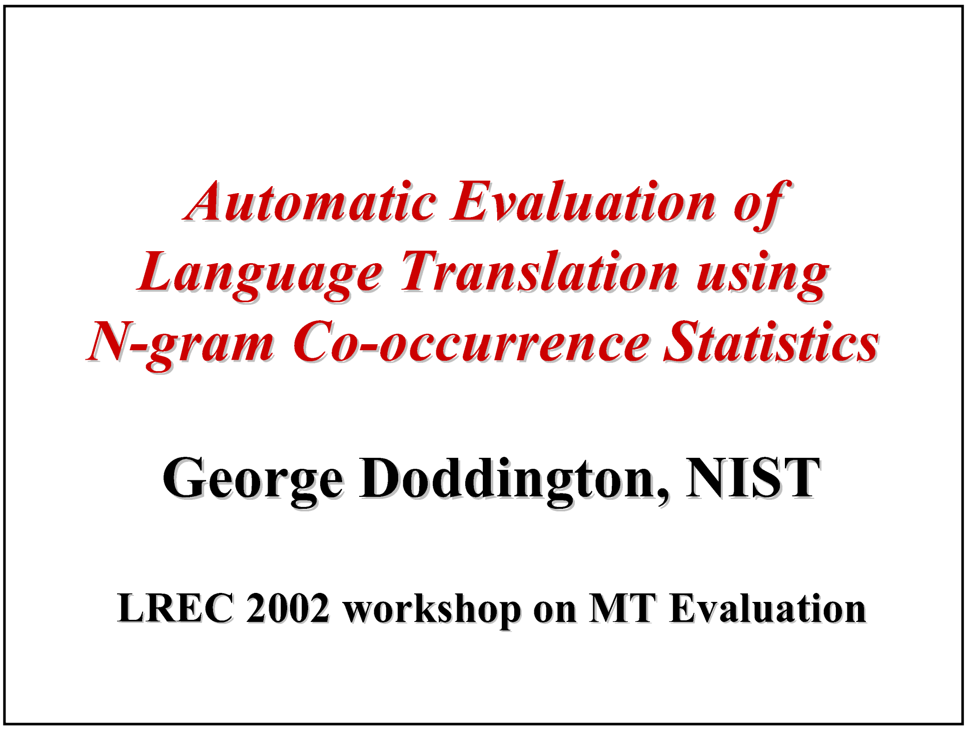
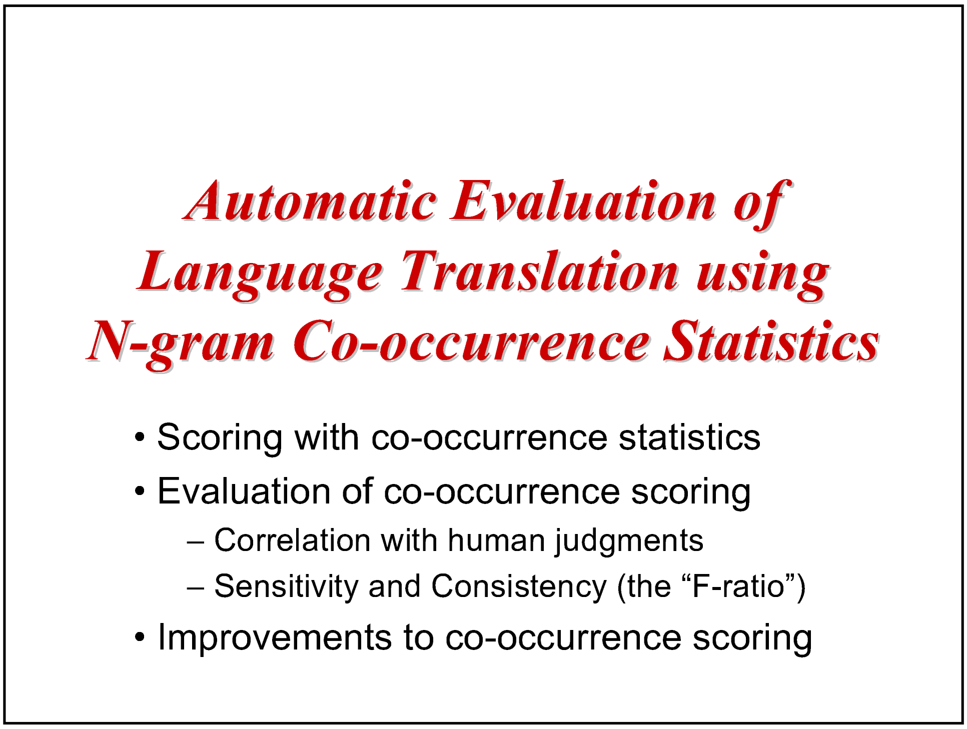


