A Continuous Relaxation of Beam Search for End-to-end Training of Neural Sequence Models

Kartik Goyal, Graham Neubig, Chris Dyer, Taylor Berg-Kirkpatrick
CMU & Deep Mind – AAAI18

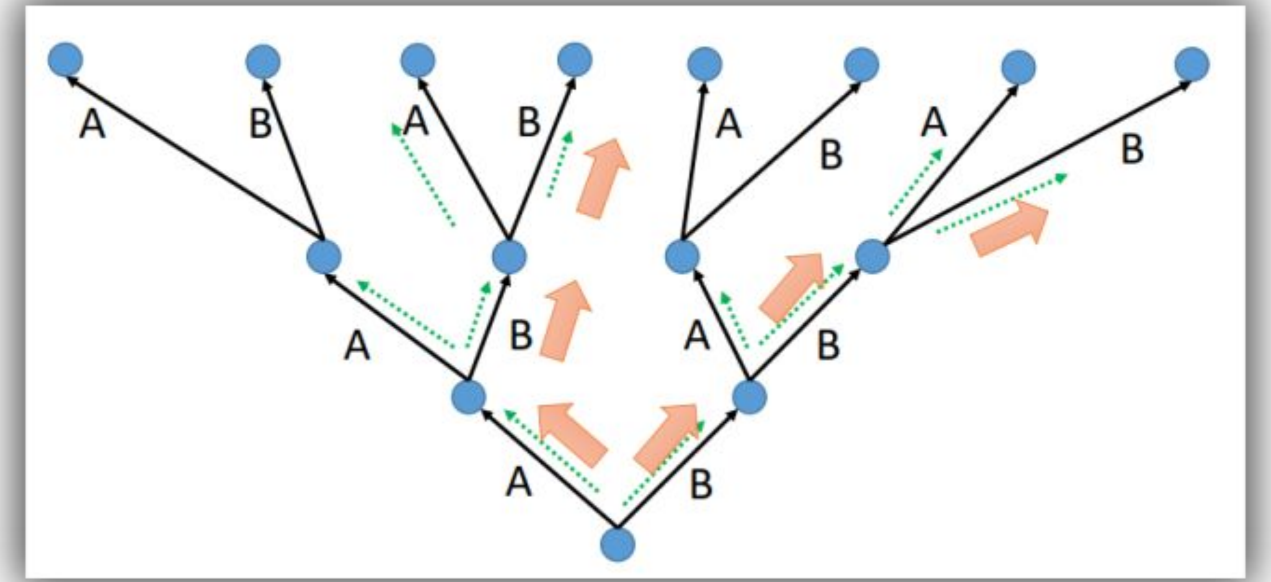
AntNLP – Tao Ji

taoji.cs@gmail.com

Outline

- Motivation
- Standard Beam Search (BS)
- Discontinuity in BS
- Continuous Approximation to BS
- Training & Decoding
- Comparison with Max-Margin Objectives
- Experiments and Results
- Discussions

