An Introduction to MT Evaluation

As Many Questions as Answers

Florence Reeder

The MITRE Corporation

19 April 2001

Agenda

- What MT Evaluation is
- Why MT Evaluation is necessary
- Why MT Evaluation is hard
- Who’s done what to whom
- Current trends in MTE
- Slings and arrows

What is MT Evaluation….

- Measuring usefulness, success, efficacy of software component that translates between two human languages (Dorr, et al. 1999)
- "What the user needs [is] the ability to characterize his or her particular needs (personal and organizational), and the ability to compare this characterization with the performance characteristics of various MT engines." (Hovy, 1999)
- "I find it very hard to talk into empty space about what counts as a good or a bad translation. I need to know what it's for and what the criteria are in that particular situation before I can even *talk* about evaluating a translation." (King, 1994)
- A demonstration of the feasibility of applying a computer to an activity (White, 2000)

What is MT Evaluation, cont’d.

- “Even today, so-called evaluations of MT technology (using ‘evaluations’ in the loose sense of the word) range from assertions that MT is an intractable problem to claims of upwards of 90% accuracy for systems, without a clear specification of what “accuracy” entails.” (Miller, 2000)
- One major disadvantage of quality assessment for MT evaluation purposes, however, is the fact the overall performance of an MT system has to be judged on more aspects than translation quality only. (Arnold, et al, 1994)
- “It has been a cliché in the field for years that machine translation evaluation is a better founded subject than machine translation.” (Wilks, 1994)

Why MT Evaluation is Hard

- No gold standard
- Wide range of parameters on which to evaluate
  - Not all have same importance to every user
    - What is acceptable to a user?
- Wide range of uses of product
  - Not all are as tolerant of failure
- MT is hard

This is Engineering?

- Comparative analysis between one or more systems (horizontal)
- Comparative analysis between one or more versions of system (vertical)

When I look at an article in Russian, I say: "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode." (W. Weaver, 1947)

Maybe it’s magic….

- No single right answer for a translation
  - Even when humans do it
  - Is a 40% solution good enough sometimes?
- Different users of evaluation have different needs from evaluation
  - Developers, MT users, money people
- MT users have different expectations
  - “Star Trek is reality” versus “Can’t do it”

Types of Evaluation (White, 1998)

- Feasibility of MT system / paradigm
- Internal evaluation of system function
- Declarative evaluation of product
- Usability evaluation
- Operational (financial or process)
- Comparison

Who’s done what to whom

- Evaluating MT systems as SYSTEMS
- Black-box evaluations
  - Measure accuracy of input/output pairs
    - Fidelity, intelligibility
- Glass-box evaluations
  - Measure data flow / architecture / methodology
- ALPAC - scales of speed, cost, quality

Evaluate MT Systems as Systems

- An important, but ignored factor
  - Coffee cup timing
  - If it crashes more than 3 times, I won’t use it.
  - I can’t copy a web page into the buffer without crashing the program.
  - What code set does this take?
- How do software standards need to be tailored to this type of software?

User Interface Issues

- Is it intuitive?
- Is it consistent?
- Does it support the intended use?
- Can it handle multiple interaction types?
  - Client/server versus stand-alone
  - On-demand versus real-time

Input Issues

- What kinds of pre-processing must be done?
  - Code set conversion / Spell checking
  - Format conversion
- What size of data chunks does it work on?
  - Sentence, paragraph, web-page?
- Does it have specialized conventions?
  - File naming, dictionary location, etc.

Output Issues

- Does it mark…
  - Words that don’t translate?
  - Translation unit boundaries?
- Does it render output...
  - In a specific code set (sometimes internal)?
- Does it interleave source and target?
- Does it generate lots of intermediate data?

Generalized Software Issues

- What is the mean-time-between-failures?
  - Does it degrade gracefully ?
- How quickly does it load lexicons?
- How quickly does it translate?
- How easy is it to install / upgrade?
  - Is it extensible to new domains?
- Will someone answer at the help desk?

