

Dual Ask-Answer Network for Machine Reading Comprehension

Han Xiao, Feng Wang, Yanjian Feng, Jingyao Zheng

Abstract

There are three modalities in the reading comprehension setting: question, answer and context. The task of question answering or question generation aims to infer an answer or a question when given the counterpart based on context. We present a novel two-way neural sequence transduction model that connects three modalities, allowing it to learn two tasks simultaneously and mutually benefit one another. During training, the model receives question-context-answer triplets as input and captures the cross-modal interaction via a hierarchical attention process. Unlike previous joint learning paradigms that leverage the duality of question generation and question answering at data level, we solve such dual tasks at the architecture level by mirroring the network structure and partially sharing components at different layers. This enables the knowledge to be transferred from one task to another, helping the model to find a general representation for each modality. The evaluation on four public datasets shows that our dual-learning model outperforms the mono-learning counterpart as well as the state-of-the-art joint models on both question answering and question generation tasks.

Introduction

Recently the task of machine reading comprehension (Rajpurkar et al. 2016; Nguyen et al. 2016) has received increasing attention from NLP and AI research communities. The task attempts to enable machines to answer questions after reading a passage. There are three modalities in this task: question, answer and context. Many successful question answering models have been proposed to fill answer modality by modeling the interactions between questions and contexts modalities (Seo et al. 2016; Yu et al. 2018). Intuitively, question and answer are quite similar in the sense that they are both short texts and strongly related to the given context. Few very recent works have recognized such relationship and explored from different perspectives (Song, Wang, and Hamza 2017; Wang, Yuan, and Trischler 2017; Tang et al. 2018). Their promising results suggest that: (i) the roles of questions and answers are switchable given context, in the sense that the same network structure can be used for both question answering (QA) and question generation (QG); (ii) this invertibility or duality can be exploited to improve QA and QG.

In this work, we consider machine reading comprehension as a dual learning problem, i.e. learning to answer versus learning to ask. It is no surprise that to utilize the commonality among

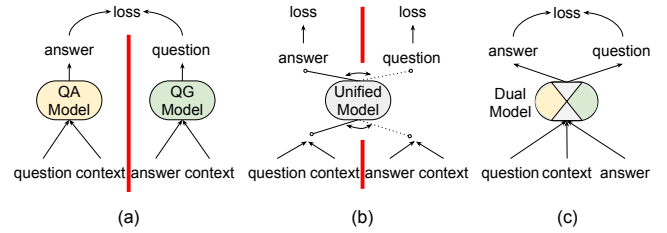


Figure 1: Three learning paradigms to exploit the task correlations of question answering and question generation. Red line represents data/model-level separation. (a) Two separated models with a joint loss function (Song, Wang, and Hamza 2017). (b) A unified model with alternated training input and two separated loss functions (Wang, Yuan, and Trischler 2017). (c) **This work**: a unified architecture with locally shared structure that can learn two tasks simultaneously. In contrast to (a) and (b), there is no data-level or model-level separation in our learning paradigm.

the tasks, some sharing scheme is required in the learning paradigm. Figure 1 visualizes previous dual learning paradigms and the one from this work. Specifically, we propose a novel two-way neural sequence transduction model that jointly solves these two tasks. The model is structurally symmetric and its components are shared between two tasks at different levels, as illustrated in Figure 2. During training, the model receives question-context-answer triplets as input and captures the cross-modal interaction via a hierarchical attention process. During testing, the model generates answer or question as sequence when given the counterpart based on context. Our contributions are summarized as follows:

- We propose a novel two-way neural sequence transduction model that learns question answering and question generation simultaneously. We tie the network components that playing similar roles in the two tasks to transfer cross-task knowledge during training. The cross-modal interaction of question, context and answer is captured by a pair of symmetric hierarchical attention processes. To the best of our knowledge, we are the first to leverage duality at this granularity under the machine reading comprehension setting.
- We demonstrate the effectiveness of our model on four public datasets and achieve promising results. Our model outperforms the mono-learning counterpart and the state-of-the-art

