IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 28, 2021, accepted March 16, 2021, date of publication March 26, 2021, date of current version April 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068993

Contextual and Semantic Fusion Network for Multiple-Choice Reading Comprehension

QIANWEI DUAN[1,2], **JUN HUANG**[2], **AND HUIYAN WU**[2]
[1]School of Microelectronics, University of Chinese Academy of Sciences, Beijing 100049, China
[2]Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China

Corresponding author: Jun Huang (huangj@sari.ac.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFC1521202.

ABSTRACT Multiple-choice reading comprehension (MCRC) aims to build an intelligent system that automatically selects an answer from a candidate set when given a passage and a question. Existing MCRC systems rarely consider incorporating external knowledge such as explicit semantic information. In this work, we propose a Contextual and Semantic Fusion Network (CSFN) which effectively integrates contextual and semantic representation. CSFN introduces explicit structured semantics from pre-trained semantic role labeling. Specially, we regard explicit semantic representation as an important feature to fuse with contextual representation, which enriches the representation of sentences. By combining with the transfer learning strategy, the CSFN model has better generalization over limited datasets. To evaluate the ability of our model, we conduct experiments on three MCRC benchmark datasets: RACE, DREAM, and MCTest. Experimental results demonstrate the effectiveness of our proposed model.

INDEX TERMS Gated mechanism, machine reading comprehension, pre-trained language model, semantic role labeling.

I. INTRODUCTION

Reading comprehension gives human beings the ability to read, process, and understand text [1]–[3]. We test whether one understands the text by asking him or her questions related to the given context. Similarly, when machines are required to comprehend text, they need to answer questions according to the given context.

Compared with other machine reading comprehension tasks where answers are usually text spans from given passages [1], the MCRC task puts no limits on answer types. Instead, the candidate options are human-generated sentences. Table 1 shows an example from one of the mainstream MCRC datasets, RACE [4].

Recently, well pre-trained language models have achieved state-of-the-art results on various machine reading comprehension (MRC) tasks because they provide high-quality word representations with context-sensitive information (e.g., Embedding from Language models (ELMO) [5], Bidirectional Encoder Representations from Transformers (BERT) [6], and Generalized Autoregressive Pretraining (XLNET) [7]). Existing approaches based on the pre-trained language model can be divided into two categories: pre-training a powerful language model by imparting general knowledge from external corpora and fine-tuning the pre-trained model on a specific task. Although training a better language model is helpful, it is also time-consuming and resource-demanding. For instance, training a 12-layer transformer requires eight P100 GPUS for 4 days. From a practical point of view, fine-tuning a pre-trained model is resource-saving. Besides, a variety of studies show that fine-tuning a pre-trained model has a great impact on MCRC performance [8]–[10].

Despite the success of these well pre-trained language models, lots of studies indicate that existing language models only lie in contextual features, which restricts the power of the pre-trained representations [11]. Moreover, in the question answering (QA) domain, plenty of answers produced by previous models are semantically incomplete, which indicates that the pre-trained language models suffer from the problem of incomplete semantics [12]. However, existing methods seldom employ explicit semantic clues, which motivates us to introduce explicit structured semantics to relieve the problem in MCRC.

It is easy for human beings to understand the meaning of a sentence based on semantic information. For example, given

The associate editor coordinating the review of this manuscript and approving it for publication was Amjad Gawanmeh.

TABLE 1. An example of passage, the related question, and options from RACE. The ground-truth answer is in bold.

Passage: *Although most weddings follow long-held traditions, there's still room for American individualism. For example, the usual place for a wedding is in a church. But some people get married outdoors in a scenic spot. A few even have the ceremony while skydiving or riding on horseback! The couple may invite hundreds of people or just a few close friends. They choose their style of colors, decorations, and music during the ceremony. But some things rarely change. The bride usually wears a beautiful long white wedding dress...*

Question: *Which of the following best shows American individualism?*

Options:

A. Most weddings observe long-held traditions.

B. Some people choose their own style of weddings.

C. People choose a church as a place for a wedding.

D. The bride wears a beautiful long white wedding dress.

the sentence *David bought some French fries in McDonald's*, we can easily figure out that *bought* is a predicate, *David* is a buyer, *French fries* is a kind of food. It gives us a clue that semantic knowledge can help machine readers understand the meaning of a sentence [13].

Explicit structured semantics can be incorporated using semantic role labeling (SRL), which is a shallow semantic parsing task aiming to annotate the core predicate and related arguments (central meaning) of the sentence. In other words, the purpose of the SRL task is to discover *who does what to whom, when, and where*, which is close to some question forms in MCRC. With the central meaning of a sentence, machine readers can understand the text better. SRL has been proved to be beneficial to a variety of natural language processing (NLP) tasks including discourse relation sense classification [14], machine translation [15], and natural language inference [11]. All the studies indicate that SRL is potentially helpful for MRC task, which is ignored in previous works.

To our knowledge, some studies have incorporated external semantic information to help their models better understand natural language text. Guo *et al.* [13] employ semantic information by modeling lexical units based on the FrameNet [16] knowledge base whose semantic label form is completely different from the PropBank frame [17]. We adopt the PropBank-style semantic frame because it can cover every verb in a sentence. Zhang *et al.* [12] regard semantic embedding as a kind of position information and concatenate it with contextual embedding in natural language inference (NLI) task, which ignores the actual meaning of the semantic labels.

