Finding the best partition (cont'd)

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

Direct maximization is **intractable!** Thus, agglomerative (bottom-up) clustering is used as a greedy heuristic.

The best merge is determined by the lowest loss in average mutual information.