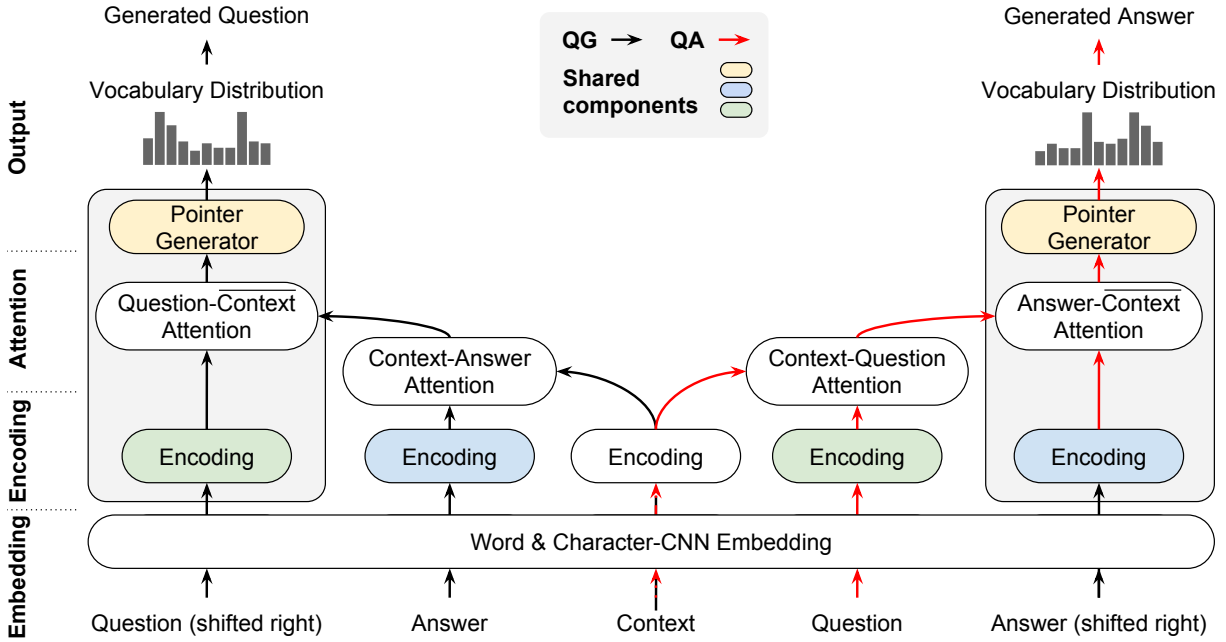


Figure 2: The model architecture of Dual Ask-Answer Network (DAANET). *Best viewed in color.* It is a hierarchical process consisted of four layers: embedding, encoding, attention and output. The rectangle super-block on the side can be viewed as a decoder for QG and QA, respectively. The computational flow of QG and QA task is plotted with black arrows and red arrows, respectively. Shared (including partially shared) components are filled with the same color, i.e. blue for the answer encoder, green for the question encoder and yellow for the pointer generator. Note that, the context encoder is also shared by QA and QG. For the sake of clarity we only draw one context encoder block. During testing, the shifted input is replaced by the model’s own generated words from the previous steps.

joint question answering models. We provide a strong evidence for bringing the duality into model structures to improve the reading comprehension ability. The code and data used in this paper are available online to enable future comparison¹.

Related Work

Our work relates to the existing works on machine reading comprehension (MRC), question answering (QA) and question generation (QG). In this section, we briefly review the previous works from these perspectives.

A great number of MRC models have been developed over the past few years, e.g. BiDAF (Seo et al. 2016), S-Net (Tan et al. 2017), R-Net (Wang et al. 2017), match-LSTM (Wang and Jiang 2016), ReasonNet (Shen et al. 2017), Document Reader (Chen et al. 2017), Reinforced Mnemonic Reader (Hu et al. 2018), FusionNet (Huang et al. 2017) and QANet (Yu et al. 2018). Most of the existing models rely on the assumption that the answer is a continuous span of a given passage. Under this assumption, an answer can be simplified as a pair of two integers, representing its start and end position in the passage respectively. In this work, we do not make such assumption. Our model generates both answer and question as sequence, so that the same model architecture can be used for both question answering and question generation. In general, we believe such generative model offers better expressivity for complicated semantics, and thus is more likely to deliver a true breakthrough in machine reading comprehension.

¹<https://github.com/hanxiao/daanet>

A large part of prior works on question generation relies on feature engineering, handcrafted templates and linguistic rules (Aldabe et al. 2006; Heilman 2011; Liu, Calvo, and Rus 2010; Bordes, Weston, and Usunier 2014; Dhingra, Pruthi, and Rajagopal 2018). Recently, encouraged by the remarkable success of the sequence-to-sequence model in machine translation (Sutskever, Vinyals, and Le 2014; Luong, Pham, and Manning 2015), parsing (Vinyals et al. 2015) and text summarization (Nallapati et al. 2016), the interest of using deep neural network to tackle QG rises rapidly (Du, Shao, and Cardie 2017; Yuan et al. 2017; Duan et al. 2017). One of the major motivations in these studies is using QG to enrich the training data of QA. For example, a corpus with 30M QA pairs is generated by transducing facts from a knowledge base into natural language questions using a deep neural network (Serban et al. 2016).

Finally, some very recent MRC works have recognized the relationship between QA and QG and exploited it differently. In spite of a similar objective shared by this study, our work is unique from the following perspectives:

- Unlike (Yang et al. 2017) that rely on the continuous span assumption of answer and (Sachan and Xing 2018; Tang et al. 2017; 2018) that consider answering as a sentence selection task, we consider both question answering and question generation as sequence transduction so that it is possible to use the same model architecture for both tasks. Unlike the extraction-then-synthesis framework proposed in (Tan et al. 2017), our model is trained completely end-to-end.
- Unlike (Song, Wang, and Hamza 2017) that train a QA

model and a QG model independently using the same architecture and (Wang, Yuan, and Trischler 2017) that alternate the training data between QA and QG examples for the same model, our two-way model directly consumes question-context-answer triplets during training. Given a triplet, the parameters will be updated “twice” in the sense that they receive gradients from both QA and QG directions. Consequently, the parameters in our model are trained more sufficiently comparing to the other two related works.

- Unlike (Tang et al. 2018) that use a collaboration detector to connect the training of two tasks, we propose a novel sharing scheme to leverage duality at the architecture level. The components and parameters that playing similar roles in the two tasks are tied together. The interaction between question, answer and context is captured by coupling two symmetric hierarchical attention processes. This sharing scheme greatly reduces the total number of parameters required for the two tasks.

Dual Ask-Answer Network

We first formulate the dual learning problem of asking and answering in the MRC setting, and then present our model: Dual Ask-Answer Network (DAANET). Finally, we make a summary about the attention and the duality used in the model.

Problem Formulation

Unlike the traditional MRC problem focusing only on QA, the problem considered in this work is bipartite: QA and QG. Specifically, the model should be able to infer answers or questions when given the counterpart based on context.

