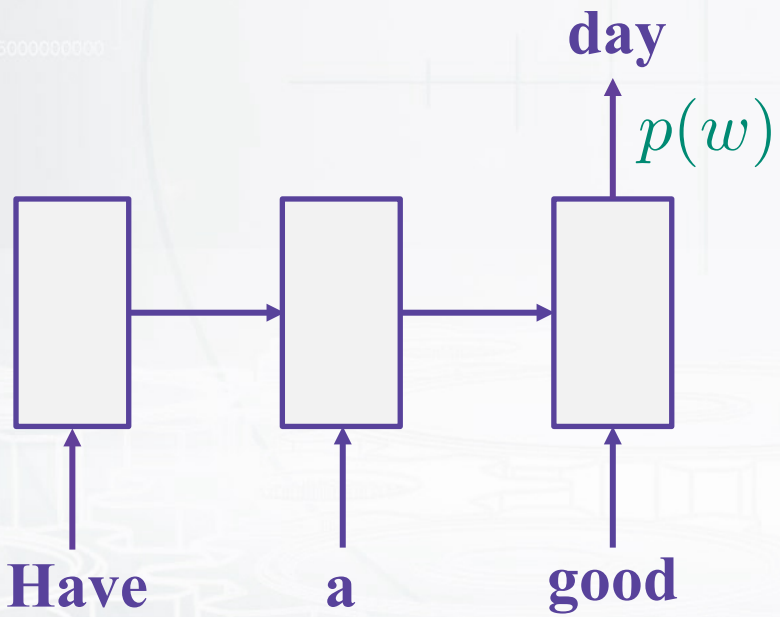
How do we train it?

Cross-entropy loss (for one position):

$$-\log p(w_i) = -\sum_{w \in V} [w = w_i] \log p(w)$$

Only one non-zero



- **Target:** word w_i
- **Output:** probabilities $p(w)$