A Study on NLP Performance Metrics With a Focus on MT

Gwenaelle Cunha Sergio

Artificial Brain Research Lab., School of Electronics Engineering,
Kyungpook National University

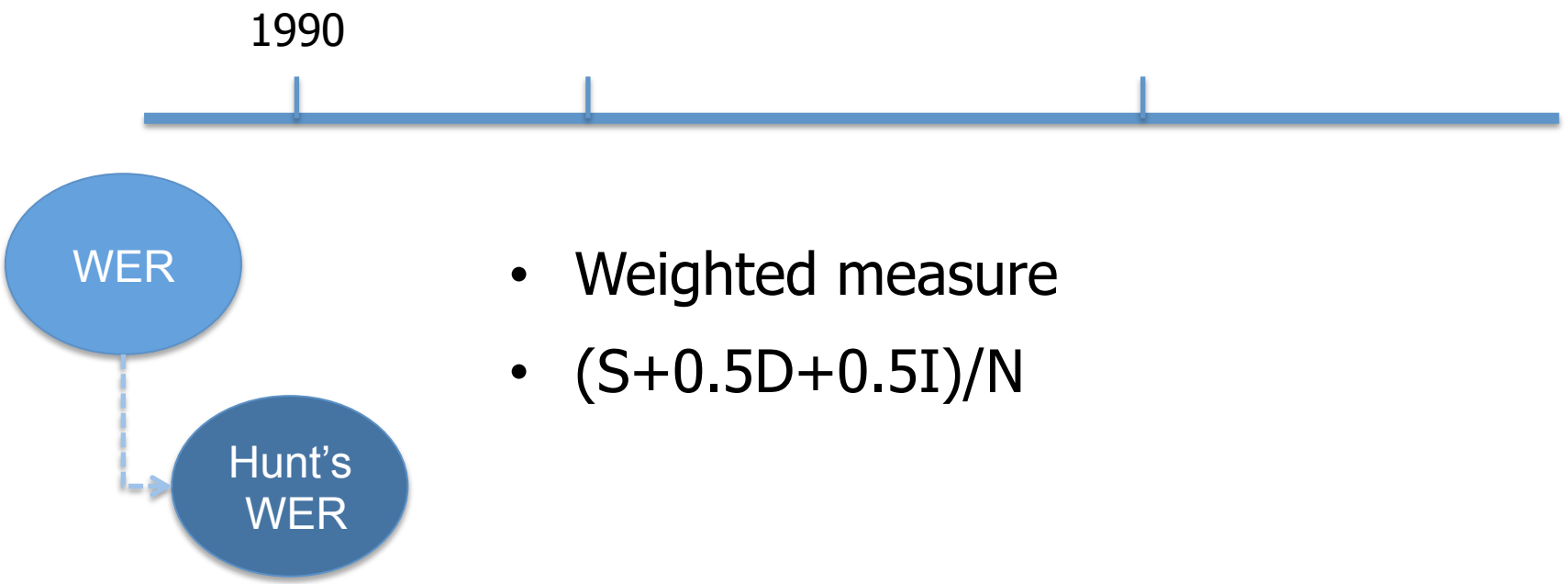
06-April-2018

Word Error Rate

WER

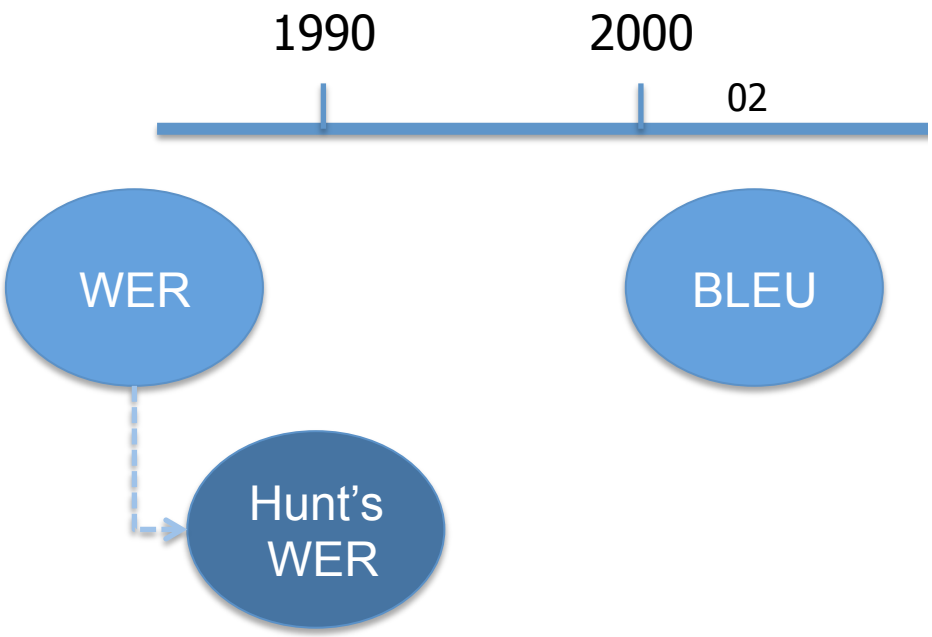
- Levenshtein distance: edit distance
- > 0 when ref=hyp
- (S+D+I)/N = (S+D+I)/(S+D+C)

Hunt's Weighted Word Error Rate



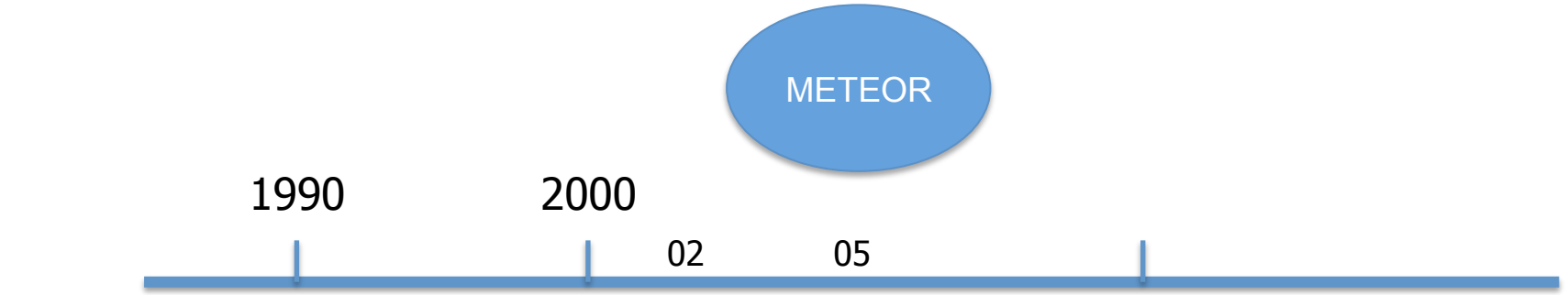