Formally, we denote a context paragraph as $C := \{c_1, \dots, c_n\}$, a question sentence as $Q := \{q_1, \dots, q_m\}$ and an answer sentence as $A := \{a_1, \dots, a_k\}$. In the sequel, we follow this notation and use n, m, k to represent the length of a context, a question and an answer, respectively. The context C is shared by the two tasks. Given C , the QA task is defined as finding the answer A based on the question Q ; the QG task is defined as finding the question Q based on the answer A . In contrast to previous MRC models that mostly assume the answer as a continuous span, we regard both QA and QG tasks as generation problems and solve them jointly in a neural sequence transduction model.

Model Description

The high level architecture of our proposed Dual Ask-Answer Network is illustrated in Figure 2. This neural sequence transduction model receives string sequences as input and processes them through an embedding layer, an encoding layer, an attention layer, and finally to an output layer to generate sequences.

1. Embedding Layer. The embedding layer maps each word to a high-dimensional vector space. The vector representation includes the word-level and the character-level information. The parameters of this layer are shared by context, question and answer. For the word embedding, we use pre-trained 256-dimensional GloVe (Pennington, Socher, and Manning 2014) word vectors, which are fixed during training. All the out-of-word vocabulary words are mapped to an $\langle \text{UNK} \rangle$ token. Besides that, there are three special tokens: $\langle \text{PAD} \rangle$,

$\langle \text{START} \rangle$ and $\langle \text{END} \rangle$. The embeddings of $\langle \text{START} \rangle$, $\langle \text{END} \rangle$ and $\langle \text{UNK} \rangle$ are trainable and initialized randomly, whereas $\langle \text{PAD} \rangle$ is fixed to an all-zero vector.

For the character embedding, each character is represented as a 200-dimensional trainable vector. Consequently, each word can be represented as a sequence of character vectors, where the sequence length is either truncated or padded to 16. Next, we conduct 1D CNN with kernel width 3 followed by max-pooling along the time axis. This gives us a fixed-size 200-dimensional vector for each word. Finally, the concatenation of the character embedding and word embedding vectors is linearly projected to a 300-dimensional space and then passed to a highway network (Srivastava, Greff, and Schmidhuber 2015) as follows:

$$\mathbf{e} := [\mathbf{e}_{\text{word}}, \mathbf{e}_{\text{char}}] \mathbf{H}_1 + v_1, \quad \mathbf{g} := \sigma(\mathbf{e} \mathbf{H}_2 + v_2), \\ \mathbf{e}_{\text{out}} := \mathbf{g} \odot \mathbf{e} + (1 - \mathbf{g}) \odot (\mathbf{e} \mathbf{H}_3 + v_3),$$

where $\{\mathbf{H}_1 \in \mathbb{R}^{456 \times 300}, \mathbf{H}_2, \mathbf{H}_3 \in \mathbb{R}^{300 \times 300} \text{ and } v_1, v_2, v_3 \in \mathbb{R}\}$ are learned parameters; σ denotes the sigmoid function. The dimension of the output is 300.

2. Encoding Layer. The encoding layer contains three encoders for context, question and answer, respectively. They are shared by QA and QG, as depicted in Figure 2. That is, given QA and QG dual tasks, the encoder of the primal task and the decoder of the dual task are forced to be the same. The parameter sharing scheme serves as a regularization to influence the training on both tasks. It also helps the model to find a more general and stable representation for each modality.

Each encoder consists of the following basic building blocks: a element-wise fully connected feed-forward block, stacked LSTMs (Gers, Schmidhuber, and Cummins 2000) and a self-attention block (Vaswani et al. 2017). The self-attention block allows each position in the encoder to attend to all positions based on the similarity measured by their dot products. The attention function used in self-attention block is described in the attention layer. Each block is followed by layer normalization (Ba, Kiros, and Hinton 2016). The final output of an encoder is a concatenation of the outputs of all blocks, as illustrated in Figure 3.

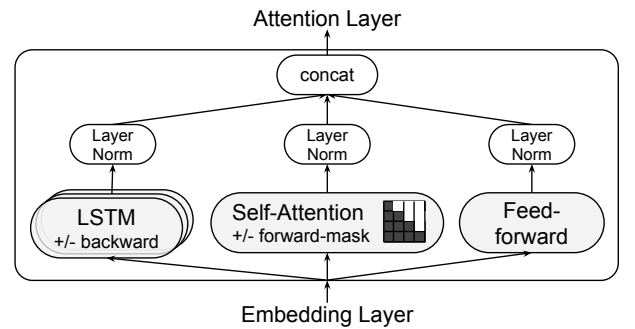


Figure 3: The architecture of a context/question/answer encoder. In the question and answer encoder, LSTM is unidirectional and a forward mask is used when computing self-attention.

The context encoder consists of a feed-forward network, a 3-layer bidirectional LSTM and a 4-head self-attention. The question and answer encoders follow the same composition but are with necessary modifications to prevent the “leakage”

of leftward information. Specifically, in question and answer encoder LSTM is *unidirectional* and a forward mask is used when computing self-attention. The forward mask keeps only attention of later position to early position, meanwhile setting the remaining to $-\infty$ as follows,

$$\text{SelfAtt}(\mathbf{e}_{\text{out}}) := \text{softmax}\left(\frac{1}{\sqrt{d}}\mathbf{e}_{\text{out}}\mathbf{R}(\mathbf{e}_{\text{out}}\mathbf{R})^\top + \mathbf{M}\right)\mathbf{e}_{\text{out}},$$

$$\mathbf{M}_{ij} := 0 \quad \text{if } i < j \quad \text{else } -\infty,$$

where \mathbf{e}_{out} is the output of the embedding layer and $\mathbf{R} \in \mathbb{R}^{300 \times d}$ is a learned parameter. Different compositions of the encoder are benchmarked in the experiment section.

