Stochastic Beams and Where to Find Them:

The Gumbel-Top-k Trick for Sampling Sequences Without Replacement

W. Kool, H. van Hoof, M. Welling 2019

ML Paper Club @Google Campus with nPlan

4th July 2019

François Steiner - francois@b3metrix.com

The idea

A magic trick to connect **sampling** and (deterministic) **beam search**, applying the “Gumbel-top-k” trick on a factorised distribution over **sequences**.

BEAM SEARCH

Sequence models – the chain rule

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\boldsymbol{y}_{1:t}) &= p_{\boldsymbol{\theta}}(y_t|\boldsymbol{y}_{1:t-1}) \cdot p_{\boldsymbol{\theta}}(\boldsymbol{y}_{1:t-1}) \\
&= \prod_{t'=1}^{t} p_{\boldsymbol{\theta}}(y_{t'}|\boldsymbol{y}_{1:t'-1}).
\end{aligned}
$$

Conditional probability: softmax

$$p_{\boldsymbol{\theta}}(y_t|\boldsymbol{y}_{1:l-1}) = \frac{\exp\left(\phi_{\boldsymbol{\theta}}(y_t|\boldsymbol{y}_{1:t-1})/T\right)}{\sum_{y'} \exp\left(\phi_{\boldsymbol{\theta}}(y'|\boldsymbol{y}_{1:t-1})/T\right)}$$

Greedy search

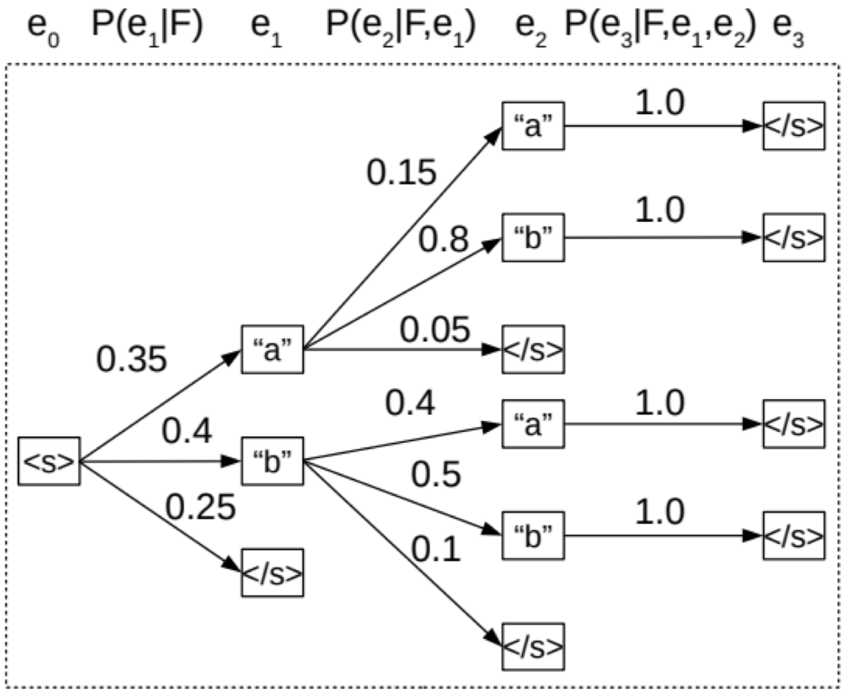


Figure 22: A search graph where greedy search fails.

In: Neural Machine Translation and Sequence-to-sequence Models: A Tutorial – Neubig 2017

