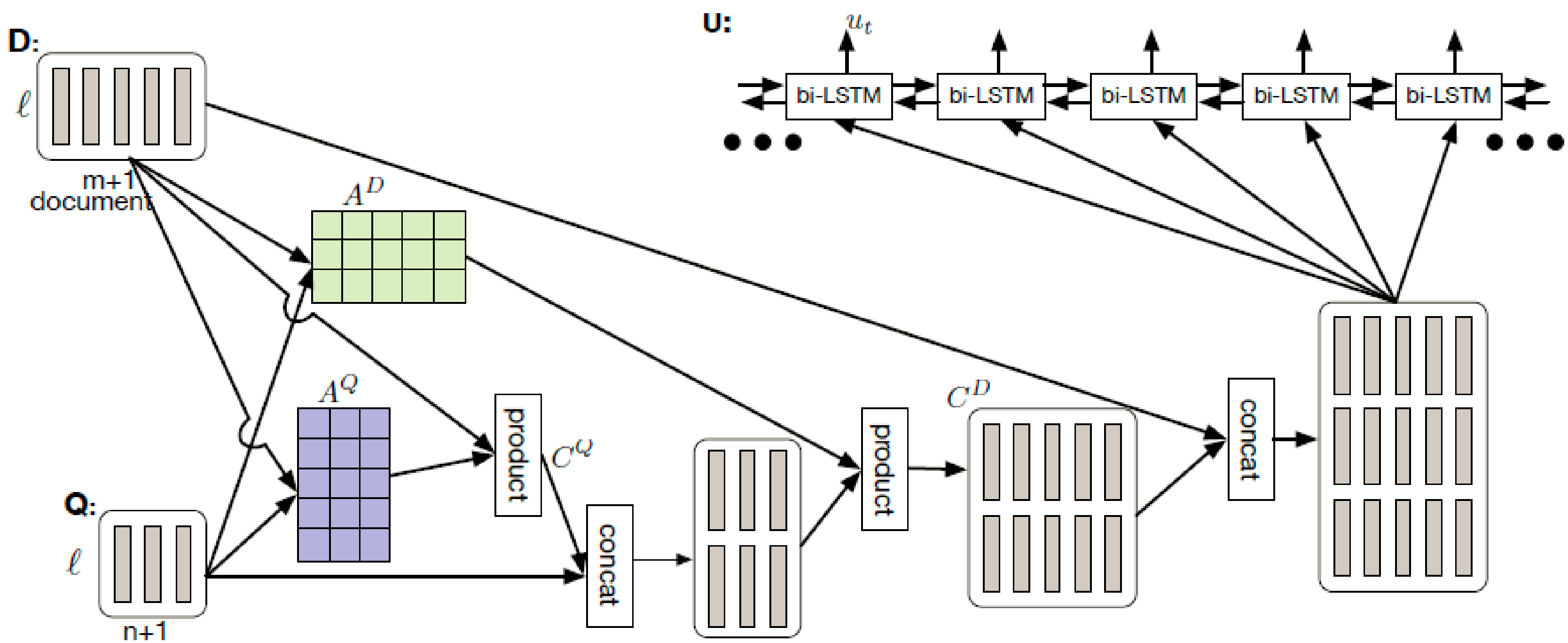
DCN+

Caiming Xiong, Victor Zhong, Richard Socher

Baseline DCN



Baseline DCN

$L = D^T Q \in R^{(m+1)*(n+1)}$

$A^Q = softmax(L) \in R^{(m+1)*(n+1)}$ and

$A^D = softmax(L^T) \in R^{(n+1)*(m+1)}$

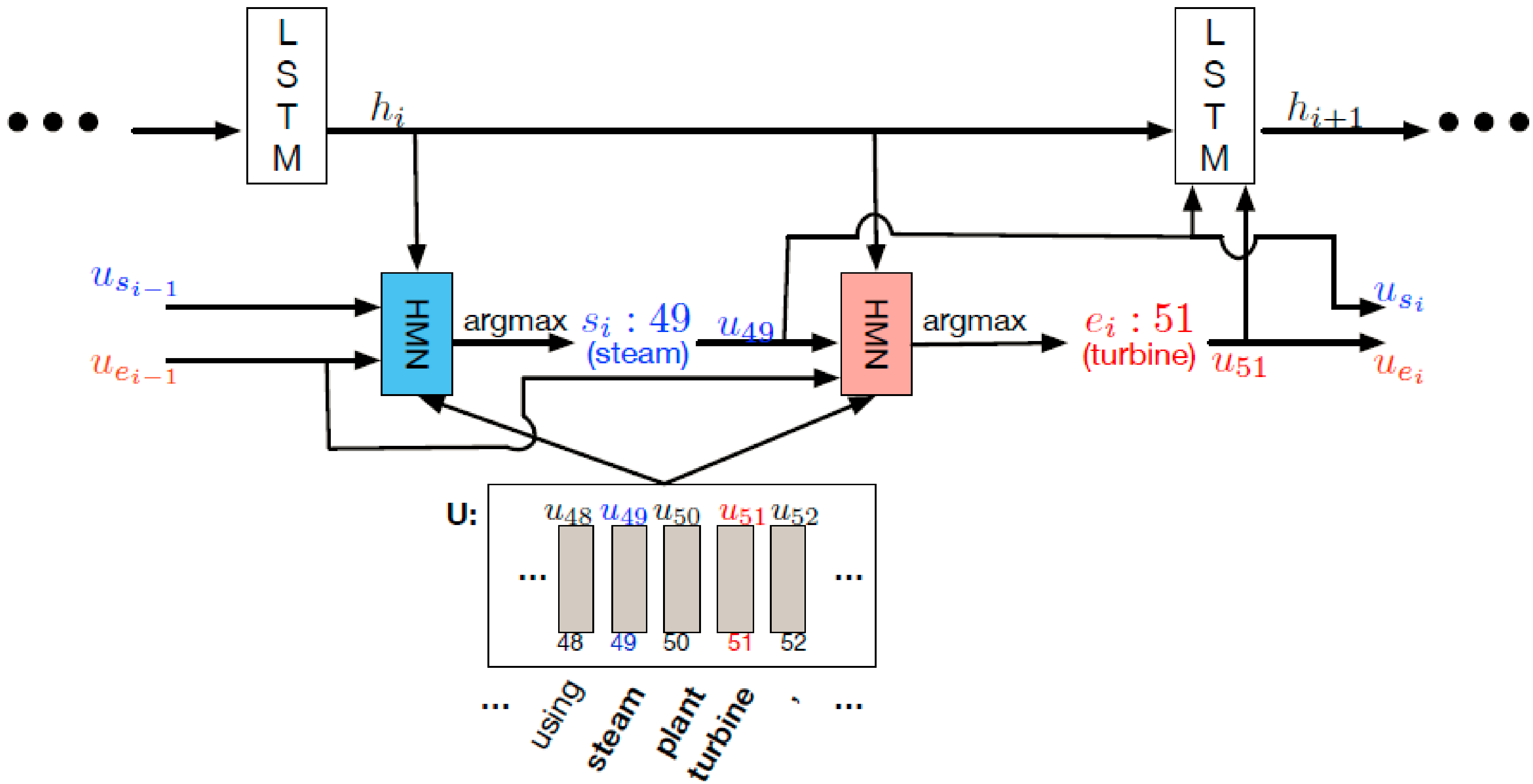
$C^Q = DA^Q \in R^{l*(n+1)}$

$C^D = [Q; C^Q]A^D \in R^{2l*(m+1)}$

$u_t = BiLSTM(u_{t-1}, u_{t+1}, [d_t; c_t^D]) \in R^{2l}$

Baseline DCN

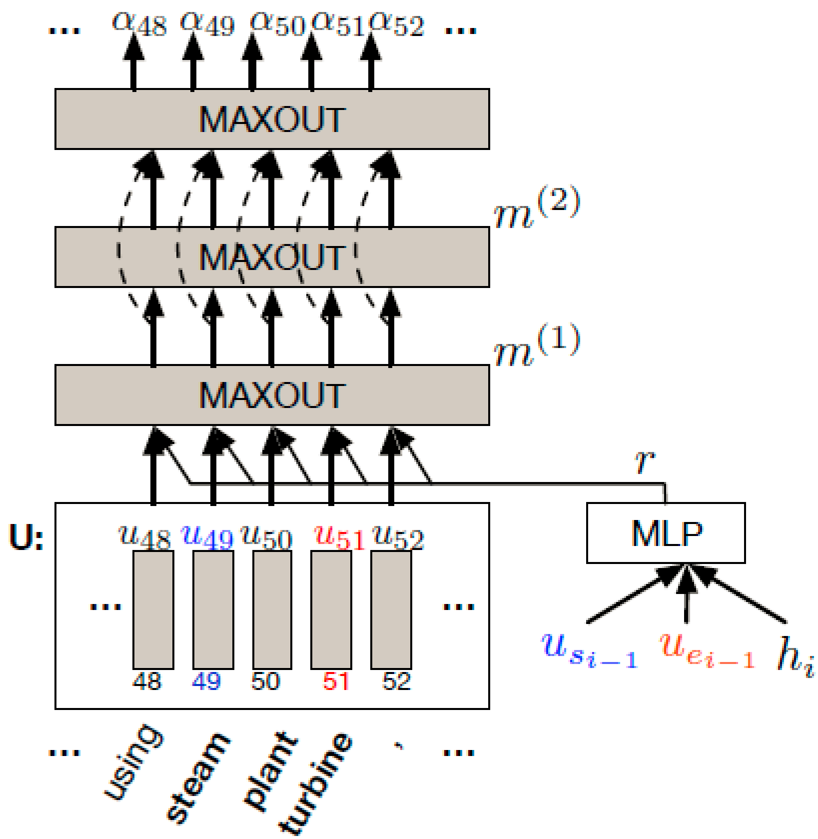
Dynamic Decoder



Dynamic Decoder

$$h_i = LSTM_{dec}(h_{i-1}, [u_{s_{i-1}}; u_{e_{i-1}}])$$
$$s_i = argmax_t(\alpha_1, ..., \alpha_m)$$
$$e_i = argmax_t(\beta_1, ..., \beta_m)$$
$$\alpha_t = HMN_{start}(u_t, h_i, u_{s_{i-1}}, u_{e_{i-1}})$$

Dynamic Deocder



$$
\begin{aligned}
\mathrm{HMN}\left(u_t, h_i, u_{s_{i-1}}, u_{e_{i-1}}\right) &= \max\left(W^{(3)}\left[m_t^{(1)}; m_t^{(2)}\right] + b^{(3)}\right) \\
r &= \tanh\left(W^{(D)}\left[h_i; u_{s_{i-1}}; u_{e_{i-1}}\right]\right) \\
m_t^{(1)} &= \max\left(W^{(1)}\left[u_t; r\right] + b^{(1)}\right) \\
m_t^{(2)} &= \max\left(W^{(2)} m_t^{(1)} + b^{(2)}\right)
\end{aligned}
$$