3. Attention Layer. The attention layer fuses all information observed so far from context encode \tilde{C} , answer encode \tilde{A} and question encode \tilde{Q} in a hierarchical manner. At time t in QG, we first fold the answer encode into the context encode, which is folded again into the question encode $\tilde{Q}_{<t}$ generated previously. The QA part follows the similar attention procedure. In this work, we denote the first fold-in step as the “context- \ast ” attention and the second fold-in step as “ \ast -context” attention, where $\overline{\text{context}}$ is the output of the first fold-in step. Figure 4 visualizes this two-step attention procedure.

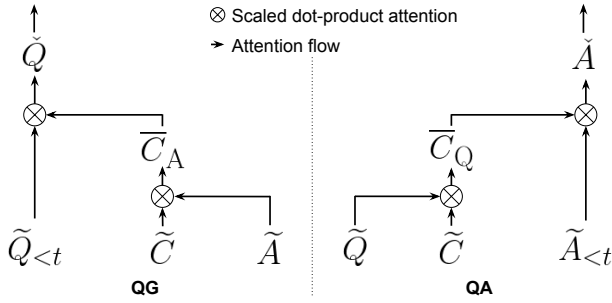


Figure 4: The two-step attention flow implemented in the attention layer, where the input $\tilde{Q}, \tilde{C}, \tilde{A}, \tilde{Q}_{<t}, \tilde{A}_{<t}$ are from the previous encoding layer.

More concretely, the context-answer attention is computed as follows: we first compute the normalized similarities between each pair of context and answer word as $s(\tilde{C}, \tilde{A}) := \text{softmax}(\frac{1}{\sqrt{d}}\tilde{C}\mathbf{U}(\tilde{A}\mathbf{V})^\top)$, where the score function $s: \mathbb{R}^{n \times d} \times \mathbb{R}^{k \times d} \mapsto \mathbb{R}^{n \times k}$. Learnable variables in s are $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d \times d'}$. This is the multiplicative attention with the scaling factor of $\frac{1}{\sqrt{d}}$, as proposed in (Vaswani et al. 2017). The context-answer attention is then computed as a weighted sum of the answer, i.e. $\overline{C}_A := s(\tilde{C}, \tilde{A}) \cdot \tilde{A}$. The context-question attention follows the same procedure, i.e. $\overline{C}_Q := s(\tilde{C}, \tilde{Q}) \cdot \tilde{Q}$. The output of the first fold-in step is in $\mathbb{R}^{n \times d}$, the same size as the context sequence.

The second fold-in step involves computing question-context and answer-context attention, which is essential for generating meaningful question and answer sequence, respectively. It allows every position in the generated sequence to attend over all positions in context sequence from the first fold-in step. At time t , the input of question-context attention consists of $\tilde{Q}_{<t}$ from

the previous encoding layer and \overline{C}_A from the context-answer attention, the output is $\tilde{Q} := s(\tilde{Q}_{<t}, \overline{C}_A) \cdot \overline{C}_A$. The answer-context attention follows the same procedure, i.e. $\tilde{A} := s(\tilde{A}_{<t}, \overline{C}_Q) \cdot \overline{C}_Q$.

4. Output Layer. The output layer generates an output sequence one word at a time. At each step the model is auto-regressive (Graves 2013), consuming the words previously generated as input when generating the next. In this work, we employ the pointer generator as the core component of the output layer (See, Liu, and Manning 2017). It allows both copying words from the context via pointing, and sampling words from a fixed vocabulary. This aids accurate reproduction of information especially in QA, while retaining the ability to generate novel words.

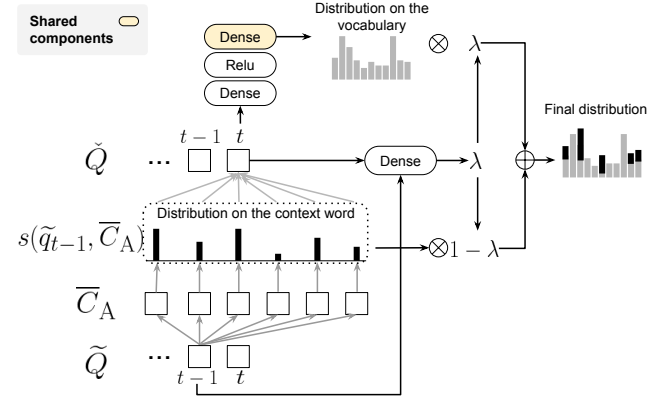


Figure 5: The architecture of the output layer for QG. Each square block represents a word. The final distribution is a mixture of two discrete distributions. QA follows the same architecture with few replacements. The dense layer in yellow is shared by QA and QG.

The architecture of the output layer is depicted in Figure 5. Specifically, at time t the probability of generated word w is defined as a mixture of two discrete distributions $P_t(w) := \lambda P_t^{\text{vocab}}(w) + (1 - \lambda) P_t^{\text{context}}(w)$, where $P_t^{\text{vocab}}(w)$ is a distribution over the predefined vocabulary, $P_t^{\text{context}}(w)$ is a distribution over words in the current context C based on the “*-context” attention score (duplicated words are merged), and λ_t is a soft switch to choose between generating a word and copying from the context. Formally, the final vocabulary distribution for QG follows the form below:

$$P_t^{\text{vocab}} := \text{softmax}(\tanh(\tilde{q}_t \mathbf{W}_1 + b_1) \mathbf{W}_{\text{shared}} + b_{\text{shared}})$$