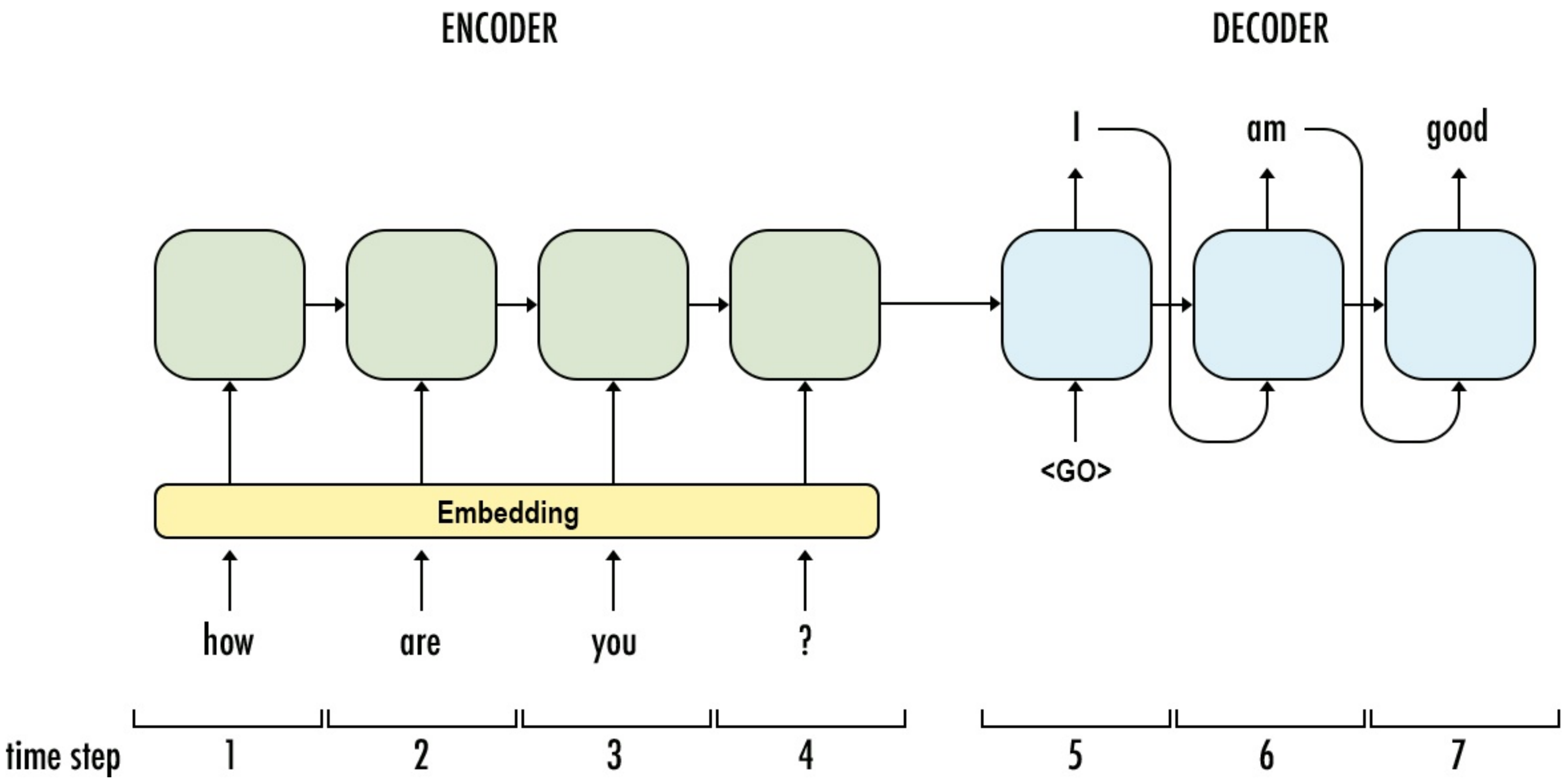
Beam Search Example



Beam = 2

- All search branches
- Beam Search extended branches
- Beam Search selected branches

Neural Sequence Model



Motivation

Advantage

- Potentially avoids search errors made by simpler greedy methods.

Problem

- Training procedures don’t consider the behavior of the decoding.
- “Direct BS loss” objective is discontinuous and difficult to optimize.

Approach

- Form a sub-differentiable surrogate objective by introducing a novel continuous approximation of the beam search decoding procedure.

