Decoding Techniques

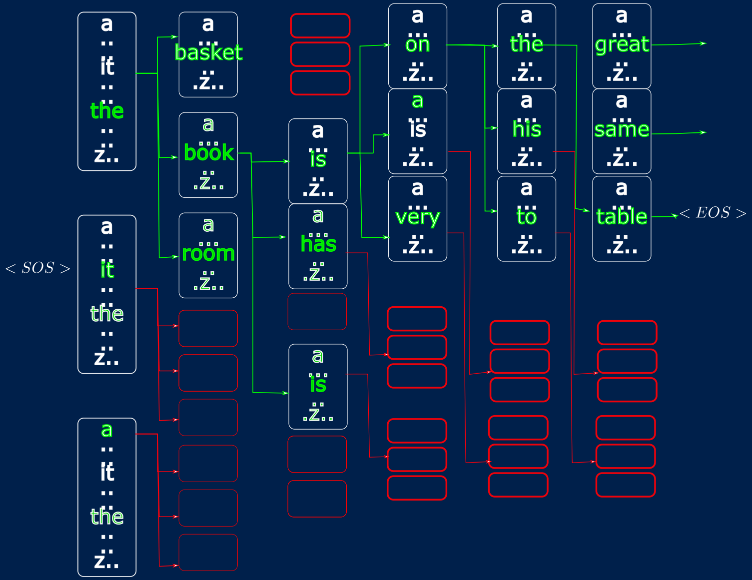
Ramaseshan Ramachandran

Beam search is a heuristic search algorithm that selects a few candidate hypothesis from $|V|$. It reduces memory requirement by using only a $M < |V|$ candidates using a score.

- Maintain M candidates/hypothesis at each time step - $C_t = (x_1^1, ..x_t^1)...(x_1^M...x_t^M)$
- Compute C_{t+1} by expanding C_t and keeping the best M candidates
- $\tilde{C} = \bigcup_{i=1}^{M} C_{t-1}^i$

Typical Beam width of size 5-10 used in NMT. The bilingual evaluation understudy (BLEU) scores computed using Beam search using B=5-10 are comparable

BEAM SEARCH - BEAM WIDTH = 3



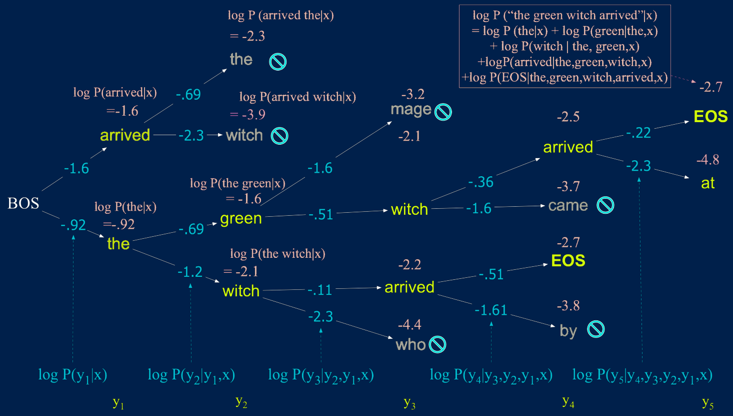


Figure: Scoring for beam search decoding with a beam width of k= 2. We maintain the log probability of each hypothesis in the beam by incrementally adding the logprob of generating each next token. Only the top k paths are extended to the next step[1].

BEAM SEARCH SUMMARY

1. Use all possible partial translations - exhaustive search
2. Beam size, $b = 1$ - greedy search - Words are predicted until the $< EOS >$ is found
3. $b > 1$ - several hypotheses
4. Each hypothesis will be produced until the $< EOS >$ is found
5. Each hypothesis will have a translation
6. The length of all hypothesis may not be the same
7. We could use different ***terminate*** conditions
   - Fixed time steps
   - Compute until $< EOS >$ is reached for each hypothesis
8. Use either log probability or product of conditional probability to find the scores for each hypothesis that maximizes
   - $P(y_1, y_2, ...y_m|\mathsf{X}) = \prod_{t=1}^{T} P(y_t| < SOS >, \ldots, y_{t-1}, \mathsf{X})$
   - $P(y_1, y_2, ...y_m|\mathsf{X}) = \sum_{t=1}^{T} \log P(y_t| < SOS >, \ldots, y_{t-1}, \mathsf{X})$