- Weighted measure
- (S+0.5D+0.5I)/N

BiLingual Evaluation Understudy



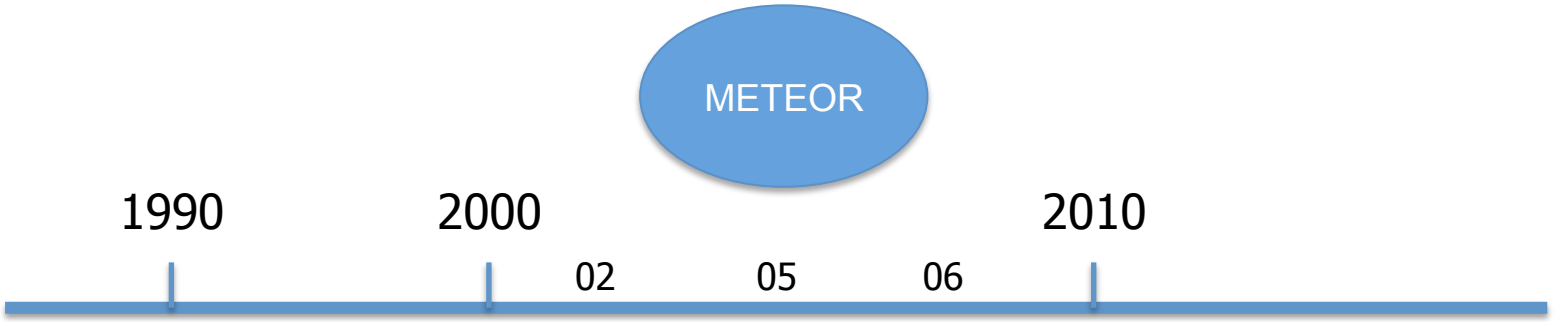
- Measures how many words **overlap** in a given translation when compared to a reference translation
- Limitation:
  - Doesn't consider different types of errors (insertions, substitutions, synonyms, paraphrase)
  - Designed to be a corpus measure, so it has undesirable properties when used for single sentences

