contentquo

MT QUALITY EVALUTION METHODS

COMPARATIVE MAP

AUTOMATIC METRICS

TRADITIONAL QUALITY METRICS (BLEU, ETC.)

(++) free
(++) very fast
(--) need reference translations
(--) poor match with human judgement
(-) some are less useful for NMT

QUALITY ESTIMATION (AI)

(+) free or cheap on a per-word basis
(+) reasonably fast
(+) no reference translations needed
(+) better match with human judgement
(-) complex to apply, few ready tools

EDIT DISTANCE (PEMT)

(++) free
(++) very fast
(--) not available for Raw MT
(-) not self-sufficient, need to do human investigation on outliers

HUMAN EVAL (HOLISTIC)

DOCUMENT-LEVEL JUDGEMENT

(++) very cheap
(++) no training required
(++) very fast
(--) very low level of detail
(---) most subjective, apply w/caution

ADEQUACY-FLUENCY (SEGMENT)

(+) cheap
(+) little training required
(+) fast
(+) reasonable level of detail
(-) subjective, 2+ evaluators and medium samples recommended

"A/B TESTING" (SEGMENT)

(+) cheap
(++) no training required
(+) fast
(-) low level of detail
(--) very subjective, 3+ evaluators and large samples recommended

"PE EFFORT PREDICTION"

(+) cheap
(-) PE training required
(+) fast
(-) low level of detail
(-) not suitable for Raw MT

HUMAN EVAL (ANALYTICAL)

11-30 CATEGORIES (DEEP)

(++) very high level of detail
(--) more training required, skilled linguists only
(--) very slow and costly
(*) risk of lower objectivity due to inter-evaluator differences

5-10 CATEGORIES (SHALLOW)

(+) high level of detail
(-) training required, linguists only
(-) slow
(++) reasonably objective with training and regular feedback

3+ SEVERITY LEVELS

(++) very high level of detail
(-) skilled linguists only
(*) good for PEMT, too much for Raw MT

2 SEVERITIES (MINOR, MAJOR)

(+) high level of detail
(-) linguists only
(+) useful already for raw MT