Human MT Metrics

What's so great about peace, love and understanding

Flo Reeder

27 May 2002

Two Types of Human Metrics

- Look at text and judge –
  - Fidelity, intelligibility, ability to do task X….
  - Holistic, point scales, DARPA-94 5-point scale
- Try to do human tasks with data
  - Cloze tests
  - Reading comprehension exams
  - Read out loud / typing test

JEIDA (1992)

- Human ratings on steroids
  - For users, questionnaires (nearly 100 questions)
  - For providers, questionnaires (3 separate)
  - For developers, questionnaires
- Questions in 14 categories are rated for category score
- Plot category scores on spider-graph
- Match spider-graphs to find best system for you

Intelligibility / fluency

- Look at text and rate intelligibility or fluency
  - How natural is the target language
- Generally a scale where
  - 1 = totally unintelligible
  - 5 = completely intelligible
- Can be either sentence-based or paragraph based
  - Sentence based may fail to capture discourse phenomena

Clarity

- How clear is the target language?
- Miller & Vanni (2001) - Scale which merges multiple individual features into single assessment (intrinsically)
  - Comprehensibility, readability, style
- In van Slype (1979) – point scales – a component of intelligibility

Adequacy / Fidelity

- How much does target convey the message of source
- Look at pairs and rate on scale (5 or 9 point)
  - Source/target for bilinguals
  - Reference/target for monolinguals
  - Two different assessment types
- Again phrasal, sentence, or text granularities

Informativeness (1)

- Rate ability to do task X with text
  - Still a human rating
  - But getting towards task-based evaluations
- Task Proficiency Scale
  - Hierarchy of downstream human tasks of increasing complexity (from binning to gisting)
- Rating on more specific dimensions of quality

Informativeness (2)

- Use texts as basis for testing
  - Reading comprehension exams
- Translate text
  - Raters take test – better translations yield better scores
- World knowledge problem
- Test item design problems

Reading Time

- Time comprehension test
  - Incorporate time spent taking test into score
- Read out loud
  - Time how long takes to read text out loud
    - From psycholinguistics
- Type it
  - For some folks, it might be typing time – could measure mistakes in typing as well…

Correction / Post-Edit time

- How long does it take to get the text into decent shape
- We have metrics for human post-editing
- Threshold below which humans will not touch data
- One possible automatable metric…

Cloze Test

- Target text with every Nth word removed (5, 8)
- Can participants reconstruct text
- Measure correct / incorrect responses
- Score
- Entropy Cloze test – condition scores by human translation scores

Issues, we got issues

- As much a human factors challenge as an MTE challenge
  - Running order effects, pizza effects
- Humans are not machines and machines are not humans
  - Good at different kinds of texts, good at different kinds of tasks
  - Sometimes end consumer of MT not a human!
- Expense not trivial
  - Hundreds of raters, hundreds of hours

But in the End

- Humans, typically, are the end-users
- Humans ARE the ones who accept the products
- Humans have intuitive knowledge we have yet to capture
- We are just starting to find metrics which correlate well with or even (gasp) mimic human judgment abilities…

NEE Score

anno-05	1	0.07692308
anno-06	1	0.07692308
anno-07	8	0.61538464
anno-01	9	0.6923077
anno-04	9	0.6923077
anno-09	9	0.6923077
anno-08	10	0.7692308
anno-10	10	0.7692308
anno-11	10	0.7692308
anno-12	10	0.7692308
anno-13	10	0.7692308
anno-02	12	0.9230769
anno-ref	13	1.0
anno-03	13	1.0

Typing test – raw times

10A	6:00
101	6:00+-
102	6:02
103	7:03
104	6:23
105	6:01
106	6:47
107	7:06
108	6:56
109	7:35
110	6:35
111	????

About the technique

- Positives:
  - Objective
  - WPM = (WORDS / TIME) – MISTAKE-PENALTY
  - Might be able to incorporate post-edit measures here -> hard to turn off editor
- Negatives:
  - Touch typist, but not a professional one
  - Make common mistakes on perfect text
  - Allow backspacing or no?
  - Editor issues (line-feed)
  - Fingers hurt
  - Could be problem for large scale evaluation