Metric for Evaluation of Translation with Explicit ORdering



- Harmonic mean of unigram precision and recall
- Includes: exact word, stem and synonym matching
- Similar performance to BLEU on corpus level as well as **sentence level**

Translation Edit Rate

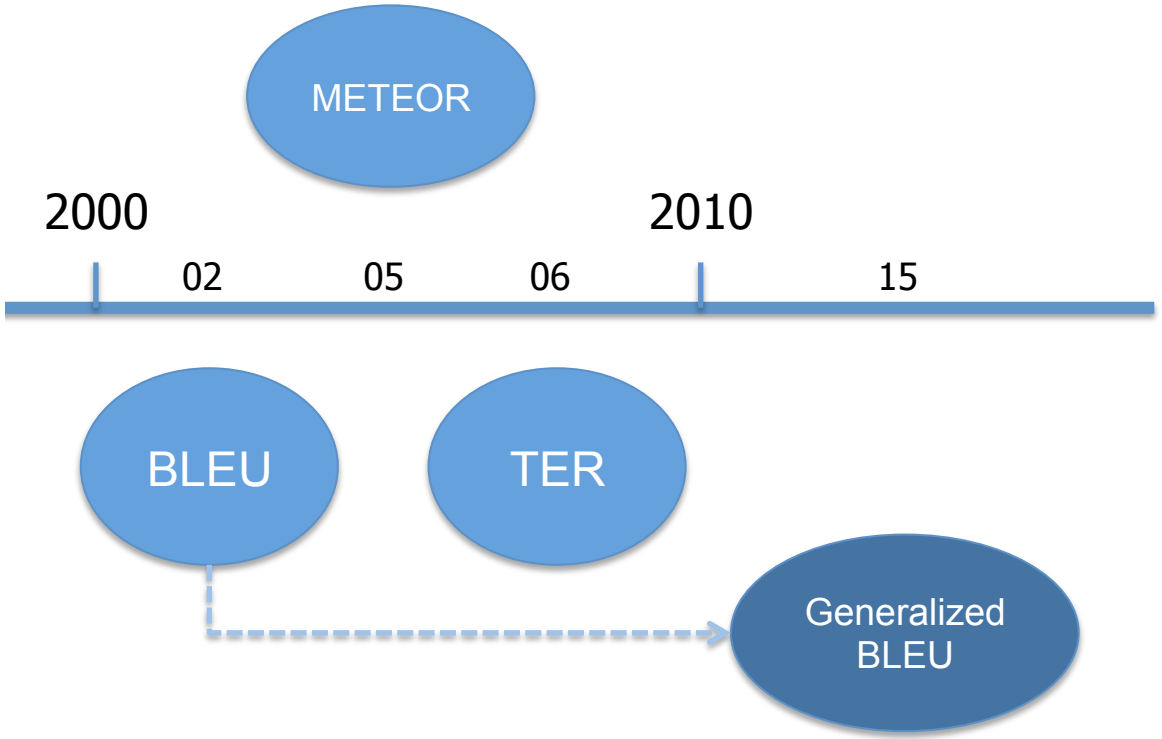


- TER = E/R
- E: minimum number of edits
- R: average length of reference text
- Better than BLEU for estimation of sentence **post-editing effort**