Beam search

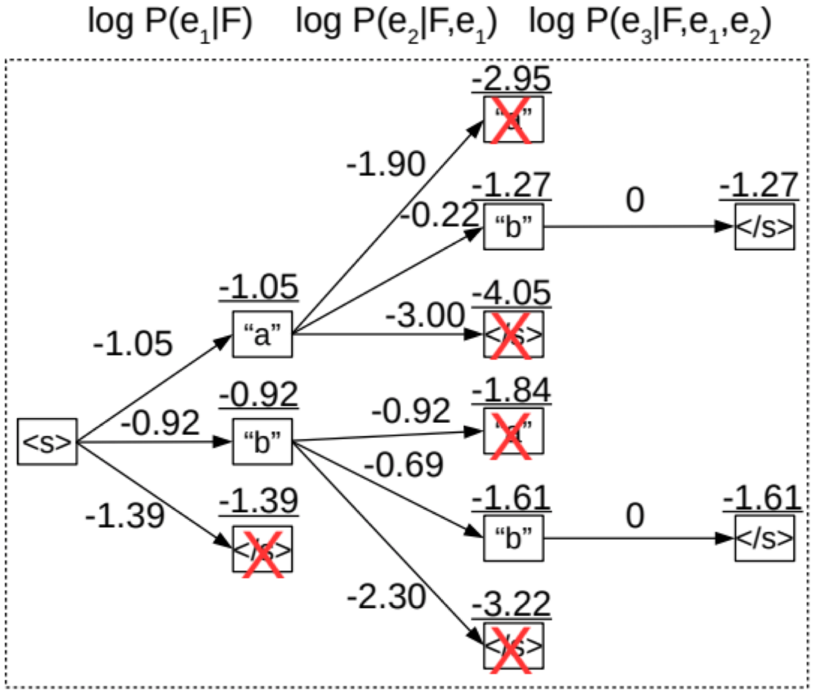


Figure 23: An example of beam search with $b = 2$. Numbers next to arrows are log probabilities for a single word $\log P(e_t \mid F, e_1^{t-1})$, while numbers above nodes are log probabilities for the entire hypothesis up until this point.

In: Neural Machine Translation and Sequence-to-sequence Models: A Tutorial – Neubig 2017

THE GUMBEL TRICK

Gumbel distribution

- Related to extreme value theory
- CDF

$$F(x;\ \phi, \beta) = \ e^{-e^{-(x-\phi)/\beta}}$$

$$F^{-1}(x; \phi, \beta) = \ \phi - \beta \log(-\log(x))$$

- If $U \sim$ Uniform(0, 1) then $G = \ \phi \ - \log(-\log(U)) \sim$ Gumbel(ϕ)
- It follows that $G' = G + \phi' \sim$ Gumbel($\phi + \phi'$)

Properties of Gumbel distribution

1 - Gumbel Trick

- Let $G_i \sim$ Gumbel(0) i.i.d. with $i \in N \subset \mathbb{N}$ and $G_{\phi_i} = \phi_i + G_i$
- Let $I^* = argmax(G_i)$
- Then $I^* \sim$ Categorical(p_i) with $p_i \propto e^{\phi_i}$

Properties of Gumbel distribution

2 - Recursion

- Let $G_i \sim$ Gumbel(0) i.i.d. with $i \in N \subset \mathbb{N}$ and $G_{\phi_i} = \phi_i + G_i$
- Then for any subset B ⊆ N
  - $\operatorname{argmax}_{i \in B} G_{\phi_i} \sim Categorical\left(\frac{e^{\phi_i}}{\sum_{i \epsilon B} e^{\phi_i}}, i \in B\right)$
  - $\max_{i \in B} G_{\phi_i} \sim Gumbel(\phi)\ with\ e^{\phi} = \sum_{i \epsilon B} e^{\phi_i}$
  - max and argmax above are independent

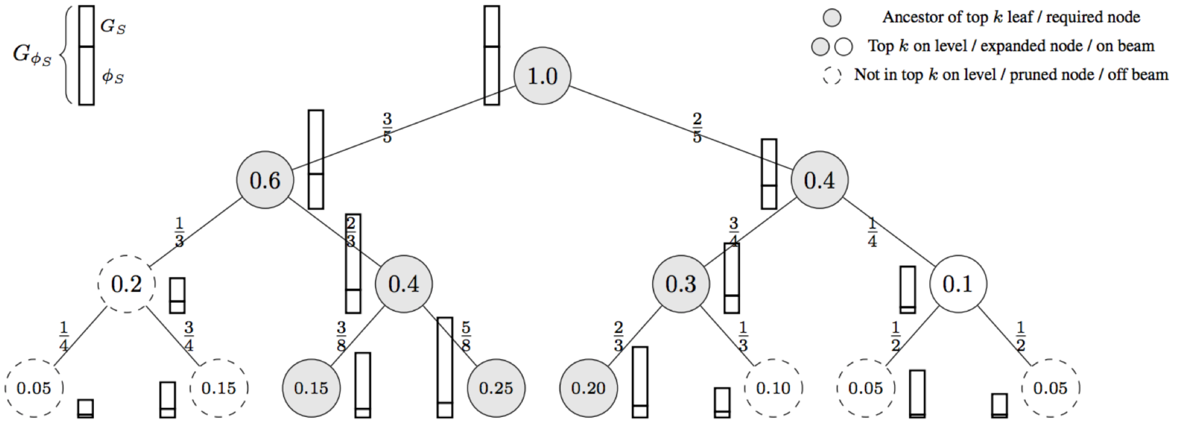
Properties of Gumbel distribution
3 – Top-k trick

