultures
University of Leeds
Leeds, LS2 9JT, United Kingdom
Tel: +44 (0)113 343 1608
Email: a.hartley@leeds.ac.uk

Maghi King
ETI/TIM/ISSCO,
University of Geneva
40, bd. du Pont-d'Arve
1211 Geneva 4, Switzerland
Tel/Fax: +41 22 379 8680 / +41 22 379 8689
Email: maghi.king@gmail.com

Andrei Popescu-Belis
IDIAP Research Institute
Av. des Prés-Beudin 20 - Case Postale 592
CH-1920 Martigny, Switzerland
Tel./Fax +41 27 721 7729 / +41 27 721 7712
Email: andrei.popescu-belis@idiap.ch

Table of Contents

Prolegomenon

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 context of use of an MT system to the quality model used to evaluate it. First defined by the Evaluation Working Group of the ISLE European/NSF (USA) project, FEMTI is in reality 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 been used in workshops including hands-on exercises in MT evaluation.

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, and would benefit from the use of a normalized reference such as FEMTI. 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 on a real life use scenario, considering what elements would go into a quality model for the MT systems embedded within that scenario, what metrics might be used to assess a system's performance relative to the various components of the quality model, and on how a particular evaluation designed for this scenario would be represented in FEMTI.

The organisers are deeply grateful to Michael Blench, who provided material for the use scenario and gave permission for its use. Any inaccuracies in representing the work of GPHIN are of course the sole responsibility of the organisers.

Participants in the tutorial are requested to read the material below before they come to the tutorial, and to reflect on possible answers to the questions at the end.

The use scenario: in summary

The goal

The goal of the Global Public Health Intelligence Network (GPHIN), established by the Public Health Agency of Canada, is to provide timely and accurate information to the World Health Organization, the European Centre for Disease Control, the 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 should support rapid assessment and response to emerging health risks around the world.

The network (GPHIN)

GPHIN is an Internet based early warning system. It monitors news wires and web sites for information on disease outbreaks and other public health events. It functions 24 hours a day, 7 days a week. Dissemination of the information selected as relevant is done in close to real time.

Monitoring is done in nine languages. Machine translation is used 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.

The use scenario: more detail

The 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.

They are in one of nine languages: Arabic, Chinese (simplified), Chinese (traditional), English, Farsi, French, Russian, Portuguese, 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
- Mis-spellings
- 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
- Chinese grammar rules

Output from the system

Processed articles deemed to be relevant are published electronically in the GPHIN database. Those requiring urgent attention are forwarded by e-mail to the end-users of the system.

Users of the MT components

Two classes of user of the may be distinguished:

- The human analysts involved in the GPHIN workflow (more below). They are
    - Multi-ethnic
    - Multi-cultural
    - Multi-lingual
    - Multi-disciplined
    - Work in close synergy

- The end users in the organizations served by GPHIN. No general observations can be made about these users.

There are also users involved in development and maintenance of the network and of the MT components. In particular, GPHIN experts extend and refine taxonomies and lexica, including the lexica used for machine translation.

Workflow

- GPHIN software pulls relevant articles every fifteen minutes from news feed aggregators (Al Bawaba and Factiva) and monitors a large number of web sites. There are currently 10'000+ sources of information

- 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
    - Unsafe products

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

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

- 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 to the GPHIN end users including the article, and the article is published to the GPHIN data base.
    - The article is published to the GPHIN data base, but no alert is given
    - The article is trashed as irrelevant

Evaluation of the GPHIN network

GPHIN has its own set of criteria for 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, 6 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 API’s
- 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 software called nTranslator^{TM}, which:

- Normalizes API’s
- Detects engine crashes, freezes and re-boots
- Overcomes incompatibilities
- Solves display problems
- Converts file formats

This tutorial is concerned with the design of an evaluation for the MT components of the system, with using the FEMTI framework as a support and guide for designing the evaluation and with using the FEMTI framework to capture the particularities of an evaluation in this specific context in order to make design decisions available to a wider public. The FEMTI framework is publicly available at http://www.issco.unige.ch/femti/

Preparatory work

Before coming to the tutorial, participants are asked to reflect on their answers to the following questions:

- Q1: what constraints and demands does the use scenario set out above place on the MT systems to be incorporated into it?

- Q2: do these constraints and demands relate, directly or indirectly, to the quality characteristics set out in the FEMTI framework?

- Q3: 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?

Bibliographic references:

Blench, M., 2007: 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

Further reading:

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

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 December 2001, pp. 345-353.

Hovy E. H., King M. and Popescu-Belis A. (2002). “Principles of Context-Based Machine Translation Evaluation”. Machine Translation, vol. 17, n° 1, pp. 1-33.

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