Scoring using NIST's "mteval-v08" MT evaluation script

A modification of IBM's BLEU score was created based on a study of N-gram co-occurrence statistics from four different sets of translations. The NIST score is the simple information-weighted sum of N-gram co-occurrences, normalized by number of N-gram occurrences and summed over 1-, 2-, 3-, 4- and 5-grams. This modified score demonstrates comparable or better performance at discriminating translation quality of different systems. System performance is characterized in terms of the F-ratio (of between-system to within-system inter-document variance) and also in terms of the correlation between the N-gram score and human judgments for different systems.

Table 1 N-gram scoring statistics for the DARPA 1994 French-English translation corpus. (Statistics based on output from 5 MT systems.)

	Correlation with Human Judgments (%)			F-ratio
	Adequacy	Fluency	Informativeness	
Human Judges	-	-	-	**87** (Adequacy) **82** (Fluency) **36** (Informativeness)
BLEU	**95.2**	**99.8**	**90.7**	**213**
Mteval-v08	**96.3**	**99.3**	**92.9**	**340**

Table 2 N-gram scoring statistics for the DARPA 1994 Spanish-English translation corpus. (Statistics based on output from 4 MT systems.)

	Correlation with Human Judgments (%)			F-ratio
	Adequacy	Fluency	Informativeness	
Human Judges	-	-	-	**63** (Adequacy) **62** (Fluency) **34** (Informativeness)
BLEU	**97.5**	**97.2**	**94.3**	**226**
Mteval-v08	**98.6**	**96.4**	**96.3**	**375**

Table 3 N-gram scoring statistics for the TIDES 2001 Chinese-English translation corpus. (Statistics based on output from 6 commercial MT systems.)

	Correlation with Human Judgments (%)		F-ratio
	Adequacy	Fluency	
Human Judges	-	-	**81** (Adequacy) **72** (Fluency)
BLEU	**96.7**	**96.7**	**56**
Mteval-v08	**99.4**	**98.2**	**147**

Table 4 N-gram scoring statistics for the TIDES 2001 Chinese-English translation corpus. (Statistics based on 7 independent professional human translations.)

	Correlation with Human Judgments (%)		F-ratio
	Adequacy	Fluency	
Human Judges	-	-	**25** (Adequacy) **52** (Fluency)
BLEU	**67.4**	**17.5**	**34**
Mteval-v08	**69.4**	**21.1**	**50**