How do we train the model?

Log-likelihood maximization:

$$\mathcal{L} = \sum_{u \in W} \sum_{v \in W} n_{uv} \log p(u|v)$$

word co-occurrence

Method:
- SGD, online by word pairs in the corpus

Problem:
- *softmax* over vocabulary is slow!