# BLEU (Papineni et al., 2002)

- Based on geometric mean of $n$-gram precision.

$\approx$ ratio of 1- to 4-grams of hypothesis confirmed by a ref. translation

| Src | Vom Glück der träumenden Kamele | Confirmed |
|---|---|---|
| Ref | On the happiness of dreaming camels | 1 2 3 4 |
| System A | > **The happiness of dreaming camels** < | 5 4 3 2 |
| System B | > Dreaming of the luck that camels < | 4 1 0 0 |
| n-grams confirmed: | none, unigram, bigram, *trigram*, **fourgram** | |

E.g. Sys B produced 6 unigrams (4 confirmed), 4 bigrams (1 confirmed), . . .

$$\text{BLEU} = \text{BP} \cdot \exp\left( {}^1\!/_4 \log\left(\frac{4}{6}\right) + {}^1\!/_4 \log\left(\frac{1}{7}\right) + {}^1\!/_4 \log\left(\frac{0}{6}\right) + {}^1\!/_4 \log\left(\frac{0}{5}\right) \right)$$

BP is "brevity penalty"

${}^1\!/_4$ is the "log-domain denominator" equivalent for $\sqrt[4]{\cdot}$ in geometric mean

# BLEU: Avoiding Cheating

- Confirmed counts "clipped" to avoid overgeneration.
- "Brevity penalty" applied to avoid too short output:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases}$$

| | |
|---|---|
| Ref 1: | The cat is on the mat . |
| Ref 2: | There is a cat on the mat . |
| Candidate: | The the the the the the the . |
| | $\Rightarrow$ Clipping: only $\frac{3}{8}$ unigrams confirmed. |
| Candidate: | The the . |
| | $\Rightarrow$ $\frac{3}{3}$ unigrams confirmed but the output is too short. |
| | $\Rightarrow$ BP $= e^{1-7/3} = 0.26$ strikes. |

The candidate length $c$ and "effective" ref. length $r$ calculated over the whole test set.

# References

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania.