Tell Me How to Ask Again: Question Data Augmentation with Controllable Rewriting in Continuous Space

Dayiheng Liu, Yeyun Gong, Jie Fu, Yu Yan,
Jiusheng Chen, Jiancheng Lv, Nan Duan, Ming Zhou

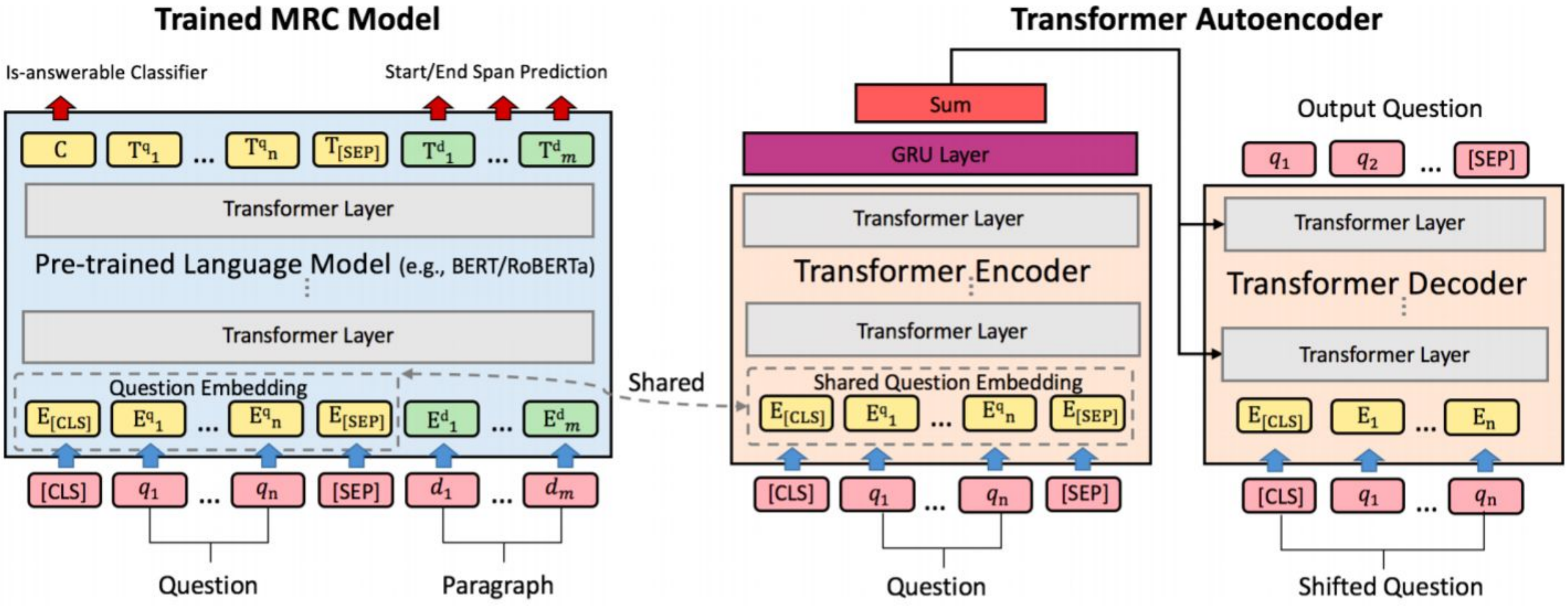
EMNLP 2020

Task

- Question answering:
  - Given a question
  - Given a document
  - Find the span in the document in which answer is provided
  - Binay classification if the question is answerable or not

- They look for a method to augment the questions:
  - To have the same answer span
  - To be answerable/unanswerable from the document
  - To be similar to the original question

Model



Trained MRC Model
Is-answerable Classifier
Start/End Span Prediction
C
T^q_1 ... T^q_n $T_{[SEP]}$ T^d_1 ... T^d_m
Transformer Layer
Pre-trained Language Model (e.g., BERT/RoBERTa)
Transformer Layer
Question Embedding
$E_{[CLS]}$ E^q_1 ... E^q_n $E_{[SEP]}$ E^d_1 ... E^d_m
[CLS] q_1 ... q_n [SEP] d_1 ... d_m
Question
Paragraph
Shared
Transformer Autoencoder
Sum
GRU Layer
Transformer Layer
Transformer Encoder
Transformer Layer
Shared Question Embedding
$E_{[CLS]}$ E^q_1 ... E^q_n $E_{[SEP]}$
[CLS] q_1 ... q_n [SEP]
Question
Output Question
q_1 q_2 ... [SEP]
Transformer Layer
Transformer Decoder
Transformer Layer
$E_{[CLS]}$ E_1 ... E_n
[CLS] q_1 ... q_n
Shifted Question

Pre-trained Language Model based MRC Model

- A BERT-based model to classify the input question is answerable or not and to find the answer span

$$\mathbf{E}^q, \mathbf{E}^d = \mathrm{BertEmbedding}(q, d),$$

$$P_a(\text{is-answerable}) = \mathrm{Sigmoid}(\mathbf{C}\mathbf{W}_c^T + \mathbf{b}_c),$$

$$P_s(i =< \text{start} >) = \mathrm{Sigmoid}(\mathbf{T}_i^d \mathbf{W}_s^T + b_s),$$
$$P_e(i =< \text{end} >) = \mathrm{Sigmoid}(\mathbf{T}_i^d \mathbf{W}_e^T + b_e),$$

$$\begin{aligned} \mathcal{L}_{\text{mrc}} &= \lambda \mathcal{L}_a(t) + \mathcal{L}_s(s) + \mathcal{L}_e(e), \qquad (5) \\ &= -\lambda \log P_a(t) - \log P_s(s) - \log P_e(e), \end{aligned}$$

