Scoring Techniques

Black-box: Judging hypotheses produced by MT systems:

- Adequacy and fluency of whole sentences.
  Somewhat revisited under the name Direct assessment (DA).
- Relative ranking (RR) of full sentences by several MT systems:
  Longer sentences hard to rank. Candidates incomparably poor.
- Ranking of constituents, i.e. parts of sentences:
  Tackles the issue of long sentences. Does not evaluate overall coherence.
- Comprehension test: Blind editing+correctness check.
- Task-based: Does MT output help as much as the original?
  Do I dress appropriately given a translated weather forecast?

Gray-box: Analyzing errors in systems' output.

- HMEANT, HUME: Is the core event structure preserved?
- MQM: Multi-dimensional quality metrics.

Glass-box: System-dependent: Does this component work?