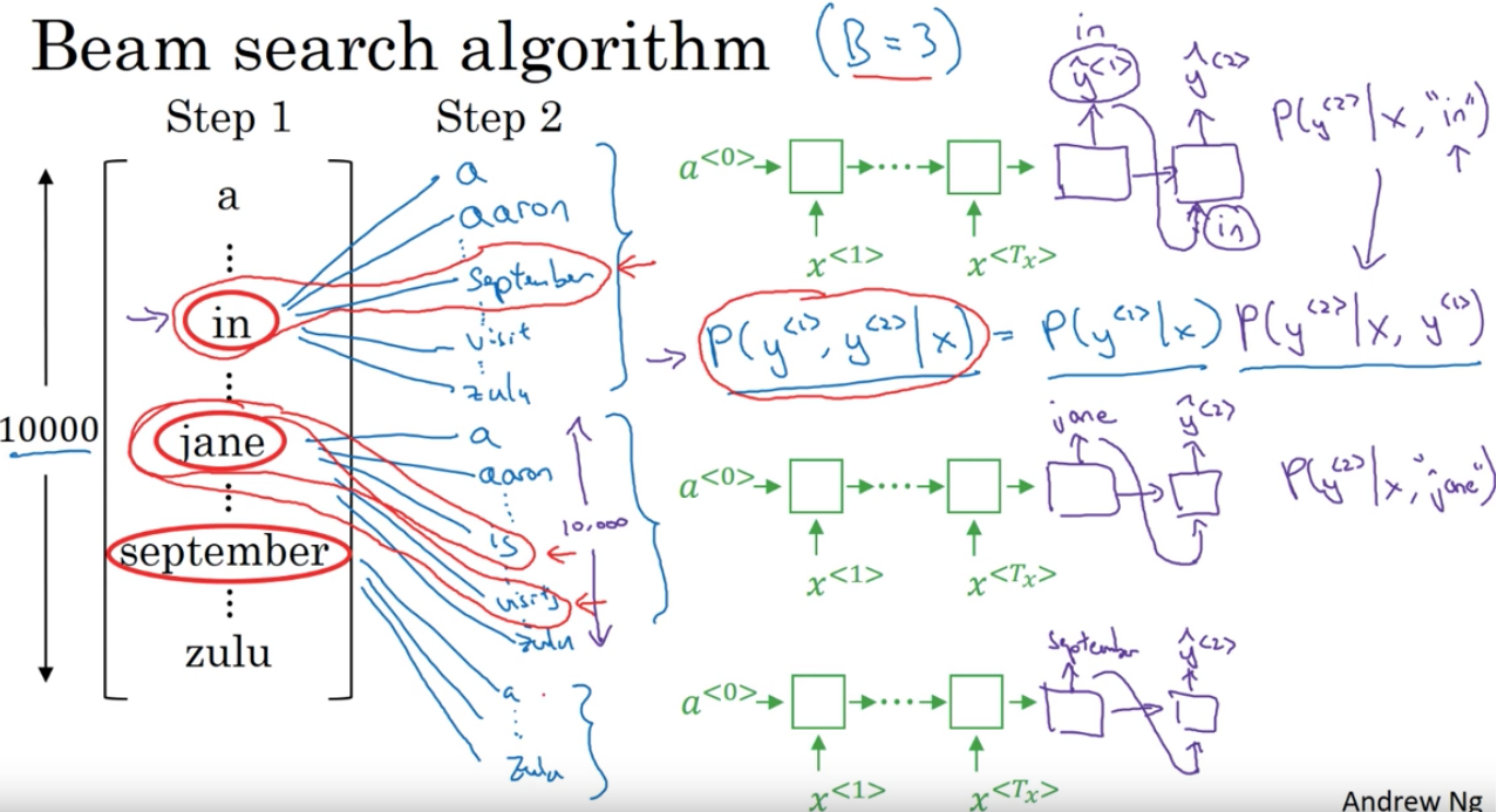
Stochastic Beams and Where to Find Them: The Gumbel-Top-k Trick for Sampling Sequences Without Replacement

