Some questions from the discussant

Andrei Popescu-Belis
ISSCO, University of Geneva
Future challenges for MT evaluation

• Who are the main users of MT evaluation?
• What do they need?

  – *MT developers* want to improve the accuracy of MT output quality

  – *MT users* (human or software) want to improve their productivity using the most suitable MT system
MT developers (1)

• Do they need better (automatic) metrics?

Example: PINK

– get plenty of source texts, human (reference) translations, MT output, human (reference) scores
– train/test PINK to best match human scores
– PINK is 1% better than BLEU at matching human judgments 😊

• but if the agreement of human raters is only about 95%, what is the sense of the 1% improvement?
MT developers (2)

• Tony said:
  – improve the measures that target the preservation of form and of content

• But:
  – should automatic MT evaluation replace MT?

  • if you have a good automatic method to measure semantic similarity, why not use it for MT?
MT developers (3)

- Do we know enough about the behavior of existing metrics of output quality?

- Do we need some kind of common packaging of output quality metrics?
  - resources
  - scoring software (automatic / human interfaces)
  - scale/range
  - reliability assessment
MT users (1): humans

• Tony said:
  – look for instance at post-editing effort

• But:
  – shouldn’t we try to assess first the range of actual (and future) uses of MT?

  ➢ new ideas & funding for MT developers
MT users (2): other software

• Tony said:
  – look at IR+MT, QA+MT, etc.

• But:
  – are there any general results about the combination of performances/flaws in a complex HLT system?

  ➢ e.g. progressive degradation in a pipeline
Summary: do we need more “standardization” of MTEval?

• Improve the reusability of metrics for output quality
  – design self-contained evaluation packages

• Improve the generality of usability studies
  – survey existing uses of MT as a component
  – methodology for reusing user-centric MTEval results
  – define a methodology for the evaluation of multi-component systems that include MT