Automatic Evaluation of Language Translation using N-gram Co-occurrence Statistics

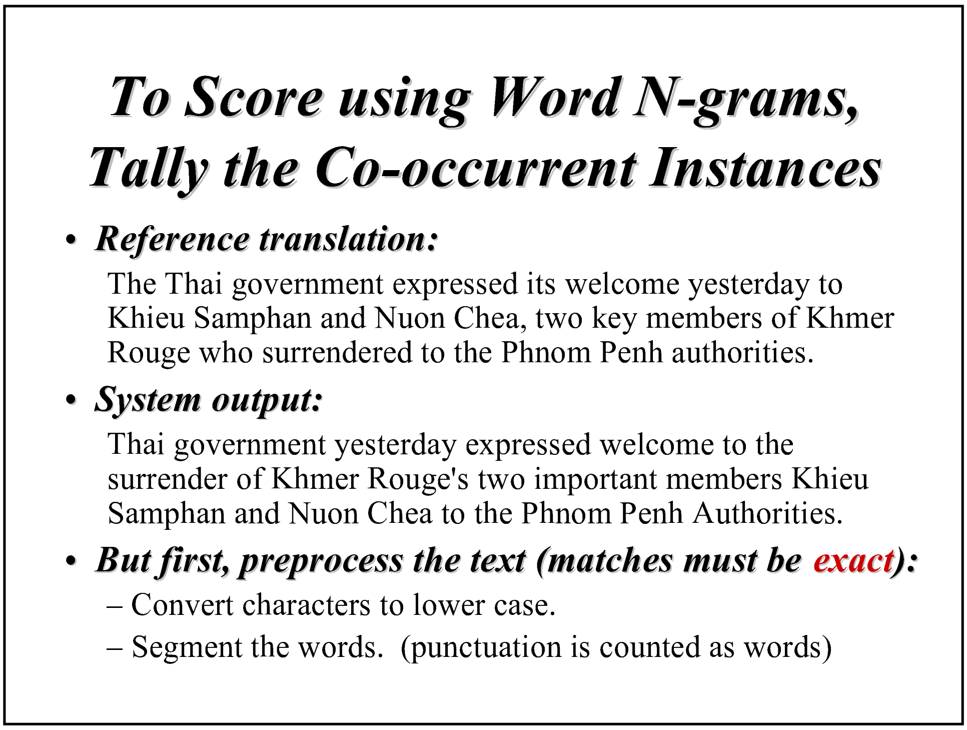
George Doddington, NIST

LREC 2002 workshop on MT Evaluation



Automatic Evaluation of Language Translation using N-gram Co-occurrence Statistics

- Scoring with co-occurrence statistics
- Evaluation of co-occurrence scoring
  - Correlation with human judgments
  - Sensitivity and Consistency (the “F-ratio”)
- Improvements to co-occurrence scoring