$$P_t^{\text{context}} := s(\tilde{q}_{t-1}, \overline{C}_A)$$

$$\lambda_t := \sigma([\tilde{q}_t, \tilde{q}_{t-1}] \mathbf{W}_2 + b_2),$$

where $s(\cdot, \cdot)$ is the score function defined in the attention layer; \tilde{q}_t and \tilde{q}_{t-1} corresponds to the t^{th} row of \tilde{Q} and the $(t-1)^{\text{th}}$ row of \tilde{Q} , respectively. Learnable variables are $\mathbf{W}_1 \in \mathbb{R}^{d \times 1024}$, $\mathbf{W}_{\text{shared}} \in \mathbb{R}^{1024 \times |V|}$, $\mathbf{W}_2 \in \mathbb{R}^{2d \times 1}$, $b_1 \in \mathbb{R}^{1024}$, $b_{\text{shared}} \in \mathbb{R}^{|V|}$ and a scalar b_2 , where $|V|$ represents the number of words in the vocabulary. Note that $\mathbf{W}_{\text{shared}}$ and b_{shared} are shared by QA and QG, whereas $\{\mathbf{W}_1, \mathbf{W}_2, b_1, b_2\}$ are task-specific parameters.

There are two advantages of this sharing scheme. First, sharing $\mathbf{W}_{\text{shared}}$ significantly reduces the number of parameters,

which helps the model to better capture the cross-modal interaction in the attention layer. Second, it forces the model to first project the decoded information from their own spaces to a common latent space before projecting to the vocabulary space. This common latent space serves as a tie that connects two tasks, providing another channel for transferring knowledge. One may also consider it as a regularization we set at the upper-level of the network to influence the training process. In the experiment section, we show that this sharing scheme indeed contributes and improves the performance.

For QA, the parametric form of $P_t(w)$ is defined similarly by first replacing \tilde{q}_{t-1} with \tilde{a}_{t-1} , \bar{C}_A with \bar{C}_Q , and \tilde{q}_t with \tilde{a}_t ; then equipping the task-specific parameters $\{\mathbf{W}_1, \mathbf{W}_2, b_1, b_2\}$; finally sharing $\mathbf{W}_{\text{shared}}$ and b_{shared} with QG.

Loss Function

During training, the model is fed with a question-context-answer triplet (Q, C, A) , and the decoded \hat{Q} and \hat{A} from the output layer are trained to be similar to Q and A , respectively. To achieve that, our loss function consists of two parts: the negative log-likelihood loss widely used in the sequence transduction model and a coverage loss to penalize repetition of the generated text, which is akin to the form in (See, Liu, and Manning 2017).

We employ *teacher forcing* strategy in training, where the model receives the groundtruth token from time $(t-1)$ as input and predicts the token at time t . Specifically, given a triplet (Q, C, A) the objective is to minimize the following loss function with respect to all model parameters:

$$\begin{aligned} \ell := & \underbrace{-\sum_{t=1}^m \left(\log p(\hat{q}_t = q_t | Q_{<t}, A, C) - \kappa \sum_{\min} (s_t^{\text{QC}}, \sum_{t'=1}^{t-1} s_{t'}^{\text{QC}}) \right)}_{\text{QG loss}} \\ & \underbrace{-\sum_{t=1}^k \left(\log p(\hat{a}_t = a_t | A_{<t}, Q, C) - \kappa \sum_{\min} (s_t^{\text{AC}}, \sum_{t'=1}^{t-1} s_{t'}^{\text{AC}}) \right)}_{\text{QA loss}} \end{aligned}$$

where s_t^{QC} and s_t^{AC} corresponds to the t^{th} row of question-context and the answer-context attention score obtained from the second fold-in step in the attention layer, respectively; \sum_{\min} represents the sum of element-wise minimum of two vectors; κ is a hyperparameter for weighting the coverage loss.

Attention in the Model

Our model exploits the attention mechanism in three places.

- In the encoding layer, we employ self-attention as one way to capture long-range dependencies. Comparing to LSTM, the maximum path length a signal has to traverse in the network is much shorter in self-attention, making it easier to learn long-range dependencies. When encoding context, we allow both forward and backward signals to traverse in the self-attention network. When encoding question and answer with self-attention, we prevent the backward signal with a mask to preserve the auto-regressive property.

- In the attention layer, we develop a two-step attention that folds in all information observed so far for generating final sequences. The first fold-in step captures the interaction between question/answer and context and represents it as a new context sequence. The second fold-in step mimics the typical encoder-decoder attention mechanisms in the sequence-to-sequence models (Bahdanau, Cho, and Bengio 2014; Gehring et al. 2017).
- Finally in the output layer, we recycle the attention score obtained from the second fold-in step and use it as the copy distribution in the pointer generator. The ultimate vocabulary is thus defined as an interpolation of a distribution over the current context words and a distribution over the predefined large vocabulary. This facilitates copying words from the context, while retaining the ability to generate new words.

Duality in the Model

Our model exploits the duality of QA and QG in two places.

- As we consider both QA and QG are sequence generation problems, our architecture is reflectional symmetric, as depicted in Figure 2. The left QG part is a mirror of the right QA part with identical structures. Such symmetry can be also found in the attention calculation and the loss function. Consequently, the answer modality and the question modality are connected in a two-way process through the context modality, allowing the model to infer answers or questions given the counterpart based on context.
- Our model contains shared components between QA and QG at different levels. Starting from the bottom, the embedding layer and the context encoder are always shared between two tasks. Moreover, the answer encoder in QG is reused in QA for generating the answer sequence, and vice versa. On top of that, in the the pointer generator, QA and QG share the same latent space before the final projection to the vocabulary space. The cycle consistency between question and answer is utilized to regularize the training process at different levels, helping the model to find a more general representation for each modality.