Wouter Kool Herke van Hoof

Max Welling

Presenter: Yevgeny Tkach

2019 Spring @
https://qdata.github.io/deep2Read/



Source: C5W3L03 Beam Search
https://www.youtube.com/watch?v=RLWuzLLSIgw

Problem definition

- This paper defines *Stochastic Beam Search (SBM)*
- The main issue in defining an SBM is how to create a *soft commit* to a sampling 'decision' made at step *t.*
- More precisely, what is $P(y^{\langle 1\rangle}, y^{\langle 2\rangle}|x)$ in the sequential sampling procedure?
- Assuming that $y^{\langle 1\rangle}$ initially a low probability to be sampled $P(y^{\langle 1\rangle}|x)$.
- So the naive use of defining $P(y^{\langle 1\rangle}, y^{\langle 2\rangle}|x) = P(y^{\langle 1\rangle}|x)P(y^{\langle 2\rangle}\big|x, y^{\langle 1\rangle})$ will mean that sequences with initial low probability will actually have much lower probability to be repeatedly sampled in the *SBM.*

Executive Summary

- For *Stochastic Beam Search (SBM)* of width *k*, at each step of the decision sequence, *k* ‘decisions’ are sampled using Gumbel-Top-k trick
- SBM sequential sampling procedure is unbiased and equivalent to sample Top-k sequences from the complete ‘decisions’ tree
- In a translation task SBM obtains more diverse and higher quality translations then other inference time methods
- SBM can be used to construct low-variance estimators for expected sentence-level BLEU score and model entropy

Outline

- Gumbel-Top-k trick
- Stochastic Beam Search (SBM) algorithm
- Unbiased SBM
- Experiments
- Conclusions

Gumbel-Top-k Trick

- Sampling from a discrete distribution parametrized by unnormalized log-probabilities: $\pi_k = \frac{1}{z}\exp(x_k)$ where $z = \sum_{j=1}^{k}\exp(x_j)$
- The *Gumbel(0)=–log(–log(Uniform(0,1))*
- The Gumbel-Max Trick:

$$y = \arg\max_{i\in\{1,\dots,k\}} G_{\varphi_i} \sim \pi, \quad G_{\varphi_i} = G_i + \varphi_i \sim Gumbel(\varphi_i), \quad G_i \sim Gumbel(0)$$

- Similarly, the Gumbel-Top-k Trick:

Theorem 1. *For $k \leq n$, let $I_1^*, \dots, I_k^* = \arg\operatorname{top} k\, G_{\phi_i}$. Then $I_1^*, \dots, I_k^*$ is an (ordered) sample without replacement from the Categorical $\left(\frac{\exp \phi_i}{\sum_{j\in N}\exp\phi_j}, i \in N\right)$ distribution, e.g. for a realization $i_1^*, \dots, i_k^*$ it holds that*

$$P\left(I_1^* = i_1^*, \dots, I_k^* = i_k^*\right) = \prod_{j=1}^{k} \frac{\exp \phi_{i_j^*}}{\sum_{\ell \in N_j^*} \exp \phi_\ell} \quad (4)$$

where $N_j^* = N \setminus \{i_1^*, \dots, i_{j-1}^*\}$ is the domain (without replacement) for the j-th sampled element.

SBM 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 $t = 1, \dots,$ 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

Pure Magic

Unbiased SBM

- At *t=1,* the Gumbel-Top-k trick works directly and a beam of width k is sampled with probability $Categorical\left(\frac{\exp \phi_i}{\sum_{j \in N} \exp \phi_j}, i \in N\right)$ where $\phi_i = \log p_\theta(y^i|x)$
- At *t=k>1*, the following condition needs to hold $G_{\phi_S} = \max_{S' \in \text{Children}(S)} G_{\phi_{S'}}$
- We need to sample a set of Gumble variables $\{\tilde{G}_{\phi_i} | \max_i \tilde{G}_{\phi_i} = T\}$ with the following procedure:

1. Sample $i^* \sim \text{Categorical}\left(\frac{\exp \phi_i}{\sum_j \exp \phi_j}\right)$. We do not need to condition on T since the arg max i^* is independent of the max T (Section 2.3).
2. Set $\tilde{G}_{\phi_{i^*}} = T$, since this follows from conditioning on the max T and arg max i^*.
3. Sample $\tilde{G}_{\phi_i} \sim \text{TruncatedGumbel}(\phi_i, T)$ for $i \neq i^*$.