To Score using Word N-grams, Tally the Co-occurrent Instances

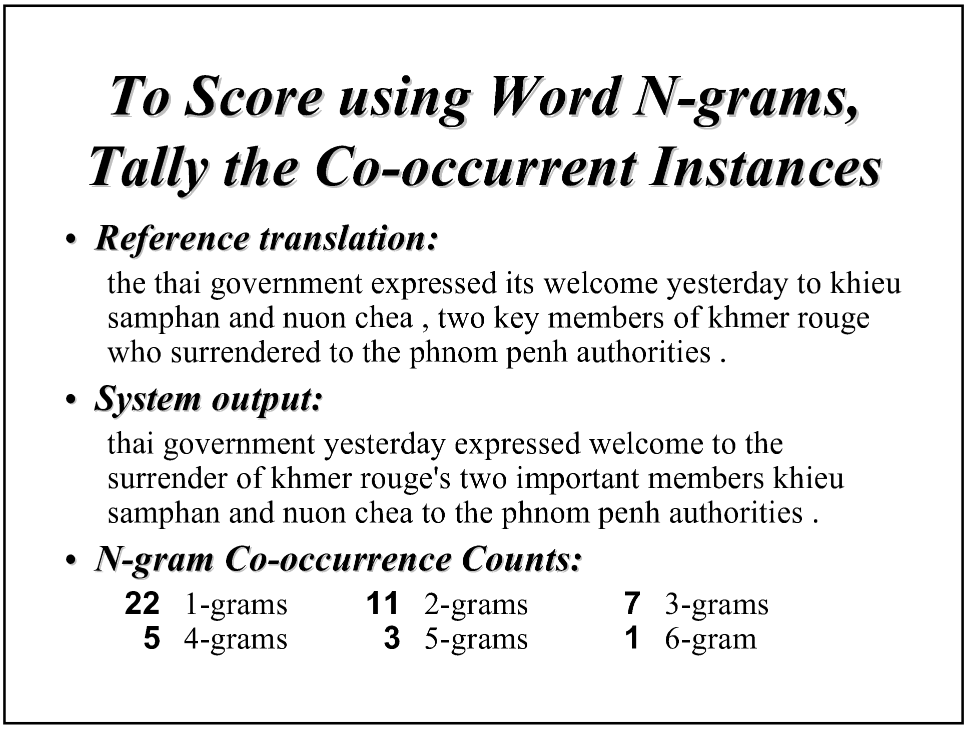
- ***Reference translation:***

  The Thai government expressed its welcome yesterday to Khieu Samphan and Nuon Chea, two key members of Khmer Rouge who surrendered to the Phnom Penh authorities.

- ***System output:***

  Thai government yesterday expressed welcome to the surrender of Khmer Rouge's two important members Khieu Samphan and Nuon Chea to the Phnom Penh Authorities.

- ***But first, preprocess the text (matches must be exact):***
  - Convert characters to lower case.
  - Segment the words. (punctuation is counted as words)



To Score using Word N-grams, Tally the Co-occurrent Instances

- ***Reference translation:***

  the thai government expressed its welcome yesterday to khieu samphan and nuon chea , two key members of khmer rouge who surrendered to the phnom penh authorities .

- ***System output:***

  thai government yesterday expressed welcome to the surrender of khmer rouge's two important members khieu samphan and nuon chea to the phnom penh authorities .

- ***N-gram Co-occurrence Counts:***

22	1-grams	**11**	2-grams	**7**	3-grams
5	4-grams	**3**	5-grams	**1**	6-gram

The IBM Score (BLEU)

$$Score = \exp\left\{ \sum_{n=1}^{N} w_n \log(p_n) \quad - \quad \max\left(\frac{L^*_{ref}}{L_{sys}} - 1, \quad 0 \right) \right\}$$

where

$$p_n = \frac{\sum_i \begin{pmatrix} \text{the number of } n\text{-grams in segment } i, \\ \text{in the translation being evaluated, with} \\ \text{a matching reference cooccurence in segment } i \end{pmatrix}}{\sum_i \begin{pmatrix} \text{the number of } n\text{-grams in segment } i, \\ \text{in the translation being evaluated} \end{pmatrix}}$$

$$w_n = N^{-1}$$

$$N = 4$$

The IBM Score (BLEU)

$$Score = \exp\left\{ \sum_{n=1}^{N} w_n \log(p_n) \quad - \quad \max\left(\frac{L^*_{ref}}{L_{sys}} - 1, \quad 0 \right) \right\}$$

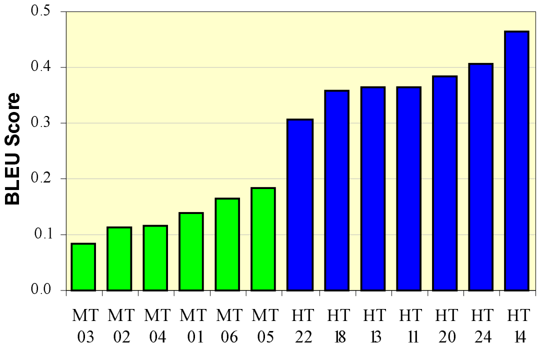
and

L^*_{ref} = the number of words in the reference translation that is closest in length to the translation being scored