Experimental Results

Implementation Details

As every component in the proposed model is differentiable, all parameters could be trained via back propagation. We randomly initialized the parameters using the ‘‘fan-avg’’ strategy, i.e. sampling from a uniform distribution whose width is the average number of inputs and output connections (Glorot and Bengio 2010). We filtered out words that appear less than 5 times in the data and replaced them with $\langle \text{UNK} \rangle$ token, resulting a vocabulary with 90,000 words. We fixed the word embedding to a pretrained 256-dimensional GloVe word vectors (Pennington, Socher, and Manning 2014). The character embedding was in the size of 66×256 and was learned during training. Dropout was mainly applied to the encoding layer with the keep rate of 0.9. The coverage loss weight κ was 1.0. We used the Adam optimizer (Kingma and Ba 2014) for minimizing the loss function. The gradient was clipped by restricting its ℓ_2 -norm less than or equal to 5.0. The learning rate was increased from zero to 0.001 with an inverse exponential function and

then fixed for the remainder of training. The batch size was empirically set to 16. During testing, we conduct auto-regressive decoding separately for QA and QG. The input at time t is its own generated words from the previous steps $< t$. Decoding is terminated when the model encounters the first $< \text{END} >$ or when the sequence contains more than 100 words.

Datasets

Our experiments were conducted on four datasets: SQuAD (Rajpurkar et al. 2016), MSMARCO (Nguyen et al. 2016), WikiQA (Yang, Yih, and Meek 2015) and TriviaQA (Joshi et al. 2017). We subsampling MSMARCO to speedup the experiments. All data are preprocessed to SQuAD format, which is available online. Table 1 summarizes the statistics of the training and testing data used in the experiments.

Dataset	#Train	#Test	\bar{n}	\bar{m}	\bar{k}
SQuAD	86,821	5,928	117.1	10.1	3.1
MSMARCO	120,000	24,000	405.8	4.1	12.3
TriviaQA	120,000	24,000	631.9	12.2	1.5
WikiQA	873	126	471.2	6.4	24.1

Table 1: Summary of the datasets. Statistics include the number of training/testing (Q, A, C) triplets; the average number of words in context (\bar{n}), question (\bar{m}) and answer (\bar{k}).

Evaluation Metrics

As both question and answer are generated in our model, we adopt BLEU-1,2,3,4 (Papineni et al. 2002) and Meteor (Denkowski and Lavie 2014) scores from machine translation, and ROUGE-L from text summarization (Lin 2004) to evaluate the quality of the generation. These metrics are calculated by aligning machine generated text with one or more human generated references based on exact, stem, synonym, and paraphrase matches between words. Higher score is preferable as it suggests better alignments with the groundtruth.

Effectiveness of Dual Learning

Table 2 summarizes the performance of our model versus the mono-learning counterpart (mono), JointQA (Wang, Yuan, and Trischler 2017) (jqqa), multi-perspective QG without reinforcement learning (Song, Wang, and Hamza 2017) (mpqg) and a simple sequence-to-sequence model with attention mechanism (s2s). Both jqqa and mpqg share the same objective as our work, i.e. leveraging the duality of QA and QG to improve each other. We implement mono by masking out all the task-irrelevant structures and parameters from DAANET. One can observe that, DAANET considerably outperforms the mono-learning counterpart with an average margin of 5.7pp in Rouge-L and 5.3pp in Bleu-4 on QA. It also outperforms the state-of-the-art joint model in most of the cases. This result provides a strong evidence for bringing the duality into model structures to improve the reading comprehension ability. One may also notice that, the improvement of DAANET is more significant in QA than in QG. This is due to the fact that there are many ways to convey the same question in QG, e.g. by replacing

interrogative, thus metrics based on literally matching may underestimate the quality of generated question. We leave the investigation of a better QG performance metric as a future work.

		QA					QG				
		DAANET	mono	jqqa	mpqg	s2s	DAANET	mono	jqqa	mpqg	s2s
SQuAD	B1	37.28	29.62	26.66	14.65	15.47	33.95	31.28	34.48	29.92	28.73
	B2	32.66	25.37	22.14	11.17	11.44	18.99	16.23	18.84	14.82	13.05
	B3	29.31	22.23	18.82	8.74	8.74	12.35	9.92	12.09	8.64	6.73
	B4	26.38	19.56	16.11	6.93	6.75	8.71	6.64	8.32	5.49	3.74
	RL	43.85	37.38	41.43	26.96	24.21	35.58	33.32	34.31	32.24	30.92
	Mt	23.21	19.73	21.70	12.64	8.95	14.18	12.52	13.89	11.70	11.41
MSMARCO	B1	44.98	41.20	41.26	35.51	29.41	58.45	59.04	56.28	49.42	53.20
	B2	38.09	34.70	34.53	28.51	22.84	45.65	46.14	43.64	37.25	40.01
	B3	34.83	31.62	31.37	25.43	19.66	35.98	37.47	34.14	28.40	30.31
	B4	32.82	29.70	29.42	23.60	17.67	28.79	29.18	26.99	22.10	22.86
	RL	44.25	42.23	41.30	35.23	32.75	57.90	58.19	56.13	50.90	53.57
	Mt	22.60	22.28	22.06	16.93	15.33	29.52	29.78	27.94	24.14	25.73
TriviaQA	B1	62.13	48.47	47.67	41.41	50.14	45.61	43.26	37.57	34.07	31.93
	B2	59.40	45.13	44.54	37.91	45.07	33.25	29.91	23.49	19.91	17.43
	B3	53.41	39.68	40.90	31.45	32.88	26.60	22.81	16.91	13.23	10.45
	B4	40.93	29.10	34.54	20.10	19.18	22.32	18.33	11.48	9.46	6.72
	RL	61.97	47.89	46.86	41.94	51.87	45.31	42.69	38.26	34.57	32.81
	Mt	38.80	29.64	28.45	24.79	31.38	21.35	19.67	16.64	14.54	13.43
WikiQA	B1	37.31	36.47	31.74	33.53	7.49	14.13	13.49	14.45	15.53	11.06
	B2	31.76	30.89	26.31	28.81	2.82	7.67	7.28	7.76	7.78	6.37
	B3	29.74	28.78	24.33	27.02	1.21	2.89	0.00	0.00	0.00	0.00
	B4	28.61	27.56	23.25	25.98	0.56	1.71	0.00	0.00	0.00	0.00
	RL	37.32	35.50	29.91	36.41	10.74	18.86	19.61	19.24	20.03	18.45
	Mt	19.50	18.90	14.95	20.08	3.47	5.71	5.71	5.63	5.47	4.27