Black-Box Evaluations

- Look at the output and rate it
- Back-translations
- Compare to language learners, translators
- “Helicopters in Vietnam”
- Categorize translation needs by input type or output use
- Process and cost in process

Look at the output and rate it

ÇáÓíÏ ÑÆíÓ ÝÑíÞ 25/89 ?ã/ÕáÇÍíÉ ÇáãÎÊÈÑ ÈÚÏ ÇáÒíÇÑÉ áãÎÊÈÑÇÊ ÇáÓíÏ ÝÇÖá ÚÈÏ ÑÍãä ¡ äæÏ Ãä äÈíä ãáÇÍÙÇÊäÇ ÃÏäÇå :

700 Scholars and Doctors at Blood Pressure Conference

Sharm al-Shaykh witnessed a major scientific event involving 700 professors of cardiac, kidney, and internal medicine …..

Output Only - Techniques

- Human rating of output - is this good?
- Rating fluency, comprehensibility, fidelity, post-editing needed
- Error analysis of output only
  - Categorize errors into classes - different ones effect use of system differently
- Edit distance measure between translations
  - Particularly good for vertical comparison

Output only analysis - issues

- Frequently subjective measures
- Need target language speakers and potentially bilinguals and domain experts
- Human intensive
    - Human factors problems of evaluation
- Does not measure fidelity
- Error analysis very hard to do - can’t just count number of “wrong” words

Back-translation

- Source1 → Target → Source2
- Compare source results with original
- Measure divergence of Source1 and Source2
- Need both sides of translation process
  - Source → target and target → source
- Comical divergences & Pathological case
- String substitution will win

Helicopters in Vietnam

- Translate helicopter maintenance manuals from English to Vietnamese
  - Some manually, some with MT
- Wait a while
- See which helicopters crash
- Apocryphal example (but objective and operational)
- Even MTE has its urban legends

Comparison to Humans

- Comparisons to human translator evaluation
- Comparisons to language learners
  - Cloze tests
  - Multiple choice reading comprehension tests
- Machines are not humans
  - Some things machines are good at and some not
  - Different kinds of errors
  - Humans have variations

Categorize translation - input

- Measure success of system as function of quality of input
  - Structure of input (formal, informal, technical)
  - Grammatical types
- Results in translatability index
- Measure success as function of language divergences (Arabic → Thai)

Categorize translation - output

- If using MT for particular task, measure success of using MT output in that task
  - Assimilation / Dissemination / Conversation
  - Filtering / Routing / Analysis / Gisting
    - MT Proficiency scale
- Measure success of humans performing tasks with output as compared to other systems or human translation
  - TOEFL experiments

Categorizing input / output

- Now have two language problems to solve
- Must re-do tests for new input type or output use
- Can be very human intensive and resource expensive
  - Finding and preparing corpora

Cost of use in process

- Measure cost of process with and without MT
- May not capture personnel availability
- Costs to factor in
  - Maintenance of system / lexicons
  - Conversion of materials to appropriate format
    - Errors introduced by each stage in process
    - Cascading errors without apparent cause

Glass-box techniques

- Being able to look inside the system figures out if success is a side-effect or a feature
- Correspondence models
- Test suites based on linguistic models

Correspondence Model

- Describe syntactic and semantic relations
- Categorize according to divergences
- Measure correspondence of models to translation pyramid
- Does not measure if good enough
- Does not measure all types of good enough
  - (Ahrenberg & Merkel, 2000)

A New Wave of Old Guard MTE

- Revisiting ghosts of the past
  - Reasons MTE failed then may not apply now
    - Corpora availability and processing power have increased
    - Expectations have changed
- Looking at the overlap between MT and other NLP fields
  - They've learned something over the years too

The New Wave - continued

- Embedded part of bigger process
  - Effect of MT as components change
  - Evaluate only parts as needed by downstream processing (name translation for IE)
  - Measuring effects of each stage
- Push towards full automation
  - Reduce amount of human effort necessary
  - Building corpora to represent range of HT/MT

A Quick Recipe (King, 1999)

- Why is the evaluation being done?
- Elaborate a task model.
- Define top level quality characteristics.
- Produce detailed requirements for the object under evaluation, using the information gained in the previous steps as a basis.
- Define the metrics to be applied to the system for the requirements produced under 4.
- Design the execution of the evaluation
- Execute the evaluation

Slings and Arrows

- Thank you for suffering what is probably review for all of you.
- On to the fun stuff……
- Questions?