Model Definition

- $\mathcal{M}(\theta)$ denote the seq2seq model parameterized.
- Assume : $L(\hat{y}, y^*) = \sum_{t=1}^{T} d(\hat{y}_t, y^*)$

$$\min_{\theta} G_{\mathrm{DL}}(x, \theta, y^*) = \min_{\theta} L(\mathit{Beam}(x, \mathcal{M}(\theta)), y^*) \quad (1)$$

$$\min_{\theta} \tilde{G}_{\mathrm{DL}}(x, \theta, y^*) = \min_{\theta} \mathit{softLB}(x, \mathcal{M}(\theta), y^*) \quad (2)$$

Standard Beam Search

Algorithm 1 Standard Beam Search

Initialize:
    $h_{0,i} \leftarrow \vec{0}$, $e_{0,i} \leftarrow \textit{embedding}(\text{<s>})$, $s_{0,i} \leftarrow 0$, $i = 1, \ldots, k$
for t = 0 to T **do**
    for i = 1 to k **do**
        for all $v \in V$ **do**
            $\tilde{s}_t[i, v] \leftarrow s_{t,i} + f(h_{t,i}, v)$ ▷ f is the local output scoring function
    $s_{t+1} \leftarrow \textit{top-k-max}(\tilde{s}_t)$ ▷ Top k values of the input matrix
    $b_{t+1,*}, y_{t,*} \leftarrow \textit{top-k-argmax}(\tilde{s}_t)$ ▷ Top k argmax index pairs of the input matrix
    for i = 1 to k **do**
        $e_{t+1,i} \leftarrow \textit{embedding}(y_{t,i})$
        $h_{t+1,i} \leftarrow r(h_{t,i}, e_{t+1,i})$ ▷ r is a nonlinear recurrent function that returns state at next step
$\hat{y} \leftarrow \textit{follow-backpointer}((b_{1,*}, y_{1,*}), \ldots, (b_{T,*}, y_{T,*}))$
$s(\hat{y}) \leftarrow \textit{max}(s_T)$

Discontinuity in BS

- Beam search decoding (referred to as the function Beam) involves discrete argmax decisions and thus represents a discontinuous function.

- The output of the Beam function, which is the input to the loss function is discrete and hence the evaluation of the final loss is also discontinuous.