Unbiased SBM

A random variable G' has a *truncated* Gumbel distribution with location ϕ and maximum T (e.g. $G' \sim \text{TruncatedGumbel}(\phi, T)$) with CDF $F_{\phi,T}(g)$ if:

$$\begin{aligned}
&F_{\phi,T}(g) \\
&= P(G' \leq g) \\
&= P(G \leq g | G \leq T) \\
&= \frac{P(G \leq g \cap G \leq T)}{P(G \leq T)} \\
&= \frac{P(G \leq \min(g, T))}{P(G \leq T)} \\
&= \frac{F_\phi(\min(g, T))}{F_\phi(T)} \\
&= \frac{\exp(-\exp(\phi - \min(g, T)))}{\exp(-\exp(\phi - T))} \\
&= \exp(\exp(\phi - T) - \exp(\phi - \min(g, T))). \qquad (20)
\end{aligned}$$

The inverse CDF is:

$$F_{\phi,T}^{-1}(u) = \phi - \log(\exp(\phi - T) - \log u). \qquad (21)$$

3. Sample $\tilde{G}_{\phi_i} \sim \text{TruncatedGumbel}(\phi_i, T)$ for $i \neq i^*$. This works because, conditioning on the max T and arg max i^*, it holds that:

$$\begin{aligned}
&P(\tilde{G}_{\phi_i} < g | \max_i \tilde{G}_{\phi_i} = T, \arg\max_i \tilde{G}_{\phi_i} = i^*, i \neq i^*) \\
&= P(\tilde{G}_{\phi_i} < g | \tilde{G}_{\phi_i} < T).
\end{aligned}$$

Equivalently, we can let $G_{\phi_i} \sim \text{Gumbel}(\phi_i)$, let $Z = \max_i G_{\phi_i}$ and define

$$\begin{aligned}
\tilde{G}_{\phi_i} &= F_{\phi_i,T}^{-1}(F_{\phi_i,Z}(G_{\phi_i})) \\
&= \phi_i - \log(\exp(\phi_i - T) \\
&\qquad - \exp(\phi_i - Z) + \exp(\phi_i - G_{\phi_i})) \\
&= -\log(\exp(-T) - \exp(-Z) + \exp(-G_{\phi_i})).
\end{aligned} \qquad (22)$$

Here we have used (20) and (21). Since the transformation (22) is monotonically increasing, it preserves the arg max and it follows from the Gumbel-Max trick (3) that

$$\arg\max_i \tilde{G}_{\phi_i} = \arg\max_i G_{\phi_i} \sim \text{Categorical}\left(\frac{\exp \phi_i}{\sum_j \exp \phi_j}\right).$$

We can think of this as using the Gumbel-Max trick for step 1 (sampling the argmax) in the sampling process described above. Additionally, for $i = \arg\max_i G_{\phi_i}$:

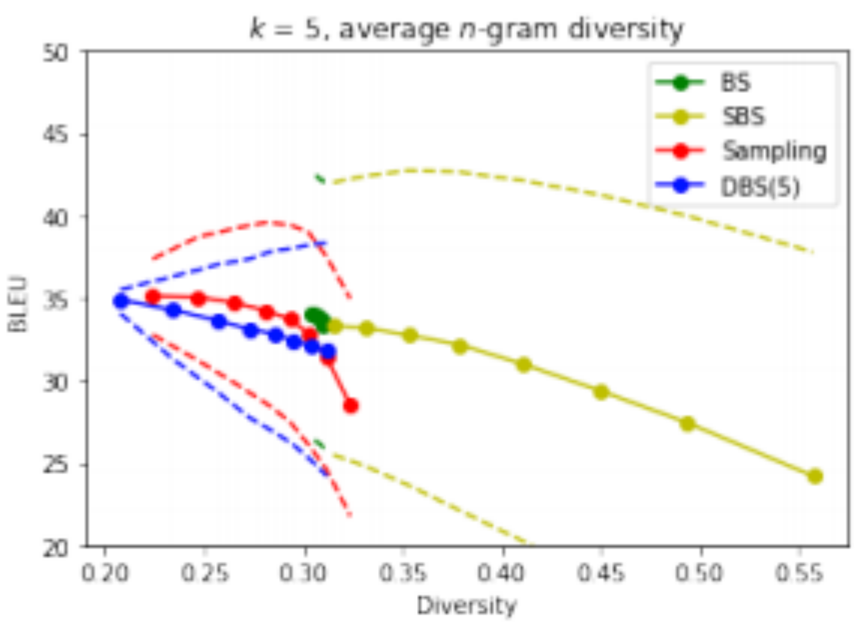
$$\tilde{G}_{\phi_i} = F_{\phi_i,T}^{-1}(F_{\phi_i,Z}(G_{\phi_i})) = F_{\phi_i,T}^{-1}(F_{\phi_i,Z}(Z)) = T$$