- Let $G_i \sim$ Gumbel(0) i.i.d. with $i \in N \ \subset \mathbb{N}$ and $G_{\phi_i} \ = \ \phi_i + G_i$
- For k ≤ |N| let $I_1^*, \dots, I_k^* = \arg \underset{i \in N}{top}\ k\ G_{\phi_i}$
- Then $I_1^*, \dots, I_k^*$ is a sample without replacement from $Categorical\left(\frac{e^{\phi_i}}{\sum_{i \epsilon N} e^{\phi_i}}\right)$

STOCHASTIC BEAM

Representing the model as a tree

- Internal nodes at level t represent partial sequences $y_{1:t}$
- Leaves y^i represent finished sequences, $i \in N = \{1, \dots, n\}$
- Normalised log-probability of y^i is $\phi_i = \log p_\theta(y_i)$



Naive computation of full tree

- Sample $G_{\phi_i} \sim \text{Gumbel}(\phi_i)$ so that G_{ϕ_i} can be seen as log-probability of sequence y^{i}
- Let $i_1^*, \ldots, i_k^* = \arg \underset{i \in N}{top}\; k\; G_{\phi_i}$
- Then $y^{i_1^*}, \ldots, y^{i_k^*}$ is a sample of sequences without replacement following the model's probability distribution

Probability distribution of internal nodes (representing partial sequences)

- Let us consider an internal node j and its direct children $i \in C_j$, then

  $$p(\text{node } j) = \sum_{i \in C_j} p(\text{node } i)$$

  $$e^{\phi_j} = \sum_{i \in C_j} e^{\phi_i}$$

  $$\phi_j = \log\left(\sum_{i \in C_j} e^{\phi_i}\right)$$

  where Φ_i is the log probability of the node i

- If each child node $i \sim \text{Gumbel}(\phi_i)$ then:

  $$\max_{i \in C_j} G_{\phi_i} \sim \text{Gumbel}(\phi_j)$$

- It follows that this relationship holds for all nodes in the tree

Actual computation

- Top-down, starting by sampling a Gumbel distribution for the root
- Until all leaves reached:
  - For each node one level down, sample a Gumbel distribution **conditioned on its max being ≤ G-value of its parent node**
  - Once complete**, select the top-k G-values** of current level (or leaves if reached)
    - Sample follows the model distribution (independence of max and argmax)
    - Necessarily have the final top-k values in their descendants
- Final sample true to model distribution
- Cost comparable to beam search: O(kVt) vs O(V^t)for full search (V size of vocabulary)

Algorithm

Algorithm 1 StochasticBeamSearch($p_{\boldsymbol{\theta}}$, k)

Input: one-step probability distribution $p_{\boldsymbol{\theta}}$, beam/sample size k
Initialize BEAM empty
add $(\boldsymbol{y}^N = \varnothing, \phi_N = 0, G_{\phi_N} = 0)$ to BEAM
for $t = 1, \ldots,$ steps **do**
  Initialize EXPANSIONS empty
  for $(\boldsymbol{y}^S, \phi_S, G_{\phi_S}) \in$ BEAM **do**
    $Z \leftarrow -\infty$
    for $S' \in \text{Children}(S)$ **do**
      $\phi_{S'} \leftarrow \phi_S + \log p_{\boldsymbol{\theta}}(\boldsymbol{y}^{S'} | \boldsymbol{y}^S)$
      $G_{\phi_{S'}} \sim \text{Gumbel}(\phi_{S'})$
      $Z \leftarrow \max(Z, G_{\phi_{S'}})$
    end for
    for $S' \in \text{Children}(S)$ **do**
      $\tilde{G}_{\phi_{S'}} \leftarrow -\log(\exp(-G_{\phi_S}) - \exp(-Z) + \exp(-G_{\phi_{S'}}))$
      add $(\boldsymbol{y}^{S'}, \phi_{S'}, \tilde{G}_{\phi_{S'}})$ to EXPANSIONS
    end for
  end for
  BEAM $\leftarrow$ take top k of EXPANSIONS according to $\tilde{G}$
end for
Return BEAM

THAT'S ALL FOR NOW

In: Les Shadoks, Rouxel et al. 1968