DCN+

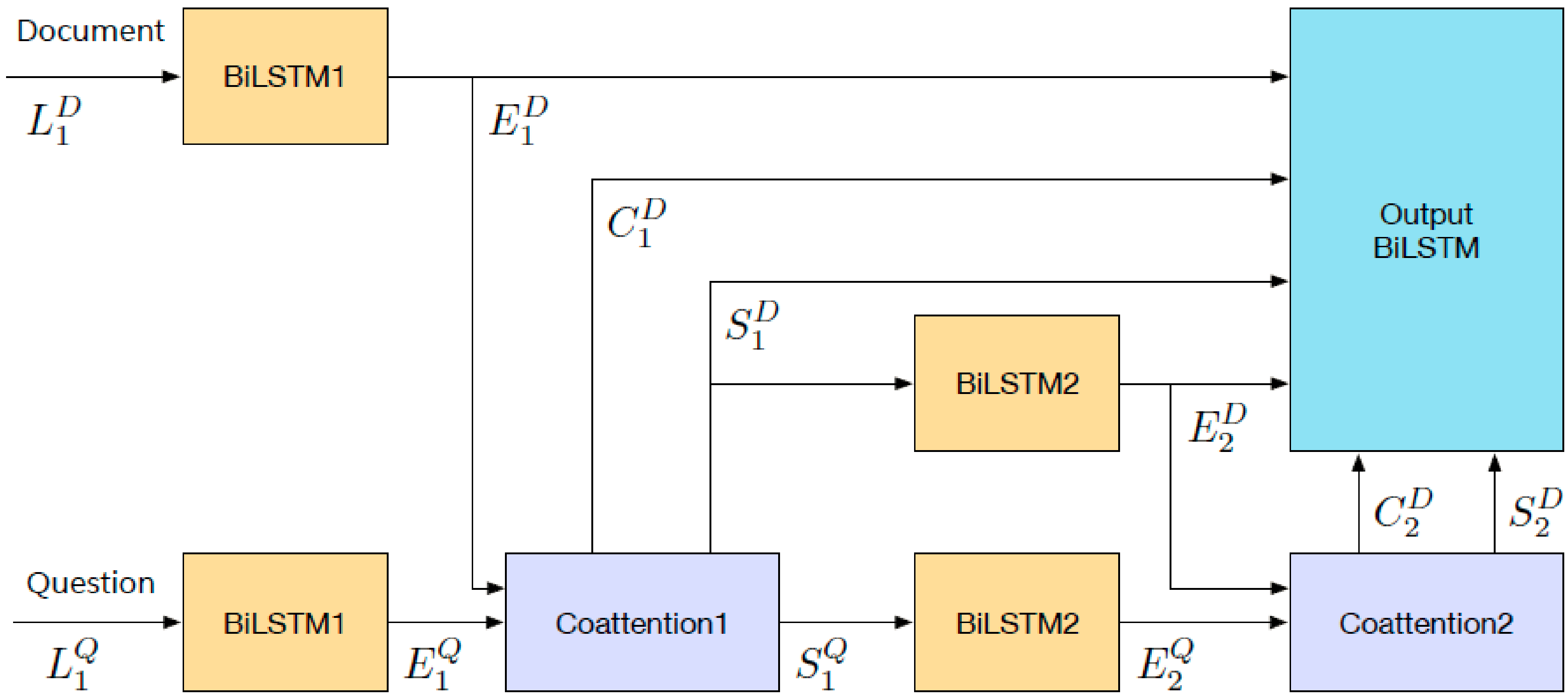


Figure 1: Deep residual coattention encoder.

First, encoding for D and Q:

$$E_1^D = biLSTM_1(L^D) \in R^{h*(m+1)}$$

$$E_1^Q = tanh(WbiLSTM_1(L^Q) + b) \in R^{h*(n+1)}$$

Secondly, counting the representation $Q2D_1$ and $D2Q_1$:

$$A = (E_1^Q)^T E_1^D \in R^{(m+1)*(n+1)}$$

$$S_1^D = E_1^D softmax(A^T) \in R^{h*(m+1)}$$

$$S_1^Q = E_1^D softmax(A) \in R^{h*(n+1)}$$

Thirdly, counting the $(D2Q_1)2D$ representation:

$$C_1^D = S_1^Q softmax(A^T) \in R^{h*m}$$

Fourthly,counting the second LSTM encoding for S_1^D and S_1^Q:

$$E_2^D = biLSTM_2(S_1^D) \in R^{2h*m}$$

$$E_2^Q = biLSTM_2(S_1^Q) \in R^{2h*n}$$

Last, doing the same coattention operation for E_2^D and E_2^Q.

DCN+

In summary,

$$coattn_1(E_1^D, E_1^Q) -> S_1^D, S_1^Q, C_1^D$$
$$coattn_2(E_2^D, E_2^Q) -> S_2^D, S_2^Q, C_2^D$$

The output of encoder is:

$$U = biLSTM(concat(E_1^D; E_2^D; S_1^D; S_2^D; C_1^D; C_2^D)) \in R^{2h*m}$$

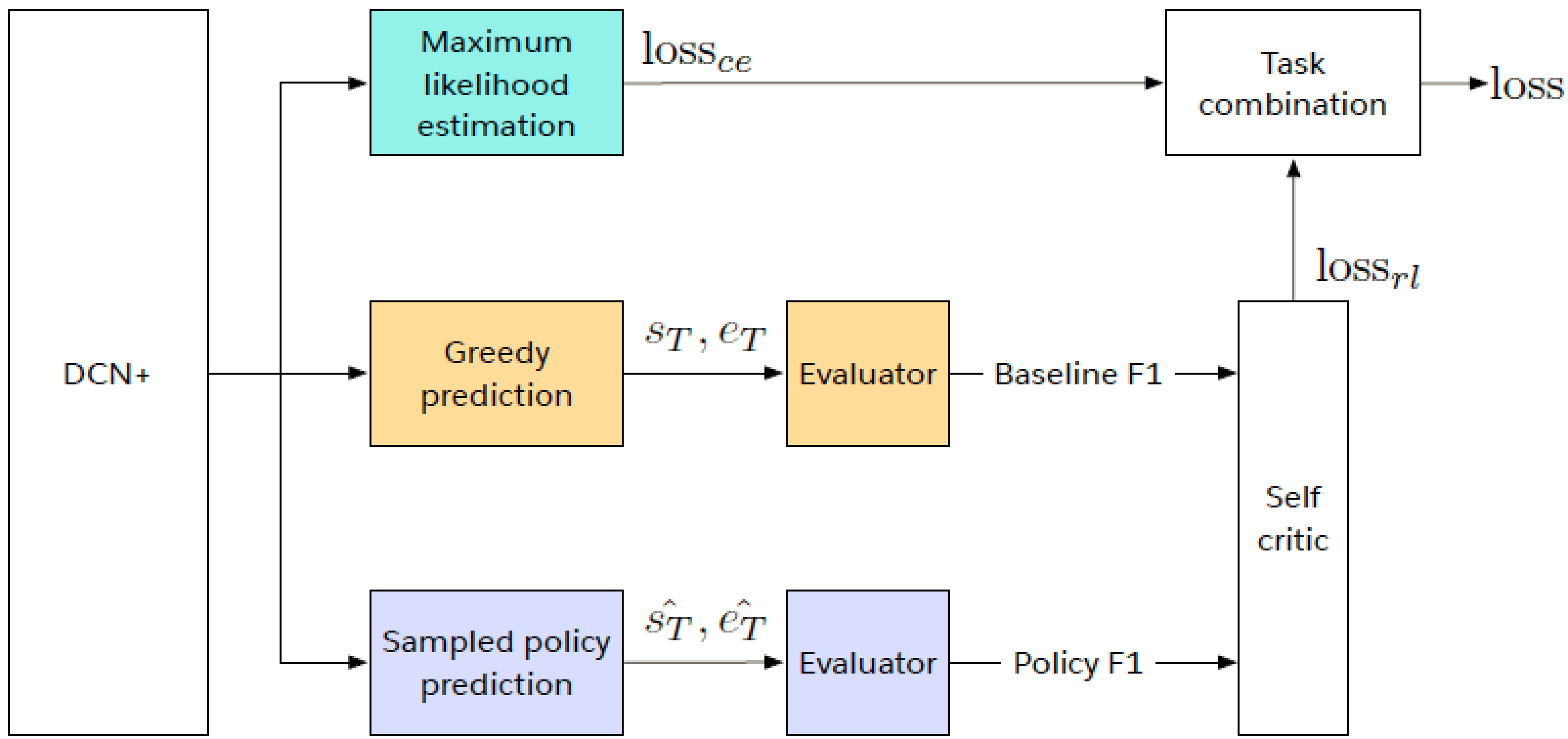
Optimization Objective

$$l_{ce}(\Theta) = -\sum_{t}\left(\log p_t^{\text{start}}\left(s \mid s_{t-1}, e_{t-1}; \Theta\right) + \log p_t^{\text{end}}\left(e \mid s_{t-1}, e_{t-1}; \Theta\right)\right)$$

$$\begin{aligned} l_{rl}(\Theta) &= -\mathbb{E}_{\hat{\tau}\sim p_\tau}\left[R\left(s, e, \hat{s}_T, \hat{e}_T; \Theta\right)\right] \\ &\approx -\mathbb{E}_{\hat{\tau}\sim p_\tau}\left[F_1\left(\mathrm{ans}\left(\hat{\mathrm{s}}_\mathrm{T}, \hat{\mathrm{e}}_\mathrm{T}\right), \mathrm{ans}\left(\mathrm{s}, \mathrm{e}\right)\right) - F_1\left(\mathrm{ans}\left(\mathrm{s}_\mathrm{T}, \mathrm{e}_\mathrm{T}\right), \mathrm{ans}\left(\mathrm{s}, \mathrm{e}\right)\right)\right] \end{aligned}$$

$$\begin{aligned} \nabla_\Theta l_{rl}(\Theta) &= -\nabla_\Theta\left(\mathbb{E}_{\hat{\tau}\sim p_\tau}[R]\right) && (14) \\ &= -\mathbb{E}_{\hat{\tau}\sim