L_{sys} = the number of words in the translation being scored

Example BLEU Scores for the 2001 DARPA Evaluation

- 80 Chinese news documents were translated to English from newswire and VOA transcripts.
- Each document was scored using 4 independent professional translations

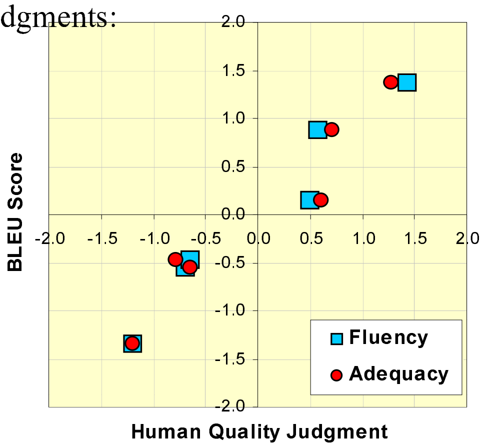


Evaluation of Automatic Scoring of Language Translation

- The score must be able to accurately predict (human judgments of) ***quality***.
  - Note that different dimensions of judgment may require different scoring algorithms.
- The score must be ***sensitive*** yet ***reliable***.
  - ***Sensitivity***: Large differences in scores should result for significantly different systems
  - ***Reliability***: Systems should always score the same, regardless of different test sets (docs and ref translations)
  - Use one measure for both sensitivity and reliability: the ***F-ratio*** = (Between-sys variance)/(Within-sys variance)

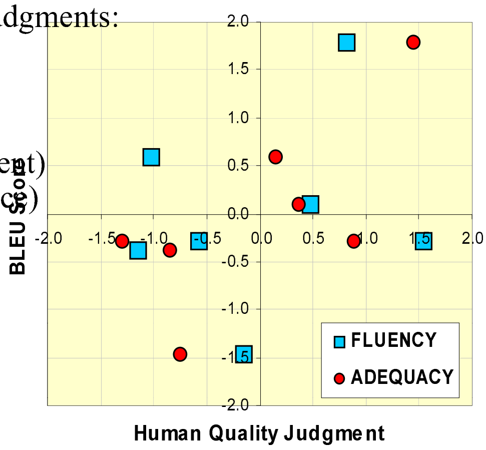
Evaluation of BLEU Scores for the 80 document Chinese corpus

- For the 6 commercial MT systems:
  - Correlation with human judgments:
    96.2% for “Adequacy”
    97.0% for “Fluency”
  - F-ratio:
    43 (document variation)
    45 (reference variation)



Evaluation of BLEU Scores for the 80 document Chinese corpus

- For 7 professional translators:
  - Correlation with human judgments:
    70.8% for “Adequacy”
    21.2% for “Fluency”
  - F-ratio:
    27 (with respect to document)
    3 (with respect to reference)



The NIST MTeval Score

$$Score = \sum_{n=1}^{N} \left\{ \sum_{\substack{\text{all } w_1 ... w_\mathrm{n} \\ \text{that co-occur}}} Info(w_1 ... w_\mathrm{n}) \right\} \cdot \exp\left\{ \beta \log^2 \left[\min\left(\frac{L_{sys}}{\overline{L}_{ref}}, 1 \right) \right] \right\}$$

where

$$Info(w_1 ... w_\mathrm{n}) = \log_2 \left(\frac{\text{the \# of occurrences of } w_1 ... w_{\mathrm{n-1}}}{\text{the \# of occurrences of } w_1 ... w_\mathrm{n}} \right)$$

