Beam Search

- Instead of taking the arg max in every, step, keep a list (or beam) of k-best scoring hypotheses.
- Hypothesis = partially decoded sentence $\rightarrow$ score
- Hypothesis score $\psi_t = (y_1, y_2 \ldots, y_t)$ is the probability of the decoded sentence prefix up to t-th word.

$$p(y_1, \ldots, y_t|h) = p(y_1|h) \cdot \cdots \cdot p(y_t|y_1, \ldots, y_{t-1}|h)$$

- Rule to compute the score of an extended hypothesis ψ_t:

$$p(\psi_t, y_{t+1}|h) = p(\psi_t|h) \cdot p(y_{t+1}|h)$$

- Prefers shorter hypotheses $\rightarrow$ normalization necessary.