Decoding Strategies

‣ LMs place a distribution $P(y_i | y_1, ..., y_{i-1})$

‣ seq2seq models place a distribution $P(y_i | \mathbf{x}, y_1, ..., y_{i-1})$

‣ Generation from both models looks similar; how do we do it?

  ‣ Option 1: max y_i $P(y_i | y_1, ..., y_{i-1})$ — take greedily best option

  ‣ Option 2: use beam search to find the sequence with the highest prob.

  ‣ Option 3: sample from the model; draw y_i from that distribution

‣ This segment: beam search

Beam Search

Beam Search

Beam Search: Applications

‣ Beam search is used frequently in seq2seq conditional generation settings like machine translation where finding the highest-probability hypothesis is really important

‣ Better models make beam search less crucial. The models are better at "planning ahead" internally so it's less likely that they assign a sequence $y_1 \ldots y_{n-1}$ high probability if there's no good y_n for it