



MT Evaluation



Workshop at LREC 2002
Conference
**Machine Translation Evaluation:
Human Evaluators Meet
Automated Metrics**

Mairead McCarthy, Mary Hearne,
Nano Gough, Andy Way
{mccarthy, away}@computing.dcu.ie



MT Evaluation



- Not a true MT evaluation exercise
 - assume reference translation is correct (both texts contain errors)
 - evaluation of systems' ability to produce paraphrases (comparing English to English)...really need to look at the source text not just the reference translation?



MT Evaluation



Some frequent (expected) problems

- Determiners – jumbled (him/her/it – can't decide between); inserts & deletes where it shouldn't
- Some systems have alternative translations in () – not always a useful facility
- De (of/from/what preposition?)



MT Evaluation



- Exercise is a bit false/artificial/postmodified?
- Alter punctuation (it is a sensible thing to do – reads better but still!)
- Inserts words – adding style to the translation



MT Evaluation



Examples of Real Translations??

- Taliban texts – no.8 & 9 so similar – can they be different systems? – only a few words in the difference.
- Children & Drugs – no.4 combines 3 sentences into 1 – style of which is too good.



MT Evaluation



Lack of consistency throughout text

- Some phrases are really good, others bad.
- Children & Drugs – no.3 word “addcition” & “addiction” appear – unlike a machine translating.
- More than 1 lexical entry in the translation dictionary –first/highest probability should always be selected? Sometimes an odd translation is chosen.



MT Evaluation



Questions

- What MT systems were used?
- Were the outputs post-edited?
- Is this really a valid exercise?