BLEU changes on test set

- regression 11%
- improvement 7%
- no change 82%