p_\tau}\left[R\nabla_\Theta \log p_\tau\left(\tau; \Theta\right)\right] && (15) \\ &= -\mathbb{E}_{\hat{\tau}\sim p_\tau}\left[R\nabla_\Theta\left(\sum_t^T\left(\log p_t^{\text{start}}\left(\hat{s}_t | \hat{s}_{t-1}, \hat{e}_{t-1}; \Theta\right) + \log p_t^{\text{end}}\left(\hat{e}_t | \hat{s}_{t-1}, \hat{e}_{t-1}; \Theta\right)\right)\right)\right] \\ &\approx -R\nabla_\Theta\left(\sum_t^T\left(\log p_t^{\text{start}}\left(\hat{s}_t | \hat{s}_{t-1}, \hat{e}_{t-1}; \Theta\right) + \log p_t^{\text{end}}\left(\hat{e}_t | \hat{s}_{t-1}, \hat{e}_{t-1}; \Theta\right)\right)\right) && (16) \end{aligned}$$

Optimization Objective



Experiment

Model	EM	ΔEM	F1	ΔF1
DCN+ (ours)	74.5%	–	83.1%	–
- Deep residual coattention	73.1%	-1.4%	81.5%	-1.6%
- Mixed objective	73.8%	-0.7%	82.1%	-1.0%
- Mixture of experts	74.0%	-0.5%	82.4%	-0.7%
DCN w/ CoVe (baseline)	71.3%	-3.2%	79.9%	-3.2%

ps.CoVe and residual coattention is important.

Cross-entropy is important for RL.