S-NET: FROM ANSWER EXTRACTION TO ANSWER GENERATION FOR MACHINE READING COMPREHENSION

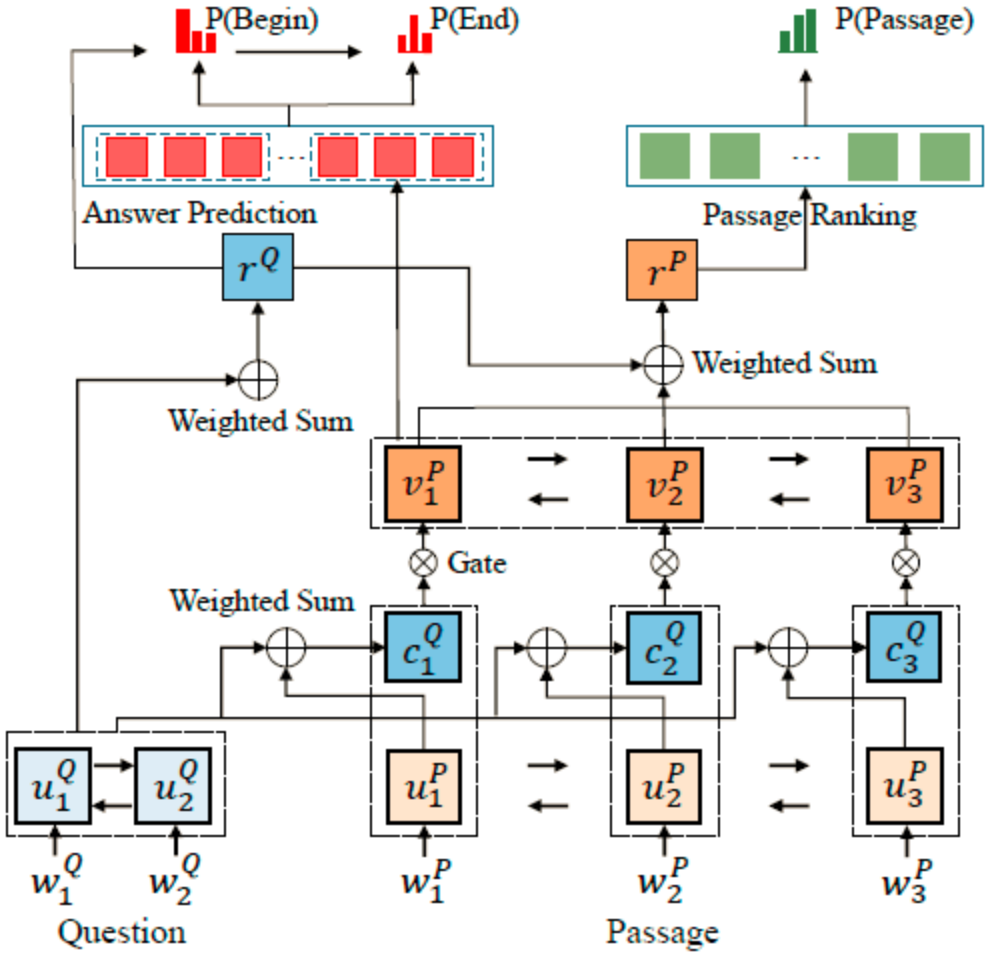
Chuanqi Tany, Furu Weiz, Nan Yang, Weifeng Lvy, Ming Zhou
State Key Laboratory of Software Development Environment,
Beihang University, Beijing, China
Microsoft Research, Beijing, China

MS-MARCO

Different from SQuAD, the answer can exit multi-passages or can not be found in passages. Therefore, the answers should be generated according multi-passages and question.

Module

Evidence Snippet Prediction



Bi-GRU Encoding for Question and Passage

$$u_t^Q = BiGRU_Q(u_{t-1}^Q, [e_tQ, char_t^Q])$$
$$u_t^P = BiGRU_P(u_{t-1}^P, [e_tP, char_t^P])$$

Question-to-Passage Representation

$$s_j^t = v^T tanh(W_u^Q u_j^Q + W_u^P u_t^P)$$
$$a_i^t = exp(s_i^t) / \sum_{j=1}^m exp(s_j^t)$$
$$c_t^Q = \sum_{i=1}^m a_i^t u_i^Q$$

Gated-ReEncoding for Passage

$$g_t = sigmoid(W_g[u_t^P, c_t^Q])$$
$$[u_t^P, c_t^Q]^* = g_t \odot [u_t^P, c_t^Q]$$
$$v_t^P = GRU(v_{t-1}^P, [u_t^P, c_t^Q]^*)$$

