CENTRE FOR TRANSLATION STUDIES
at the University of Leeds

MT Evaluation – the little we know

Tony Hartley

Centre for Translation Studies

Leeds, UK

a.hartley@leeds.ac.uk

UNIVERSITY OF LEEDS

CENTRE FOR TRANSLATION STUDIES
at the University of Leeds

MT Evaluation – a big space

- Requirements
  - Task: assimilation, dissemination, ...
  - Text: type, provenance, ...
  - User: translators, consumers, ...
- Quality attributes
  - Internal: architecture, resources, ...
  - External: readability, fidelity, well-formedness, ...

Current research vision

Developer's perspective

- Test suites
  - Syntactic coverage, degradation
- *N*-gram metrics
  - BLEU (2001), NIST, WNM/LTV, RED ...
  - One or more reference translations
  - Correlation with human judgments
  - Automatable

The trouble with *n*-grams

- Correlate reliably with human rankings based on *adequacy* and *fluency* across languages
- But ...
  - Over-rate SMT
  - Complex relationship with perceived *acceptability / suitability* of output – re-calibrate for each language pair and text type
  - Poor reference translations improve scores
  - Based on flawed model of translation

Equivalence in HT

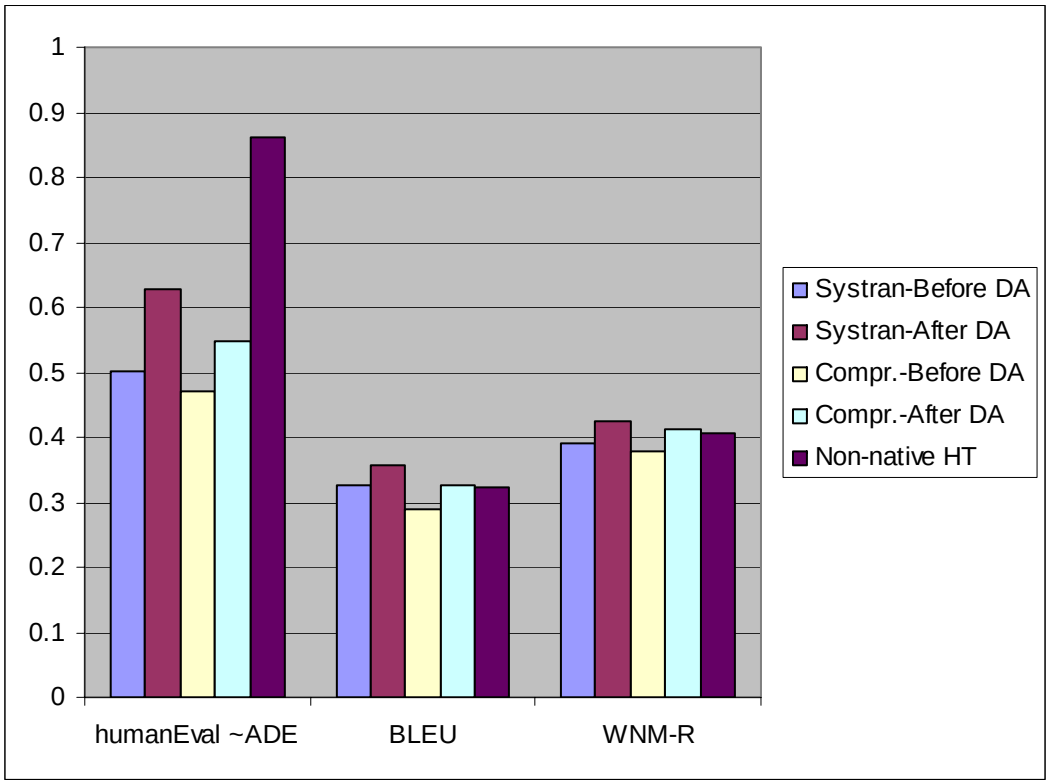
From German

Do not bring infected animal products into the country.

From Italian

Let's keep infected animal products out of the country.

MT is not translation



Where next?

- Beyond similarity metrics
  - FEMTI offers a rich palette of techniques
- Beyond *adequacy* and *fluency*
  - Too generic / abstract for specific tasks?
  - Consider MT output in its own right
- Beyond conventional uses of MT as surrogate human translation (*emulation*)
  - MT as a component in a workflow

Restore a sense of purpose

Texts are meant to be used.

*There are **no absolute standards** of translation quality but **only more or less appropriate translations for the purpose** for which they are intended.*
(Sager 1989: 91)

Revisit MT proficiency (White 2000)

- View MT output as a genre
  - Characterise ***in****adequacy*, ***dis****fluency*, ***ill-****formedness*
- Embed MT and adapt (to) the environment
  - in IE, CLIR, CLQA, Speech2Speech
  - in pre- and post-editing

MT evaluation agenda

- Create common set of multilingual data to assess performance on (generic) tasks with and without MT: IR, IE, QA ...
- Improve metrics that do not rely on *n*-gram similarity, e.g. X-score, D-score
- Challenge of evaluating (mutually) embedded HLTs
  - NER and/or WSD within MT
  - IE before MT
  - Generation of multiple solutions