TER

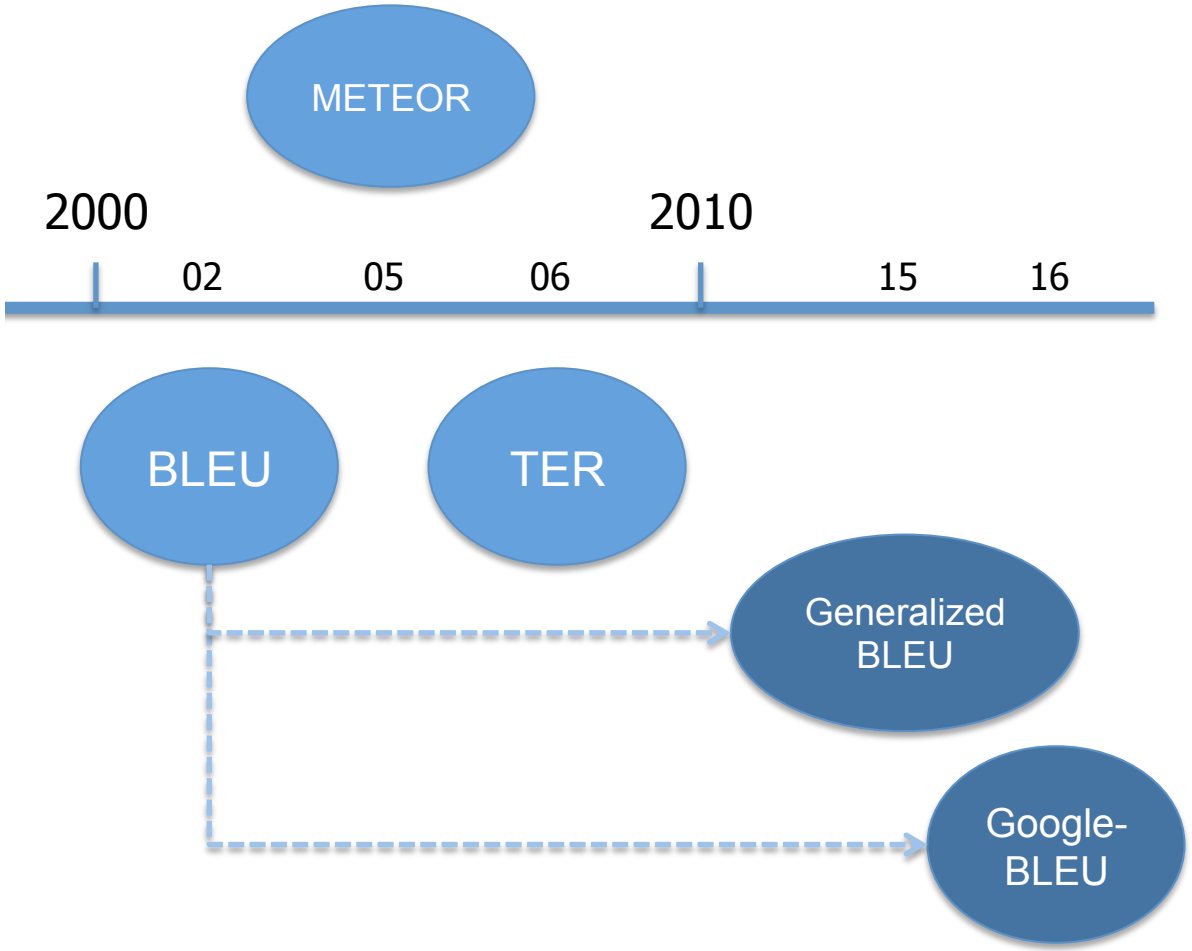
Generalized BLEU

- Closer to human judgments than BLEU
- Computes n-gram precisions over the reference and assigns more weight to n-grams that have been correctly changed from the source