Continuous Approximation to argmax

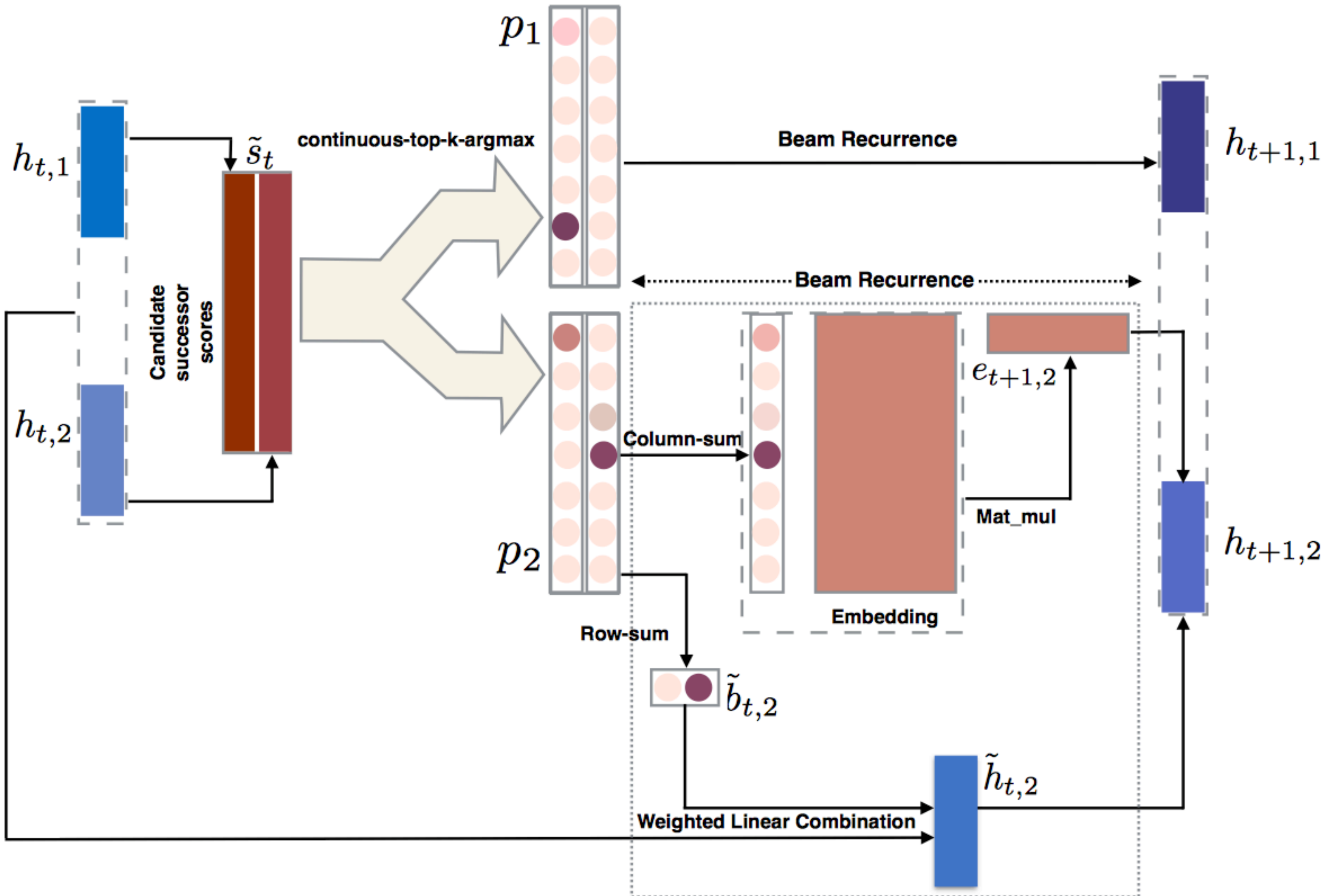
Algorithm 2 continuous-top-k-argmax

1: **Inputs:**
   $s \in \mathbb{R}^{k \times |V|}$
2: **Outputs:**
   $p_i \in \mathbb{R}^{k \times |V|}$, s.t. $\sum_j p_{ij} = 1, i = 1, \ldots, k$
3: $m \in \mathbb{R}^k = \textit{top-k-max}(s)$
4: **for** i = 1 to k **do** ▷ *peaked-softmax* will be dominated by scores closer to m_i
5: $p_i = \textit{peaked-softmax}_\alpha(-(s - m_i \cdot \mathbf{1})^2)$ ▷ The square operation is element-wise

$$\hat{z} = \sum_i z_i \mathbb{1}[\forall i' \neq i, \;\; s_i > s_{i'}], \qquad \tilde{z} = \sum_i z_i \frac{\exp(\alpha \; s_i)}{\sum_{i'} \exp(\alpha \; s_{i'})} = z^T \cdot \frac{\text{elem-exp}(\alpha \; s)}{\sum_{i'} \exp(\alpha \; s_{i'})}$$

$$= z^T \cdot \textit{peaked-softmax}_\alpha(\mathbf{s})$$

Continuous Approximation to BS



Continuous Approximation to BS

Algorithm 3 Continuous relaxation to beam search

Initialize:

  $h_{0,i} \leftarrow \vec{0}$, $e_{0,i} \leftarrow \textit{embedding(<s>)}$, $s_{0,i} \leftarrow 0$, $D_t \in \mathbb{R}^k \leftarrow \vec{0}$, $i = 1, \ldots, k$

for t = 0 to T **do**

  for all $w \in V$ **do**

    for i=1 to k **do**

      $\tilde{s}_t[i, w] \leftarrow s_{t,i} + f(h_{t,i}, w)$ ▷ f is a local output scoring function

      $\tilde{D}_{t,w} = d(w)$ ▷ $\tilde{D}_t$ is used to compute D_{t+1}

  $p_1, \ldots, p_k \leftarrow \textit{continuous-top-k-argmax}(\tilde{s}_t)$ ▷ Call Algorithm 2

  for i = 1 to k **do**

    $\tilde{b}_{t,i} \leftarrow \textit{row\_sum}(p_i)$ ▷ Soft back pointer computation