$N = 5$

The NIST MTeval Score

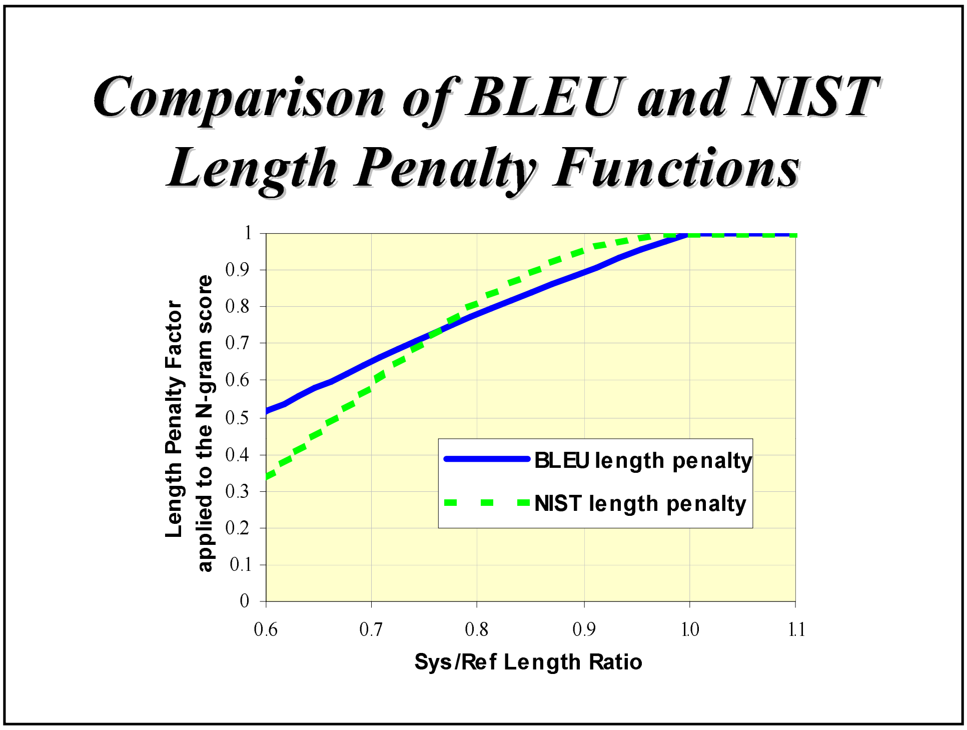
$$Score = \sum_{n=1}^{N} \left\{ \sum_{\substack{\text{all } w_1 ... w_\mathrm{n} \\ \text{that co-occur}}} Info(w_1 ... w_\mathrm{n}) \right\} \cdot \exp\left\{ \beta \log^2 \left[\min\left(\frac{L_{sys}}{\overline{L}_{ref}}, 1 \right) \right] \right\}$$

and

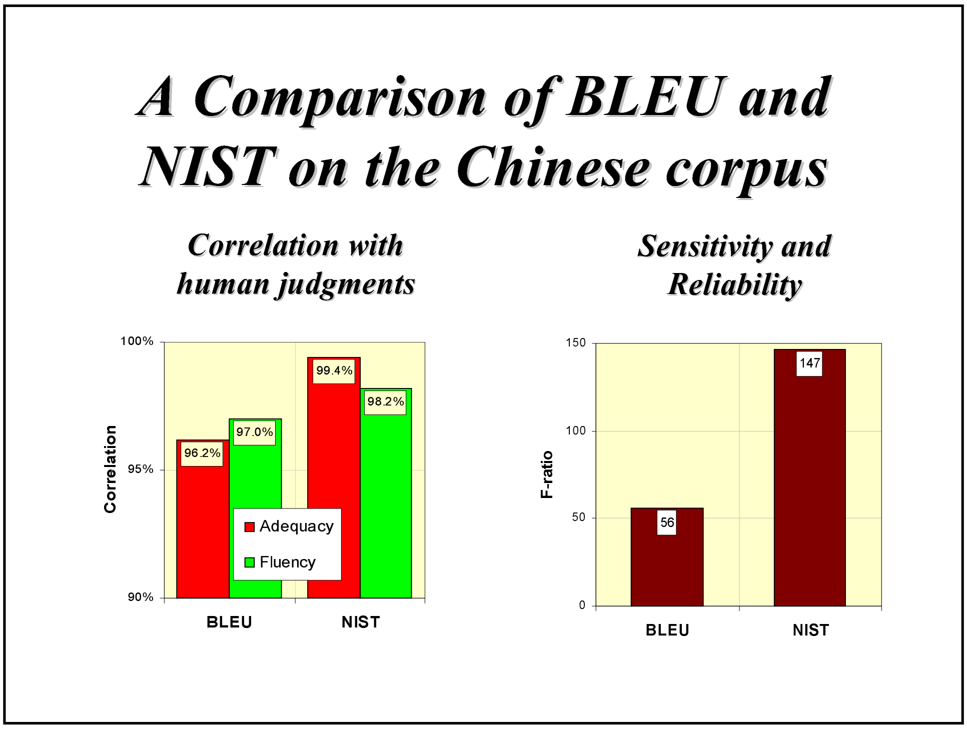
β is chosen to make the length penalty factor = 0.5 when the # of words in the system output is 2/3$^{\text{rds}}$ of the average # of words in the reference translation

