Finding the best partition

$$\pi^* = \arg\max_{\pi} P(\mathbf{W} \mid \pi) = \arg\max_{\pi} \log P(\mathbf{W} \mid \pi)$$

One can derive[2] $L(\pi)$ to be

$$L(\pi) = \underbrace{\sum_{w} n_w \log n_w}_{\substack{\text{(nearly) unigram entropy} \\ \text{(fixed w.r.t. } \pi\text{)}}} + \underbrace{\sum_{c_i, c_j} n_{c_i, c_j} \log \frac{n_{c_i, c_j}}{n_{c_i} \cdot n_{c_j}}}_{\substack{\text{(nearly) mutual information} \\ \text{(varies with } \pi\text{)}}}$$

[2] Sven Martin, Jörg Liermann, and Hermann Ney. "Algorithms for Bigram and Trigram Word Clustering". In: *Speech Commun.* 24.1 (Apr. 1998), pp. 19–37.