    $a_i \in \mathcal{R}^{|V|} \leftarrow \textit{column\_sum}(p_i)$ ▷ Contribution from vocabulary items

    $e_{t+1,i} \leftarrow a_i^T \times E$ ▷ Peaked distribution over the candidates to compute e, D, S

    $D_{t+1,i} \leftarrow a_i^T \cdot \tilde{D}_t$

    $s_{t+1,i} = \textit{sum}(\tilde{s}_t \odot p_i)$

    $\tilde{h}_{t,i} \leftarrow \vec{0}$

    for j = 1 to k **do** ▷ Get contributions from soft backpointers for each beam element

      $\tilde{h}_{t,i} += h_{t,j} * \tilde{b}_{t,i}[j]$

      $D_{t+1,i} += D_{t,j} * \tilde{b}_{t,i}[j]$

    $h_{t+1,i} \leftarrow r(\tilde{h}_{t,i}, e_{t+1,i})$ ▷ r is a nonlinear recurrent function that returns state at next step

$L = \textit{peaked-softmax}_{\alpha}(s_T) \cdot D_T$ ▷ Pick the loss for the sequence with highest model score on the beam in a soft manner.

Training & Decoding

Training

$$\tilde{G}_{\mathrm{DL},\alpha}(x, \mathcal{M}(\theta), y^*) \xrightarrow[p]{\alpha \to \infty} G_{\mathrm{DL}}(x, \theta, y^*) \qquad (3)$$

- Starting with non-peaked softmax moving toward peaked-softmax across epochs.

Decoding

- soft beam search
- hard beam search

Comparison with Max-Margin Objectives

$$G_{hinge} = \max(0, \max_{y \in \mathcal{Y}}(\Delta(y, y^*) + s(y)) - s(y^*))$$

$$\tilde{s}_t[i, w] \leftarrow s_{t,i} + d(w) + f(h_{t,i}, w)$$

$$s_{max} = \textit{peaked-softmax}_{\alpha}(s_T) \cdot s_T$$

$$\tilde{G}_{\text{hinge},\alpha} = \max(0, s_{max} - s(y^*))$$

Experiments and Results

- **CCG Supertagging**
- The output vocabulary length (label space) is 1284.
- Beam size = 3

Experiments and Results

Training procedure	Greedy		Hard Beam Search		Soft Beam Search	
	Dev	Test	Dev	Test	Dev	Test
Baseline CE	80.15	80.35	82.17	82.42	81.62	82.00
$\tilde{G}_{\text{hinge},\alpha}$ annealed α	-	-	83.03	83.54	82.82	83.05
$\tilde{G}_{\text{hinge},\alpha} \alpha=1.0$	-	-	83.02	83.36	82.49	82.85
$\tilde{G}_{\text{DL},\alpha} \alpha=1.0$	-	-	83.23	82.65	82.58	82.82
$\tilde{G}_{\text{DL},\alpha}$ annealed α	-	-	**85.69**	**85.82**	85.58	85.78

Experiments and Results

Named Entity Recognition

- The output vocabulary length (label space) is 10.
- Beam size = 3

Experiments and Results

Training procedure	CE Greedy		Hard Beam Search		Soft Beam Search	
	Dev	Test	Dev	Test	Dev	Test
Baseline CE	50.21	54.92	46.22	51.34	47.50	52.78
$\tilde{G}_{\text{hinge},\alpha}$ annealed α	-	-	41.10	45.98	41.24	46.34
$\tilde{G}_{\text{hinge},\alpha}$ α=1.0	-	-	40.09	44.67	39.67	43.82
$\tilde{G}_{\text{DL},\alpha}$ α=1.0	-	-	49.88	54.08	50.73	54.77
$\tilde{G}_{\text{DL},\alpha}$ annealed α	-	-	51.86	56.15	**51.96**	**56.38**

Q&A