

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 $\psi_t = (y_1, y_2 \dots, y_t)$ is the probability of the decoded sentence prefix up to t -th word.

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