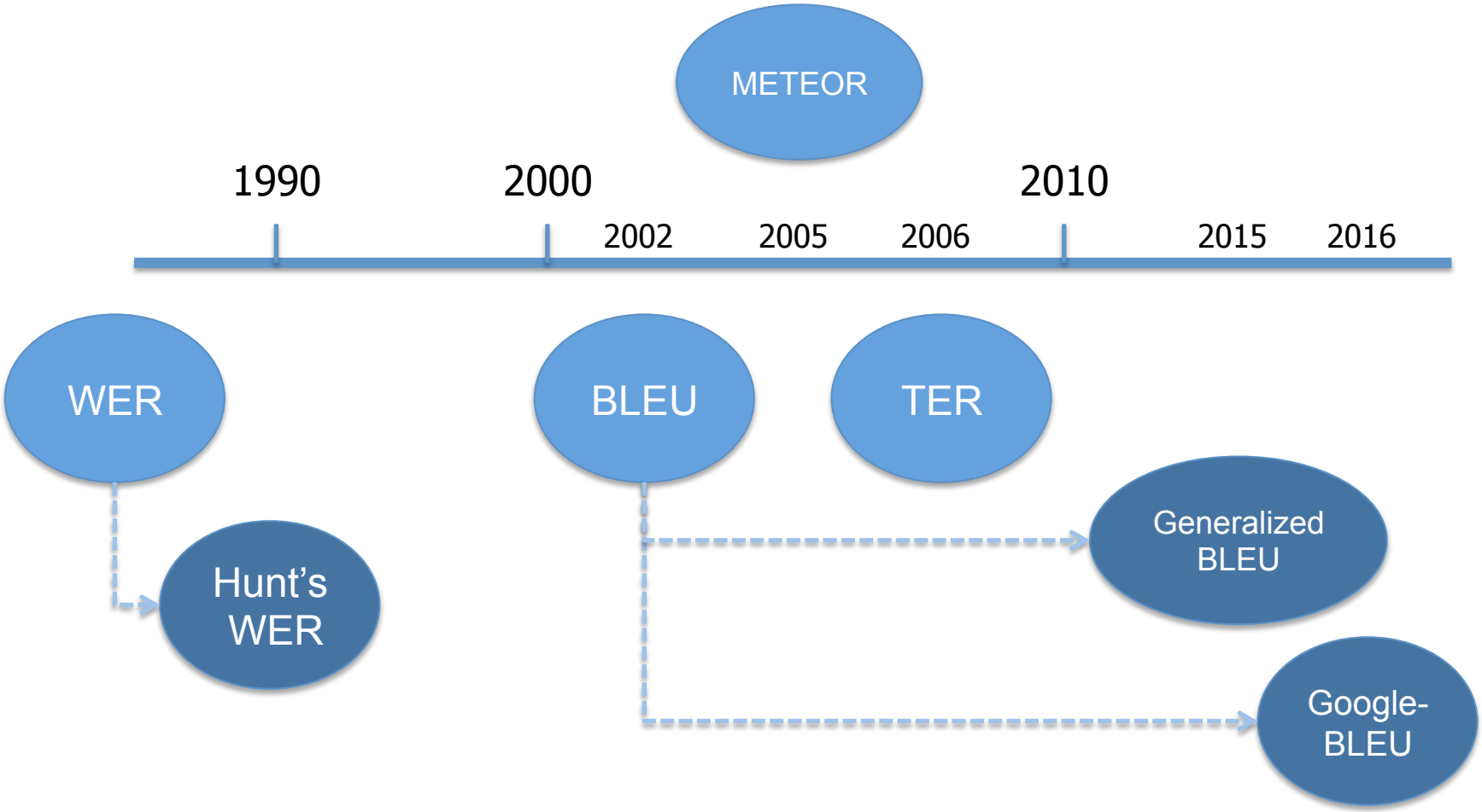


Google-BLEU

- Minimum of BLEU recall and precision applied to 1, 2, 3 and 4 grams
- Similar performance to BLEU on corpus level as well as **sentence level**



MT Metrics Timeline



Other NLP Metrics