Experiments

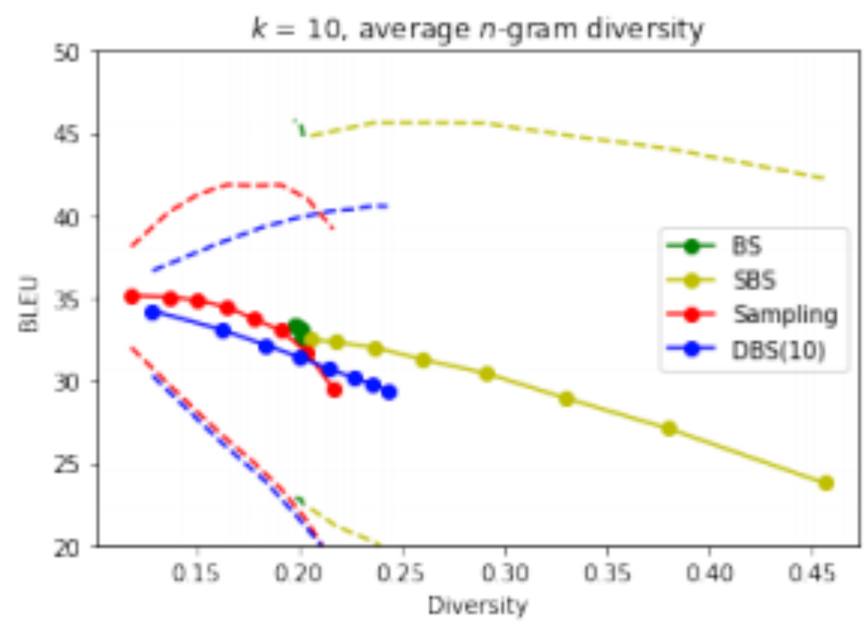
- **Diverse Beam Search -** The task, in the context of neural machine translation, is to obtain a diverse set of translations for a single source sentence x.
- **BLEU score estimation -** The task is to evaluate the expected sentence level BLEU score for a translation y given a source sentence x, by sampling without replacement different translations
- **Conditional Entropy estimation -** Similar to the BLEY score estimation above

Diverse Beam Search

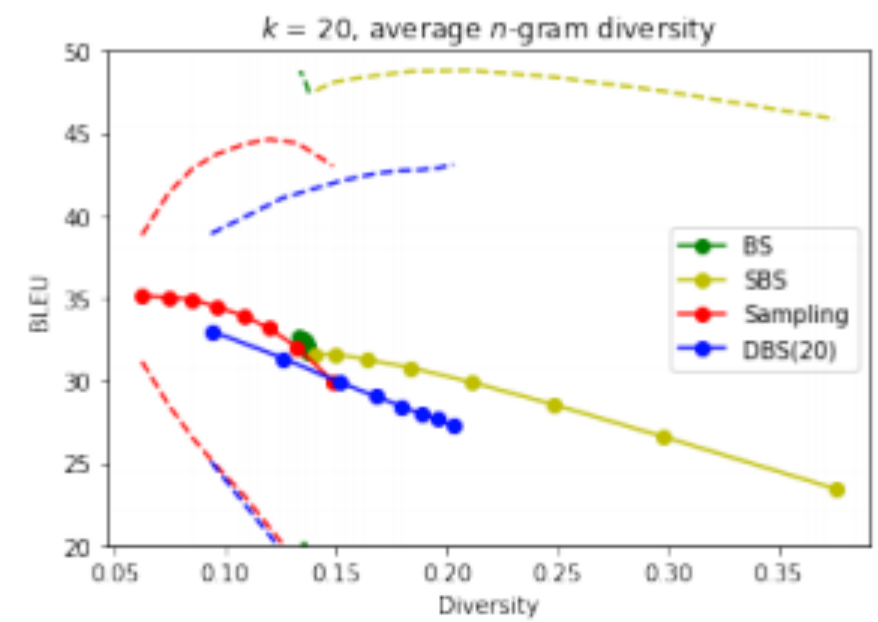
- Experiments are run against Beam Search (BS), Sampling, Stochastic Beam Search(SBS) (sampling without replacement) and Diverse BeamSearch with G groups (DBS(G))
- Average n-gram diversity is defined as: $$d_n = \frac{\text{\# of unique } n\text{-grams in } k \text{ translations}}{\text{total \# of } n\text{-grams in } k \text{ translations}}$$



