BLEU (Papineni et al., 2002)

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

\[ \text{BLEU} = \text{BP} \cdot \exp \left( \frac{1}{4} \log \left( \frac{4}{6} \right) + \frac{1}{4} \log \left( \frac{1}{7} \right) + \frac{1}{4} \log \left( \frac{0}{6} \right) + \frac{1}{4} \log \left( \frac{0}{5} \right) \right) \]

BP is “brevity penalty”

$\frac{1}{4}$ is the “log-domain denominator” equivalent for $\sqrt[4]{\cdot}$ in geometric mean

≈ ratio of 1- to 4-grams of hypothesis confirmed by a ref. translation

<table>
<thead>
<tr>
<th>Src</th>
<th>Vom Glück der träumenden Kamele</th>
<th>Confirmed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>On the happiness of dreaming camels</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>System A &gt;</td>
<td>The happiness of dreaming camels &lt;</td>
<td>5 4 3 2</td>
</tr>
<tr>
<td>System B &gt;</td>
<td>Dreaming of the luck that camels &lt;</td>
<td>4 1 0 0</td>
</tr>
</tbody>
</table>

n-grams confirmed: none, unigram, bigram, trigram, fourgram

E.g. Sys B produced 6 unigrams (4 confirmed), 4 bigrams (1 confirmed), . . .
**BLEU: Avoiding Cheating**

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

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

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 the.*

⇒ Clipping: only \(\frac{3}{8}\) unigrams confirmed.

Candidate: *The the.*

⇒ \(\frac{3}{3}\) unigrams confirmed but the output is too short.

⇒ \(BP = e^{1-7/3} = 0.26\) strikes.

The candidate length \(c\) and “effective” ref. length \(r\) calculated over the whole test set.