The place of automatic evaluation metrics in external quality models for machine translation

Andrei Popescu-Belis

ISSCO / TIM / ETI

University of Geneva

Workshop on Automatic Procedures in MT Evaluation

MT Summit XI, Copenhagen, 11 September 2007

What is translation evaluation? ☺

- Given
  - a sentence S_n in a source language
  - a sentence T_n in a target language

- Determine
  - a score $\mathbf{s}(S_n , T_n)$ such as
    - $\mathbf{s}$ = 1 iff T_n is a <u>perfect</u> translation of S_n
    - $\mathbf{s}$ = 0 iff T_n is <u>clearly not</u> a translation of S_n
    - $\mathbf{s}(S_n , T_n) > \mathbf{s}(S_n , T_k)$ iff
      T_n is a <u>better</u> translation of S_n than T_k

Issues and answers

- What does "better translation" mean?
  - go and ask people (= language users)
- Could **s** be computed automatically, directly from S_n and T_n?
  - *but this is also the goal of MT!*
  - so, could **s** be approximated? with what supplementary knowledge?
- A consistently high **s** is not the <u>only</u> desirable property of an MT system
  - → FEMTI

Plan

- A principled view of MT evaluation: FEMTI
  - quality models: characteristics, attributes, metrics
- Two types of justifications for automatic MT evaluation metrics
  - structural reasons (“glass-box”)
  - empirical reasons (“black-box”)
- Empirical distance-based metrics
  - arguments for or against them
- Task-based evaluation
  - proposal for automatic task-based evaluation

Principled view of MT evaluation: FEMTI

- FEMTI: Framework for the evaluation of MT, started within the ISLE project

  http://www.issco.unige.ch/femti

- Two classifications / surveys
  - characteristics of the context of use
  - quality characteristics and metrics

- Helps to define evaluation plans
  - support interfaces: specify context of use, then generate contextualized quality model

Important ISO-inspired notions

- ISO/IEC 9126 and 14598, SQUARE framework
- Quality
  - "the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs" (ISO/IEC 9126)
  - decomposed into quality characteristics, then into measurable attributes, each with internal/external metrics
  - six categories of quality characteristics: functionality, reliability, usability, efficiency, maintainability, portability
- Metric
  - "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 14598)

FEMTI refinement of ISO quality characteristics for MT (Hovy, King & Popescu-Belis, 2002)

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
    - 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 TM 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*

FEMTI refinement of ISO quality characteristics for MT (Hovy, King & Popescu-Belis, 2002)

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 Tr. Speed
  - 2.4.1.4 Post-processing time
    - 2.4.1.4.1 Post-editing time
    - 2.4.1.4.2 Code set conversion
    - 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 (Introduction, Maintenance, Other)

Examples of metrics from FEMTI

- For <2.1.1.2 Fidelity>
  - assessment of the correctness of the information transferred by human judges
- For <2.4.1.3 Input to Output Translation Speed>
  - number of translated words per unit of time
- For <2.1.3.2 Punctuation errors>
  - percentage of correct punctuation marks
- For <2.5.2.3 Ease of dictionary update>
  - time OR effort necessary to update dictionary

- Some metrics require human judges that cannot be replaced with software (#1 above)
- Some metrics can be applied both by human judges or software (#2), but software is more precise & cheaper
- Some require human judges or complex software (#3)
- Some metrics require human users of the system (#4)

This workshop: “Automatic procedures in MT evaluation”

- Underlying assumption: look only at automatic metrics for the quality of MT output such as BLEU, WER, etc.

➔ FEMTI Part II, under <2.1 Functionality>

- current metrics require human judges
- could they all be automated? No obvious solutions!

Place of automatic metrics in FEMTI

- Do automatic metrics which were independently proposed belong in FEMTI? Where?