(a) $k = 5$



(b) $k = 10$



(c) $k = 20$

BLEU Score estimation

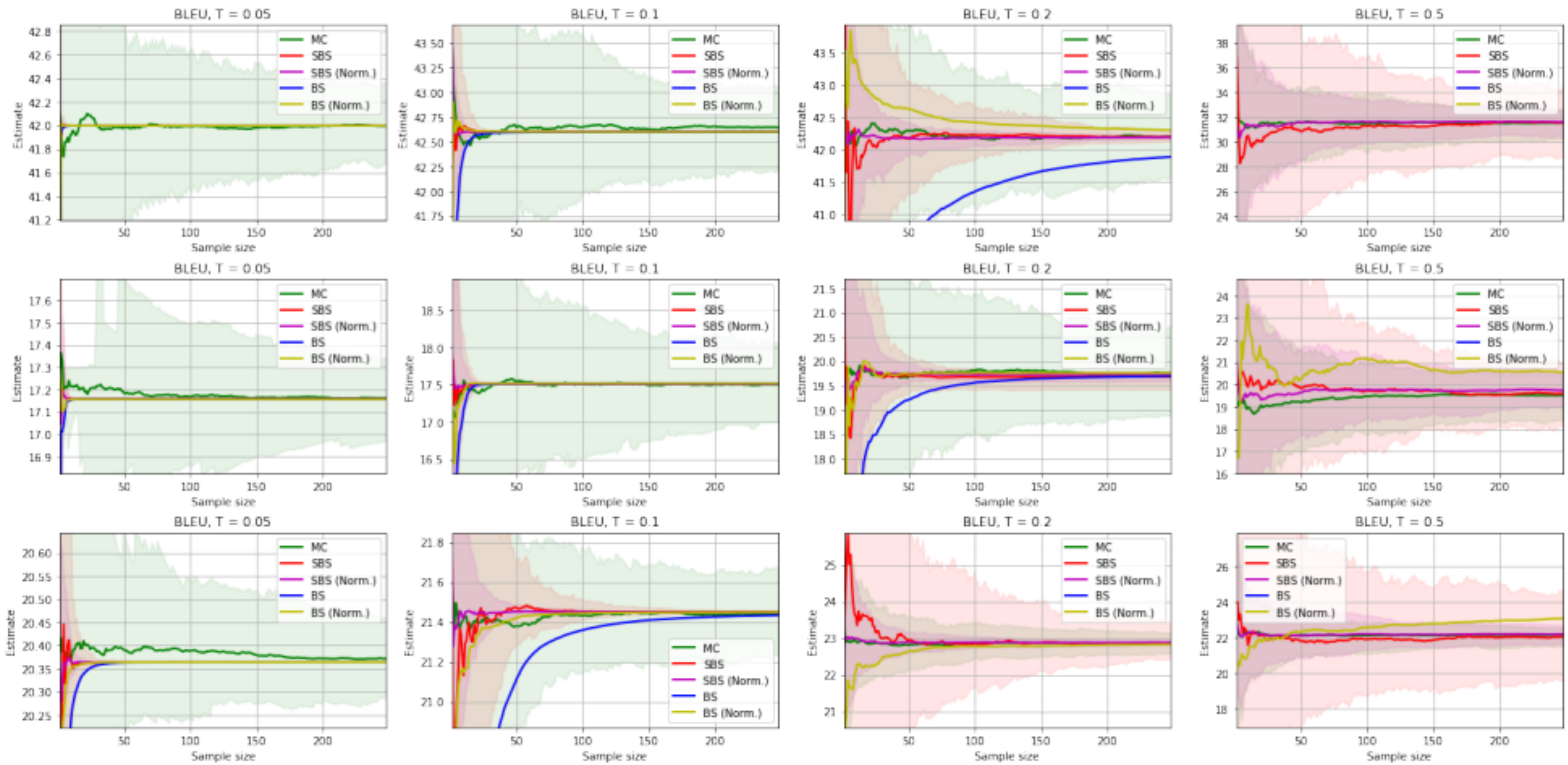


Figure 3. BLEU score estimates for three sentences sampled/decoded by different estimators for different temperatures.