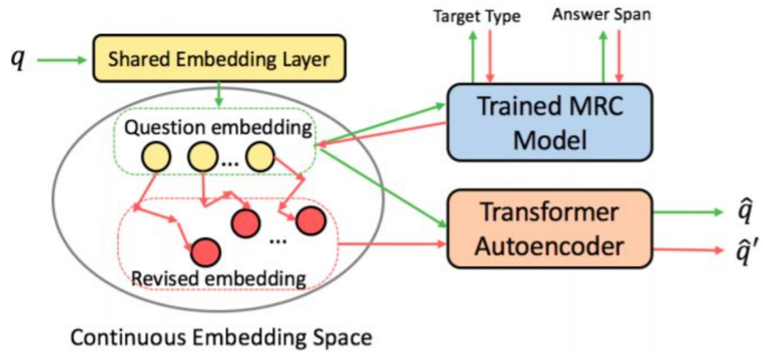
Transformer-based Autoencoder

- Encode the question
  - Use the same embedding as the pre-trained model
- Compute a vector representation of the input question
- Decode the question vector using a Decoder

$$\mathbf{H}_{enc} = \mathrm{TransformerEncoder}(q),$$
$$\mathbf{z} = \mathrm{Sum}(\mathrm{GRU}(\mathbf{H}_{enc})),$$
$$\hat{q} = \mathrm{TransformerDecoder}(\mathbf{z}).$$

Rewriting Question with Gradient-based Optimization

- Three objectives for rewriting:
  - Be unanswerable or have the same span
  - Should not be trapped by local optimum
  - Should be similar to Q



- Unanswerable question: $\mathbf{E}^{q'} = \mathbf{E}^{q} - \eta(\nabla_{\mathbf{E}^q} \mathcal{L}_a(t'))$
- Same Span: $\mathbf{E}^{q'} = \mathbf{E}^{q} - \eta\left(\nabla_{\mathbf{E}^q}(\lambda \mathcal{L}_a(t) + \mathcal{L}_s(s) + \mathcal{L}_e(e))\right)$
- Update step-size for avoiding local optimum
- Use unigram overlap rate for choosing similar questions: $\mathcal{J}(q, \hat{q}') = \frac{\text{count}(w_q \cap w_{\hat{q}})}{\text{count}(w_q \cup w_{\hat{q}})},$

Rewriting Question with Gradient-based Optimization

Algorithm 1 Question Rewriting with Gradient-based Optimization.

Input: Data tuple (q, d, s, e, t); Original question embedding $\mathbf{E}^q$; pre-trained MRC model and Transformer autoencoder; A set of step size $S_\eta = \{\eta_i\}$; Step size decay coefficient β_s; the target answerable or unanswerable label t'; Threshold $\beta_t, \beta_a, \beta_b$;

Output: a set of new answerable and unanswerable question data tuples $\mathcal{D}' = \{(\hat{q}', d, s, e, t'), .., (\hat{q}', d, s, e, t)\}$;

$\mathcal{D}' = \{\}$;
for each $\eta \in S_\eta$ **do**
    for max-steps **do**
        revise $\mathbf{E}^{q'}$ by Eq. (10) or Eq. (9)
        $\hat{q}' = \textbf{TransformerAutoencoder}\left(\mathbf{E}^{q'}\right)$
        if $P_a(t') > \beta_t$ and $\mathcal{J}(q, \hat{q}') \in [\beta_a, \beta_b]$ **then**
            add $(\hat{q}', d, s, e, t')$ to $\mathcal{D}'$;
        end if
        $\eta = \beta_s \eta$;
    end for
end for
return $\mathcal{D}'$;

Results

Methods	EM	F1
BERT$_{\text{large}}$ (Devlin et al., 2018) (original)	78.7	81.9
+ EDA (Wei and Zou, 2019)	78.3	81.6
+ Back-Translation (Yu et al., 2018)	77.9	81.2
+ Text-VAE (Liu et al., 2019a)	75.3	78.6
+ AE with Noise	76.7	79.8
+ 3M synth (Alberti et al., 2019)	80.1	82.8
+ UNANSQ (Zhu et al., 2019)	80.0	83.0
+ CRQDA (ours)	**80.6**	**83.3**

Results

Methods	EM	F1
$\text{BERT}_{\text{base}}$	73.7	76.3
+ CRQDA	**75.8** (+2.1)	**78.7** (+2.4)
$\text{BERT}_{\text{large}}$	78.7	81.9
+ CRQDA	**80.6** (+1.9)	**83.3** (+1.4)
$\text{RoBERTa}_{\text{base}}$	78.6	81.6
+ CRQDA	**80.2** (+1.6)	**83.1** (+1.5)
$\text{RoBERTa}_{\text{large}}$	86.0	88.9
+ CRQDA	**86.4** (+0.4)	**89.5** (+0.6)

Thanks