- If a function **s**(S , T) : SL x TL → [0; 1] is to be called a quality metric, one should indicate what quality it measures
  - it must be possible to integrate this (external) quality into the ISO/FEMTI classification, most likely under <Functionality>, if not present yet

Two types of justifications for automatic MT evaluation metrics (1/2)

- Structural = "glass-box"
  - the definition of the score **s** indicates that it measures the same quality attribute as a recognized metric applied by humans

  → hence place **s** in FEMTI under the same quality attribute

- An infrequent justification…

Two types of justifications for automatic MT evaluation metrics (2/2)

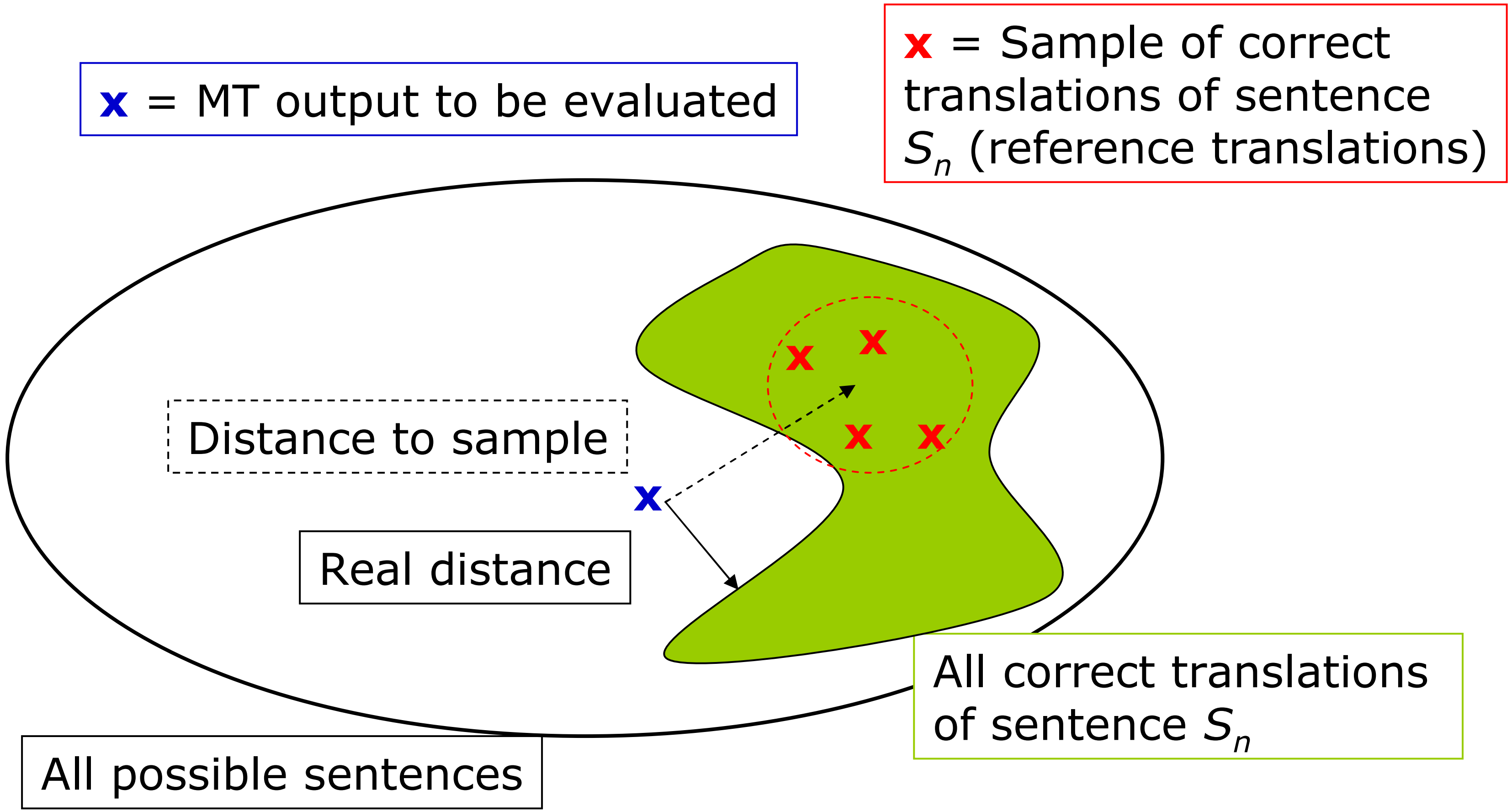
- Empirical (and frequent) justification = "black-box"
  - the values of score **s** on a given test set are statistically correlated with a recognized metric applied by human judges → assume that the two metrics measure the same quality
- Reverse engineering: how to construct such a score **s**?
  - start with a set of MT sentences that are already scored by humans according to a metric $\mathbf{s_h}$, i.e. start with a large set of triples (S_n, T_n, $\mathbf{s_h}(n)$)
  - train a statistical model to approximate $\mathbf{s_h}$ and then estimate its error using cross-validation → new automatic metric!
- But this is the same problem as statistical MT! ($\mathbf{s_h}$ = 1)
  - too difficult… → need to use supplementary information about *correct translation(s) of the evaluation data set*

