Beam Search

- LM's Place distribution $P(y_i \mid y_1, \cdots, y_{i-1})$
- Seq2Seq models place distribution $P(y_i \mid x, y_1, \cdots, y_{i-1})$
- Generation from both models looks similar; how do we do it?

Decoding strategies →

- Option 1: $\max_{y_i} P(y_i \mid y_1, \cdots, y_{i-1})$ - take greedily best option
- Option 2: Use beam search to find sequence w/ highest probability

Doesn't Work ✗ Option 3: Sample from model; draw y_i from that distribution

- Option 4: Nucleus Sampling

- **Beam Search**

Assume vocab size $|V|$ and sequence of length n

$P(y_1)$

The	0.01
A	0.005
one	0.004
I	0.003
⋮	⋮

→ The dog 0.0004
→ The cat 0.0003
→ The fish 0.0002

End up with $|V|^n$ total sequences ↑ too expensive

- **Beam search**: approximates exhaustive search w/ less compute
  - keep the top K hypotheses at each step

K=3

The ←	The dog
A ←	The cat
one ←	A fish
~~I~~	one second
	✗

} End up with $K|V|$ total sequences

Runtime: $K \times n$ transformer calls

$K \times |V| \times n$ hypotheses

- **Drawbacks** of beam search:
  - **Degeneration**: If a fragment is repeated 2-3x, it has a high probability to keep repeating. Model fails because it's locally normalized

- **Drawbacks** of Sampling: Long Tail

$P(y \mid$ they live in a desert near $)$

0.01	roads
0.01	towns
0.01	people
0.005	straits
...	
0.0005	town

} Good options, account for ~90% of total probability mass; 90% chance of good prediction

] - Long tail w/ 10% of mass

solution → Nucleus Sampling

* On average, every 10 words, we get something from 10% tail of distribution. We don't want this!