Empirical Confidence Intervals

In statistics, confidence intervals indicate how well was a parameter (e.g. the mean) of a random variable with *known/assumed distribution* estimated from a set of repeated measurements.

- We don't want to assume any distribution!
- How to "repeat" experiments with a deterministic MT system?

Use "bootstrapping" (Koehn, 2004):

1. Obtain 1000 different test sets:
   Randomly select sents., repeat some, ignore some, preserving test set size.
2. Sort by the score.
3. Drop top and bottom 2.5% (i.e. 25 out of 1000) results.

$\Rightarrow$ The lowest and highest remaining scores are 95% empirical confidence interval around the score obtained on the full test set.