Evaluating an MT system without knowledge of the source language

Donna M. Gates

Josemina Magdalen

Keith Miller

Nancy L. Underwood

Evaluation Design

- ISO/EAGLES

  1. Why is the evaluation being done?
  2. Elaborate a task model

- ISLE Taxonomy

- Declarative Evaluation of an MT system to be used for gisting with unknown source languages.

- Black box evaluation

- Should the customer acquire the system?

- Scenario: a librarian retrieving texts in an unknown language.

User Requirements

- Translation Task: Assimilation
- User characteristics:
  - No knowledge of source language
  - Little or no linguistics education
  - Ideally native proficiency in target language
- Input characteristics:
  - Chemical warfare treaty
  - Nothing can be assumed about the author

- System: S1
  - Chinese => English (evaluated against HT)
  - French =>English
  - Spanish => English

- System characteristics to be evaluated: not concerned with internal characteristics unless they influence external behavior (e.g., automatic learning algorithms)

Data

- Chem corpus:
  - Chemical weapons treaty
  - Zh, En, (Fr, Sp)
- Broken into sections (automated):
  - approximately 150 lines per section
  - 40 sections
- Chose 4 sections for «testing the tests »
  - determine feasibility and applicability of metrics

ISLE Characteristics to be measured

<u>Comprehensibility:</u>

- Is the text understandable?
- Metrics:
  - Cloze test
  - Subjective judgement per sentence (0/1)

- Readability (clarity?) :
  - Ease of reading text.
  - Metric: timing readers.
- Fidelity:
  - Most important characteristic
  - Metric: Subjective 4 point scale for each sentence (averaged)

- Coverage:
  - Corpus based problems
  - Cross-language phenomena unknown.
  - Metric: % of translated words.
- Terminology:
  - Identify terms in gold standard text
  - Metric: % of translated *terms*
- Utility of output: acid test not possible here.

Ordering of tests important when carried out by the same evaluators

Results

- Comprehensibility:
  - 0/1 test -

    text1&2 44/117 (Donna) 37.6%

    46/102 (Nancy) 45.1%

    text3 38/57 (Josemina) 66.7%

    text4 27/54 (Josemina) 50%

  - Cloze test -tbd later

Results cont.

Coverage :

text 1 total words 912
untranslated 22
coverage 97.6%

text2 total words 794
untranslated 64
coverage 91.9%

text3 total words 1210
untranslated 18
coverage 98.5%

text4 total words 1153
untranslated 28
coverage 97.6%

- Terminology: (see examples)

General observation: Verbal forms were translated into NPs

signatory state = State Party ;
accumulation destruction = cumulative destruction ;
1st kind = Category 1 ;
completes destruction = completion of destruction ;
lengthens = extention ;
installation = facility ;
technology secretariat = Technical Secretariat ;
destruction time = destruction period ;
joint pledge = Convention ;

carries out council = Executive Council ;
proposal extends long-term = The duration of the proposed extension;
chemical weapon destroys = chemical weapons destruction ;

Future Work

- Finish vetting the tests;
- Carry out the tests;
- Finding a correlation between objective and subjective measures;
- Investigate difference between the evaluation for known vs. unknown SLs;
- Find a correlation between individual measures and task performance (integrating the MT into a whole workflow)
- **Hope to find automated scoring correlations for subjective scoring**
- Feedback results to ISLE taxonomy

Thank you for your attention.

comments

- Hans Caldrin? Time to read text out loud.
- Distinguish not knowing source language vs not knowing "about" SL
- Cloze test done on other sections
- unique tokens for coverage: should we leave repeated words?
- Segmentations? TM uses paragraph for segmentation. 1 or more characters for a word.

More comments

- JM: problem: vps not found in HT
    - consistant translations
    - Cloze test may be very difficult
    - difficult to match the terms
- Spanish and French may be easier
- proper names, longer phrases as terms risks bringing syntax into terminology seperate syntactic from terminilogical.
- MK:working in windows in the text
- KM: Ngrams in HT vs MT
- Flo: mutual information collocations in different size windows,

More comments

- Anna: fidelity? How did we measure.
  - 0 nothing, 3 = all info, 1 < 50%, 2 > 50%
- MK: What counts as information?
  - What do you do with the content?
  - Influence how you feel about output.