Trainable distance-based metrics

- Distance-based NLP evaluation
  - the evaluation data set (test set) contains desired output associated to the input data
  - evaluation metrics are defined as distances between a system's output and the desired output, averaged over all items of input data

- Situation for MT
  - no unique desired output for an input sentence
  - frequent proposal: compute a distance between a system's output and a sample of correct outputs (often up to 4)
  - replace score $\mathbf{s}(S_n , T_n)$ with $\mathbf{d}(\{T_{ref(1)}, \ldots, T_{ref(4)}\}, T_n)$

Graphical representation



Training automatic metrics

- How to construct a distance-based automatic metric **d**?
  - start with a set of machine-translated sentences (T_n) that are already scored by humans according to a metric $\mathbf{s_h}$
  - each source sentence is accompanied by reference translation(s)
  - i.e. start with a large set of t-uples ($\{T_{ref(1)}, \ldots, T_{ref(k)}\}, T_n, \mathbf{s_h}(n)$)

- Find a distance **d** that approximates $\mathbf{s_h}$
  - that is, $\mathbf{d}(\{T_{ref(1)}, \ldots, T_{ref(k)}\}, T_n) \approx \mathbf{s_h}(n)$

- Essential point: role of (machine) learning
  - either the statistical model **d** was explicitly trained to approximate $\mathbf{s_h}$
  - or several distances $\mathbf{d}_i$ were tried & the one closest to $\mathbf{s_h}$ was selected
  - in both cases, error of the model was estimated using cross-validation

Advantages and drawbacks of trainable (empirical) distance-based metrics

- Advantages
  - low application cost
  - high speed
  - reproducible (*vs*. human judges who may vary)

- Drawbacks
  - correlation with reference (human) metric holds mainly for data that is similar to the training (or validation data)
    → unknown behavior for different (unseen) types of data
  - unclear/variable correlation with ISO-style qualities
  - need training data (which may have imperfect inter-judge agreement)

An alternative: task-based evaluation

- Measure utility of MT output for a given task
  - e.g. performance of human subjects on a task using human *vs*. machine-translated text
  - closer to ISO's quality in use
  - increasingly popular as limits of BLEU become visible

\+ OK if system intended for specific application

– Expensive, time-consuming

- Idea
  - <u>automatic task-based evaluation</u>
  - use MT output for another NLP module for which good automatic metrics are available
    - e.g. reference resolution, document retrieval

Conclusions: two views of the future

- Utilitarian view
  - a "better" system means only "better adapted to the users who wish to pay for it" – no absolute metrics
  - task-based metrics do work, and could be automated
  - but could this really be the whole story?

- Cognitive view
  - why did the quest for MT evaluation metrics become just another NLP problem?
    - with machine learning techniques, annotated data, etc.
  - the invariants of translation aren't well understood
    - good candidates for ground truth
    - components of meaning: logical form, inferences