Table 2: Results of question answering and question generation on four datasets. B* represents Bleu-1,2,3,4; RL is Rouge-L; Mt is Meteor. Higher is better; best in bold.

Ablation Study

Table 3 summarizes the performance of DAANET and its ablations on our SQuAD test set, in which we compare different compositions of encoder, with and without attention mechanisms and different sharing schemes. First, one can observe that LSTM is the most important piece in encoding, without which the performance decreases around 10pp on QA and 2.5pp on QG. Self-attention also contributes to the improvement with 3pp on QA and 0.6pp on QG. Next, removing context-* attention leads to a catastrophic performance. This is in line with our intuition as no context-* attention means that the answer generation is independent to the given question and depends only on the given context. Finally, the last part of Table 3 demonstrates the effectiveness of the proposed sharing scheme.

Case Study

Table 4 presents some generated questions and answers from DAANET and the mono-learning models. Question or answer is generated given the gold counterpart based on context. In the first two samples, DAANET works perfectly, whereas the

Ablation	QA				QG			
	Bleu-1	Bleu-4	Rouge-L	Metero	Bleu-1	Bleu-4	Rouge-L	Metero
Encoder without LSTM	20.80	11.22	30.84	14.84	31.60	6.30	32.75	12.38
Encoder without self-attention	33.24	22.69	39.90	21.31	33.56	8.14	34.10	13.65
No context-* attention	4.96	0.38	4.61	1.81	25.74	2.05	26.99	8.47
No copy mechanism	8.24	0.94	11.23	3.99	26.32	2.22	27.29	8.83
Unshared question & answer encoders	29.82	19.43	37.75	19.47	33.23	8.51	35.26	13.86
Unshared context encoder	22.90	13.45	29.33	14.67	33.37	8.17	34.94	13.75
Unshared output vocabulary projection	32.64	21.46	40.37	20.77	33.71	8.15	34.77	13.85
DAANET	37.28	26.38	43.85	23.21	33.95	8.71	35.58	14.18

Table 3: The performance of DAANET and its ablations on SQuAD. Higher is better; best in bold

mono-learning model fail to provide the desired output. In the third sample, the generated question from DAANET is more readable comparing to the groundtruth. Empirically, we find that DAANET poses questions that are semantically similar to the referenced questions but phrased differently. Although human may consider these as good cases, DAANET still scores poorly on QG under all alignment-based evaluation metrics. This explains the scores on QG are overall lower than QA and the improvements on QG are less significant than on QA. We attach more generated samples in the supplementary document.

Conclusion

We proposed Dual Ask-Answer Network, a two-way neural sequence transduction model, which solves question answering and question generation for machine reading comprehension. We exploited the duality of QA and QG tasks by sharing local structures at different levels. The attention mechanism is used in multiple layers to capture the interaction of context, question and answer. We demonstrated the effectiveness of our model on four public datasets, providing a strong evidence for bringing the duality into model structures to improve the reading comprehension ability. One interesting future direction is to study the collaboration of multiple DAANET. For example, one can stack DAANET multiple times, such that the generated question and answer from the lower network are fed to the upper network. This “bootstrapping” strategy can be considered as an implicit way of data augmentation, alleviating the shortage on labeled MRC data.

References

- Aldabe, I.; De Lacalle, M. L.; Maritxalar, M.; Martinez, E.; and Uria, L. 2006. Arikitturri: an automatic question generator based on corpora and nlp techniques. In *International Conference on Intelligent Tutoring Systems*, 584–594. Springer.
- Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv:1607.06450*.
- Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*.
- Bordes, A.; Weston, J.; and Usunier, N. 2014. Open question answering with weakly supervised embedding models. In *ECML-PKDD*, 165–180. Springer.
- Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 2017.

Context	In the course of the 10th century, the initially destructive incursions of Norse war bands into the rivers of France evolved into more permanent encampments that included local women and personal property. The Duchy of Normandy, which began in 911 as a fiefdom, was established by the treaty of Saint-Clair-sur-Epte between King...
Question	▷ When was the Duchy of Normandy founded?
Answer	◊ 911
DAANET	▷ when did the duchy of normandy open? ◊ 911
mono	▷ when did duchy of normandy begin? ◊ 10th century
Context	... Melbourne is home to a number of museums, art galleries and theatres and is also described as the “sporting capital of Australia”. The Melbourne Cricket Ground is the largest stadium in Australia, and the host of the 1956 Summer Olympics and the 2006 Commonwealth Games...
Question	▷ What is the largest stadium in Australia?
Answer	◊ Melbourne Cricket Ground
DAANET	▷ what is largest stadium in australia ? ◊ melbourne cricket ground
mono	▷ what is largest gross state product (AFL)? ◊ the melbourne cricket ground
Context	Sky UK Limited (formerly British Sky Broadcasting or BSkyB) is a British telecommunications company which serves the United Kingdom. Sky provides television and broadband internet services and fixed line telephone services to consumers and businesses in the United Kingdom. It is the UK’s largest pay-TV broadcaster with 11 million customers as of 2015...
Question	▷ Sky UK Limited is formerly known by what name?
Answer	◊ British Sky Broadcasting
DAANET	▷ what is name of sky uk limited? ◊ british sky broadcasting or BSkyB
mono	▷ what is canada’s largest pay-TV service? ◊ BSkyB

Table 4: Selected outputs from DAANET and mono. Yellow text is question-related; green text is answer-related.