Answer Prediction Point-Network

$$s_j^t = v^T tanh(W_h^P v_j^P + W_h^a h_{t-1}^a)$$
$$a_i^t = exp(s_i^t) / \sum_{j=1}^n exp(s_j^t)$$
$$p_t = argmax(a_1^t, ..., a_n^t)$$

Ps.t = 1,2

h_0 = r^Q (question alignment representation)

$$h_1 = GRU(h_0^a, c_t)\ (c_t = \sum_{i=1}^n a_i^t v_i^P)$$

Passage Ranking

Encode for Passage

$$s_j = v^T tanh(W w_v^P v_j^P + W_v^Q r^Q)$$

$$a_i = exp(s_i) / \sum_{j=1}^{n} exp(s_j)$$

$$r^P = \sum_{i=1}^{n} a_i v_i^P$$

Two fully-connected layers for a matching score

$$g = v_g^T (tanh(W_g [r^Q, r^P]))$$

Sotf-Max

$$\hat{g}_i = exp(g_i) / \sum_{j=1}^{k} exp(g_j)$$

Answer Synthesis

Bi-GRU Re-Encoding for Passage and Question

$$h_t^P = BiGRU(h_{t-1}^P, [e_t^p, f_t^s, f_t^e])$$
$$h_t^Q = BiGRU(h_{t-1}^Q, e_t^Q)$$

Ps.e_t indicates word embedding, f_t^s indicates whether the word is the start of evidence.

f_t^e indicates whether the word is the end of evidence.

Decoder

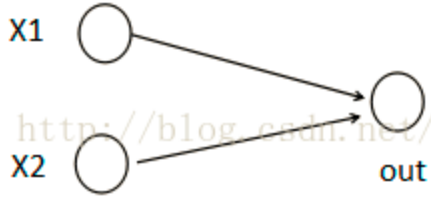
Att-seq-to-seq

$$d_t = \text{GRU}(w_{t-1}, c_{t-1}, d_{t-1})$$
$$d_0 = \tanh(W_d[\overleftarrow{h}_1^P, \overleftarrow{h}_1^Q] + b)$$

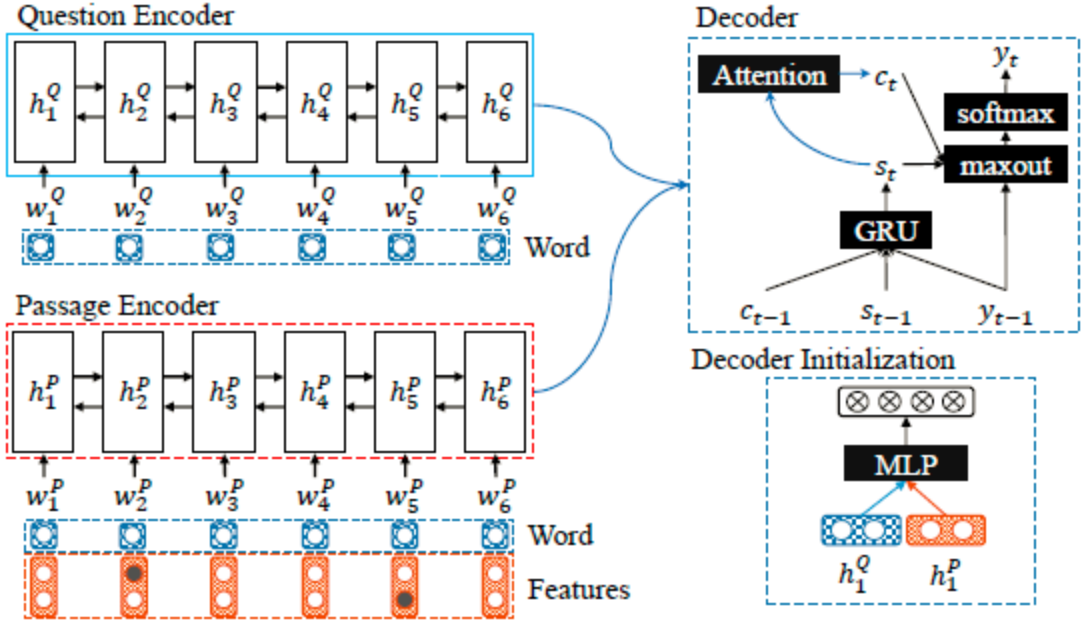
Output

$$r_t = W_r w_{t-1} + U_r c_t + V_r d_t$$
$$m_t = [\max\{r_{t,2j-1}, r_{t,2j}\}]^{\mathrm{T}}$$
$$p(y_t|y_1, \ldots, y_{t-1}) = \mathrm{softmax}(W_o m_t)$$

MaxOut



Answer Synthesis Model



Question?

I don know ranking can produce influence on test.