L_{sys} = the number of words in the translation being scored

$\overline{L}_{ref}$ = the average number of words in a reference translation, averaged over all reference translations



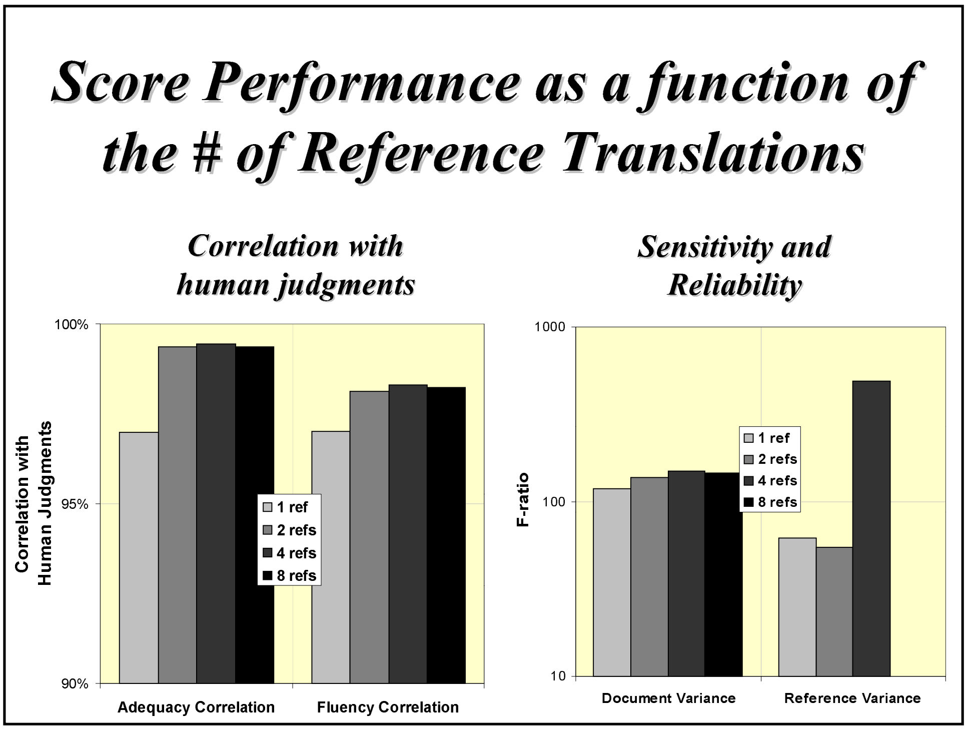
Comparison of BLEU and NIST Length Penalty Functions