Conditional Entropy estimation

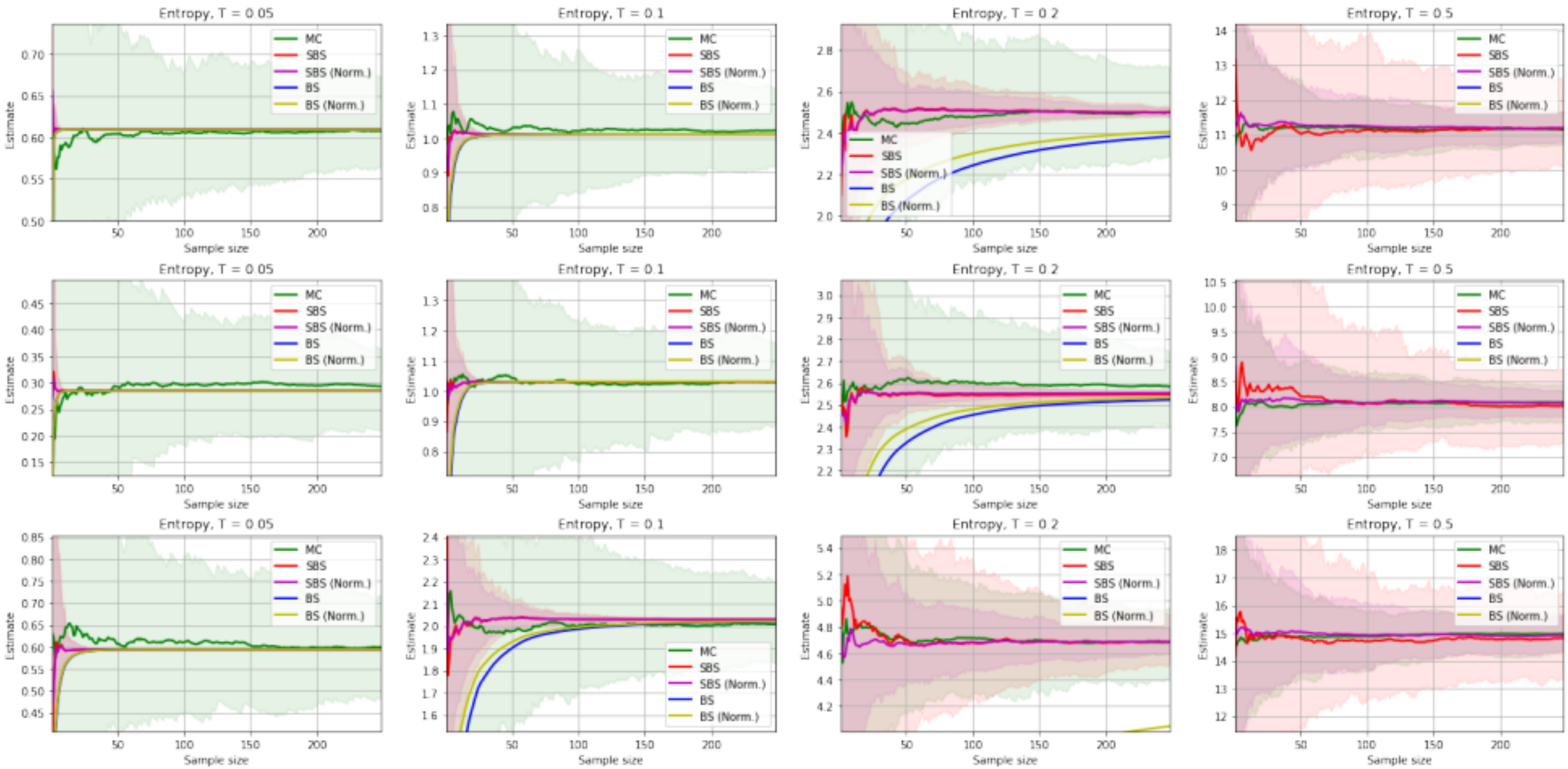


Figure 4. Entropy score estimates for three sentences sampled/decoded by different estimators for different temperatures.

Discussion

- Stochastic Beam Search is a powerful novel technique that offers unbiased sampling of top-K candidates without calculating the complete ‘decisions’ tree
- This also works as a good sampling technique since with a good choice of k, the top-k choices may offer a good estimate to the probability mass
- This approach can also be leveraged as a RL technique