Reading wikipedia to answer open-domain questions. In *ACL*, 1870–1879.

- Denkowski, M., and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Statistical Machine Translation@EACL*.
- Dhingra, B.; Pruthi, D.; and Rajagopal, D. 2018. Simple and effective semi-supervised question answering. In *NAACL-HLT*.
- Du, X.; Shao, J.; and Cardie, C. 2017. Learning to ask: Neural question generation for reading comprehension. In *ACL*.
- Duan, N.; Tang, D.; Chen, P.; and Zhou, M. 2017. Question generation for question answering. In *EMNLP*, 866–874.
- Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. 2017. Convolutional sequence to sequence learning. In *ICML*.
- Gers, F. A.; Schmidhuber, J.; and Cummins, F. A. 2000. Learning to forget: Continual prediction with lstm. *Neural Computation* 12:2451–2471.
- Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 249–256.
- Graves, A. 2013. Generating sequences with recurrent neural networks. *arXiv:1308.0850*.
- Heilman, M. 2011. Automatic factual question generation text.
- Hu, M.; Peng, Y.; Huang, Z.; Qiu, X.; Wei, F.; and Zhou, M. 2018. Reinforced mnemonic reader for machine reading comprehension. In *IJCAI*, 4099–4106.
- Huang, H.-Y.; Zhu, C.; Shen, Y.; and Chen, W. 2017. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. *arXiv:1711.07341*.
- Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*. Vancouver, Canada: Association for Computational Linguistics.
- Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.
- Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.
- Liu, M.; Calvo, R. A.; and Rus, V. 2010. Automatic question generation for literature review writing support. In *International Conference on Intelligent Tutoring Systems*, 45–54. Springer.
- Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*, 1412–1421.
- Nallapati, R.; Zhou, B.; dos Santos, C. N.; aglar Gülehre; and Xiang, B. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL*.
- Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; and Deng, L. 2016. Ms marco: A human generated machine reading comprehension dataset.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318. Association for Computational Linguistics.
- Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.
- Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*.
- Sachan, M., and Xing, E. 2018. Self-training for jointly learning to ask and answer questions. In *NAACL*, volume 1, 629–640.
- See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.
- Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2016. Bidirectional attention flow for machine comprehension. *arXiv:1611.01603*.
- Serban, I. V.; García-Durán, A.; Gülçehre, Ç.; Ahn, S.; Chandar, S.; Courville, A. C.; and Bengio, Y. 2016. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. In *ACL*.
- Shen, Y.; Huang, P.-S.; Gao, J.; and Chen, W. 2017. Reasonet: Learning to stop reading in machine comprehension. In *ACM SIGKDD*, 1047–1055. ACM.
- Song, L.; Wang, Z.; and Hamza, W. 2017. A unified query-based generative model for question generation and question answering. *arXiv:1709.01058*.
- Srivastava, R. K.; Greff, K.; and Schmidhuber, J. 2015. Highway networks. *arXiv:1505.00387*.
- Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NIPS*, 3104–3112.
- Tan, C.; Wei, F.; Yang, N.; Du, B.; Lv, W.; and Zhou, M. 2017. S-net: From answer extraction to answer generation for machine reading comprehension. *arXiv:1706.04815*.
- Tang, D.; Duan, N.; Qin, T.; Yan, Z.; and Zhou, M. 2017. Question answering and question generation as dual tasks. *arXiv:1706.02027*.
- Tang, D.; Duan, N.; Yan, Z.; Zhang, Z.; Sun, Y.; Liu, S.; Lv, Y.; and Zhou, M. 2018. Learning to collaborate for question answering and asking. In *NAACL*, volume 1, 1564–1574.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*, 5998–6008.
- Vinyals, O.; Kaiser, Ł.; Koo, T.; Petrov, S.; Sutskever, I.; and Hinton, G. 2015. Grammar as a foreign language. In *NIPS*, 2773–2781.
- Wang, S., and Jiang, J. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv:1608.07905*.
- Wang, W.; Yang, N.; Wei, F.; Chang, B.; and Zhou, M. 2017. Gated self-matching networks for reading comprehension and question answering. In *ACL*, volume 1, 189–198.
- Wang, T.; Yuan, X.; and Trischler, A. 2017. A joint model for question answering and question generation. *arXiv preprint arXiv:1706.01450*.
- Yang, Z.; Hu, J.; Salakhutdinov, R.; and Cohen, W. W. 2017. Semi-supervised qa in generative domain-adaptive nets. In *ACL*.
- Yang, Y.; Yih, W.-t.; and Meek, C. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *EMNLP*, 2013–2018.
- Yu, A. W.; Dohan, D.; Luong, M.-T.; Zhao, R.; Chen, K.; Norouzi, M.; and Le, Q. V. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv:1804.09541*.
- Yuan, X.; Wang, T.; aglar Gülehre; Sordani, A.; Bachman, P.; Subramanian, S.; Zhang, S.; and Trischler, A. 2017. Machine comprehension by text-to-text neural question generation. In *Rep4NLP@ACL*.