束搜索

贪心搜索

- 在seq2seq中我们使用了贪心搜索来预测序列
  - 将当前时刻预测概率最大的词输出
- 但贪心很可能不是最优的：

贪心： $0.5\times0.4\times0.4\times0.6=0.048$

Time step	1	2	3	4
A	0.5	0.1	0.2	0.0
B	0.2	0.4	0.2	0.2
C	0.2	0.3	0.4	0.2
<eos>	0.1	0.2	0.2	0.6

很好的选项：$0.5\times0.3\times0.6\times0.6=0.054$

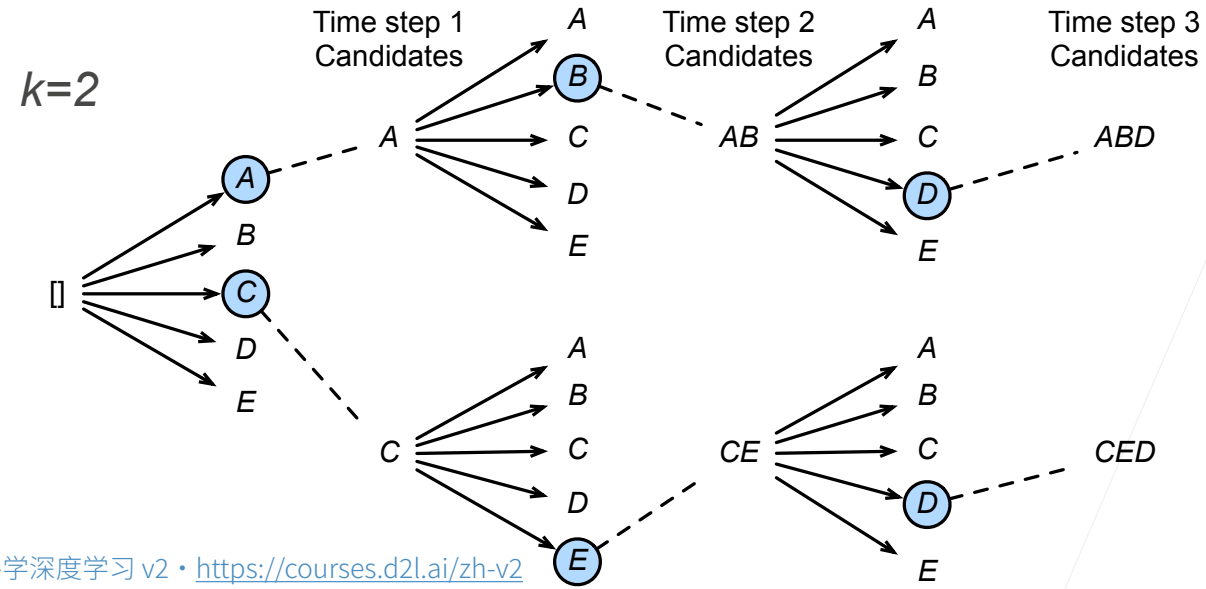
Time step	1	2	3	4
A	0.5	0.1	0.1	0.1
B	0.2	0.4	0.6	0.2
C	0.2	0.3	0.2	0.1
<eos>	0.1	0.2	0.1	0.6

穷举搜索

- 最优算法：对所有可能的序列，计算它的概率，然后选取最好的那个
- 如果输出字典大小为 n，序列最长为 T，那么我们需要考察 n^T 个序列
  - $n = 10000, \quad T = 10: \quad n^T = 10^{40}$
  - 计算上不可行

束搜索

- 保存最好的 k 个候选
- 在每个时刻，对每个候选新加一项（n 种可能），在 kn 个选项中选出最好的 k 个



束搜索

- 时间复杂度 $O(knT)$
  - $k = 5, \quad n = 10000, \quad T = 10: \quad knT = 5 \times 10^5$
- 每个候选的最终分数是：

$$\frac{1}{L^\alpha} \log p(y_1, \ldots, y_L) = \frac{1}{L^\alpha} \sum_{t'=1}^{L} \log p(y_{t'} \mid y_1, \ldots, y_{t'-1}, \boldsymbol{c})$$

  - 通常 $\alpha = 0.75$

总结

- 束搜索在每次搜索时保存 k 个最好的候选
  - $k = 1$ 时是贪心搜索
  - $k = n$ 时是穷举搜索