A Comparison of BLEU and NIST on the Chinese corpus

Correlation with human judgments

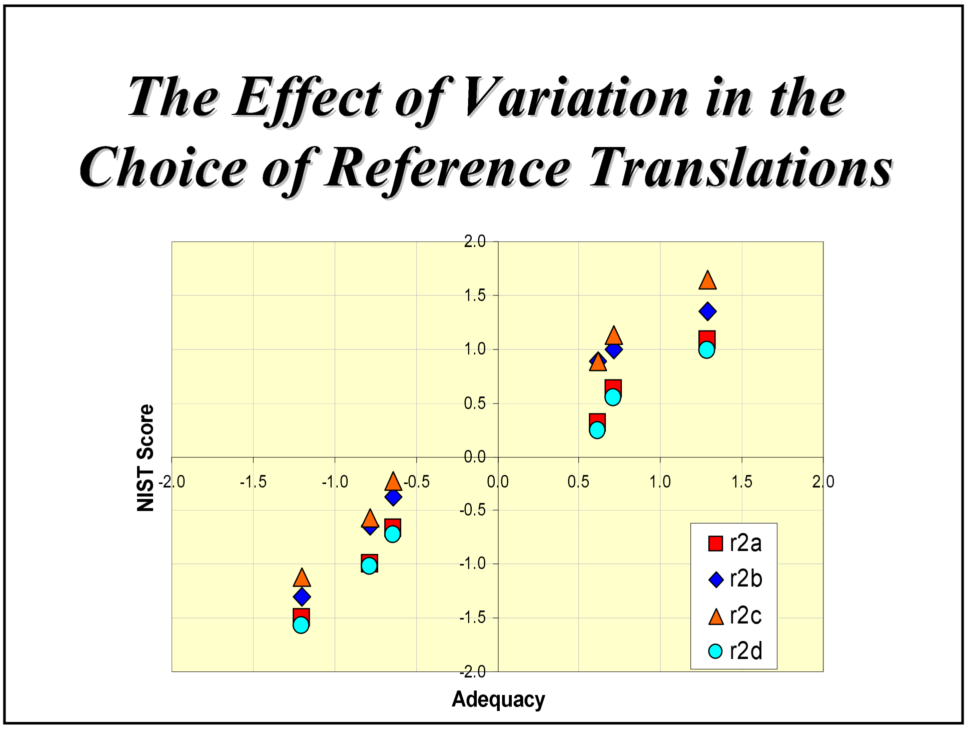
Sensitivity and Reliability



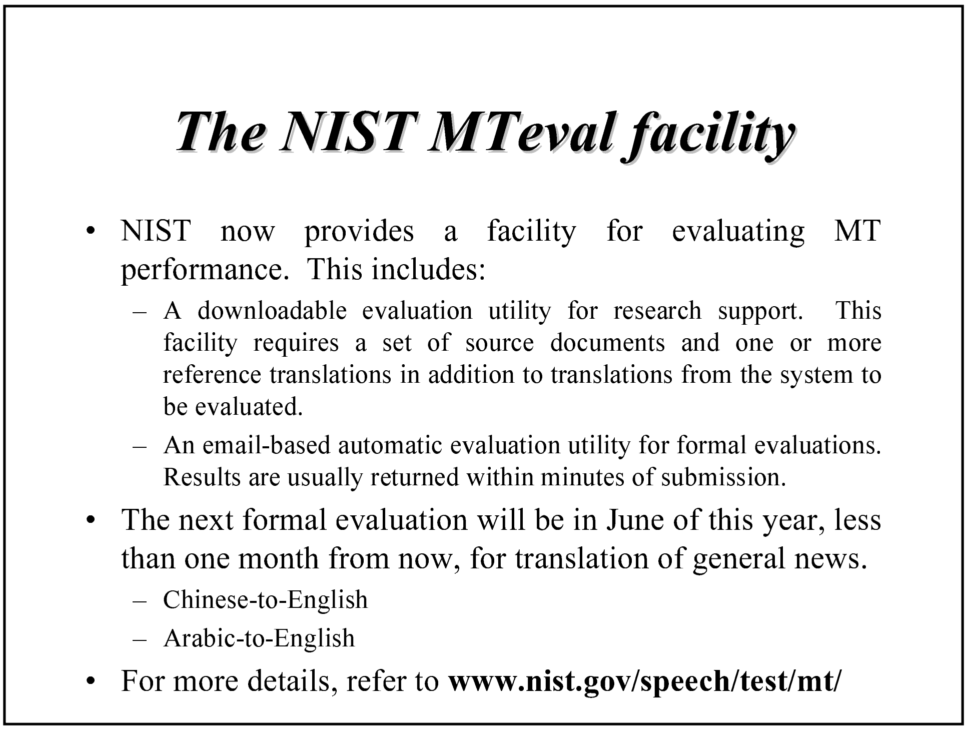
Score Performance as a function of the # of Reference Translations

Correlation with human judgments

Sensitivity and Reliability



The Effect of Variation in the Choice of Reference Translations



The NIST MTeval facility

- NIST now provides a facility for evaluating MT performance. This includes:
  - A downloadable evaluation utility for research support. This facility requires a set of source documents and one or more reference translations in addition to translations from the system to be evaluated.
  - An email-based automatic evaluation utility for formal evaluations. Results are usually returned within minutes of submission.
- The next formal evaluation will be in June of this year, less than one month from now, for translation of general news.
  - Chinese-to-English
  - Arabic-to-English
- For more details, refer to **www.nist.gov/speech/test/mt/**