Fixing Fundamenal Issues of BLEU

Evaluate coarser units:

- Lemmas or deep-lemmas instead of word forms:
  - e.g. SemPOS (Kos and Bojar, 2009): bags of t-lemmas.
- Sequences of characters:
  - e.g. chrF3 (Popović, 2015): F-score of character 6-grams.
- Use shorter of gappy sequences:
  - e.g. BEER (Stanojevic and Sima'an, 2014) uses characters and also pairs of (not necessarily adjacent) words.

Use better references:

- Using more references alone helps.
- Post-edited references serve better.
  - e.g. HTER (Snover et al., 2006): Measuring edit distance to manually corrected output.