This paper makes the first attempt to employ the PropBank-style semantic labels to MCRC. Moreover, we regard semantic information as a separate feature to fuse with the contextual representation because it reflects the central meaning of a sentence. Also, considering that different semantic labels are of different importance (for example, predicate-argument tuples represent the core meaning of a sentence and contain more important information than the non-argument labels), we obtain the semantic representation by a weighted sum of the semantic embedding. The weights are the degree to which the network attends to a particular semantic label.

Besides, the ability of the pre-trained language model is limited in some datasets because of data insufficiency. To alleviate the problem, we follow the philosophy of transfer learning. Similar to Jin *et al.* [18], we divide the transfer learning process into two stages. The first stage is coarse-tuning the model in the natural language inference (NLI) task. The second stage is multi-task learning, which uses two datasets (large-scale source dataset and limited target dataset) to fine-tune the model. Specially, different from Jin *et al.* [18], we share not only the pre-trained language model parameters but the semantic embedding parameters for two datasets in the stage of multi-task learning. The sharing of semantic embedding parameters further promotes the formation of strong semantic representation. By combining CSFN with multi-stage and multi-task (MM) strategies, we achieve new state-of-the-art results on several representative datasets. The key contributions of this paper are as follows:

- We effectively introduce the explicit structured semantics to the MCRC task by proposing a contextual and semantic fusion network (CSFN).
- We design a Contextual-to-Semantic (C2S) fusion to help the model obtain separate semantic representation. Then, by combining the separate semantic and contextual representation, the sentence representation is effectively facilitated.
- We combine the CSFN model with multi-stage and multi-task learning strategies. Eventually, the proposed model achieves a remarkable improvement of 3.3% to 27.0% in several datasets from directly BERT-Base.

II. RELATED WORK

A. THE INTRODUCTION OF LARGE-SCALE DATASETS

The success of recent MRC systems is mainly due to the emergence of large-scale datasets. There are several distinct formats of MRC datasets that are different in answer forms. The answers of extractive MRC datasets are text spans from the given passage (e.g., CNN/Daily Mail [19], SQuAD [20], and NewsQA [21]). The answers of abstract MRC datasets are free-generated by human beings based on given passage (e.g., MS MARCO [22], QuAC [23], CoQA [24]). However, as annotators tend to directly copy spans as answers, the majority of answers are still extractive [24]. The answers of MCRC datasets are from a set of candidate options (e.g., RACE [4], DREAM [25], MCTest [26], MultiRC [27]). Generally, the MCRC task requires more advanced reasoning skills and is closest to the setting of human reading comprehension because it does not restrict the answers to text spans. Besides, the answers of MCRC are in the form of open, which allows rich question types such as common sense, logical reasoning, and summarization. So it is challenging for machine readers to perform well in the MCRC task.

B. METHODS FOR MULTIPLE-CHOICE MACHINE READING COMPREHENSION

We mainly study the approaches applied to MCRC tasks. Typical MCRC systems based on the recurrent neural network (RNN) mostly focus on exploring the relationship among passage, question, and option with a matching module called attention in the neural network community. Yin *et al.* [28] first concatenate the question and candidate option and then match them against the passage by attention. Lai *et al.* [4] first-step match the passage against the question and then select an answer using the first matching result. Wang *et al.* [29] simultaneously match question and option to the given passage.

Recent MCRC systems based on the pre-trained language model also focus on modeling the relationship among passage, question, and option. Zhang *et al.* [30] model the passage-question, passage-option, and question-option pair-wise relationship simultaneously and bidirectionally for each triplet. Imitating human beings, Ran *et al.* [31] compare options with question information to identify the options correlations, and then reread the passage with the option correlation features.

However, these models only capture the relationship among passage, question, and option at the word level and rarely take deep sentence semantics into consideration, which limits the model performance. Therefore, we aim to tackle the MCRC problem in a way that resembles how humans solve it: using semantic knowledge.

C. SEMANTIC ROLE LABELING

There are several semantic frames including PropBank [17] and FrameNet [16]. The PropBank-style semantic frame is widely used in academia and industry because it can cover every verb in a sentence. The structural properties (including V (predicate verb), ARG0 (prototypical agent), ARG1 (Prototypical Patient or Theme), ARG2 (scope of the predicate), and O (non-argument word)) are unique to the PropBank-style semantic frame. By SRL, we can obtain the semantic relationship and then grasp the central meaning of the given text. Fig. 1 shows an example of PropBank-style SRL, given the text [*Product and geography are what make cream skimming work*], there are three predicate-argument sequences:

[ARG1: Product and geography] [V: are] [ARG2: what make cream skimming work].

[O: Product and geography are] [ARG0: what] [V: make] [ARG1: cream skimming work].

[O: Product and geography are what make] [ARG1: cream] [V: skimming] [ARG0: work].

V and O represent the predicate and non-argument words. ARG0 represents the argument showing features of a Prototypical Agent, while ARG2 represents a Prototypical Patient or Theme [32].

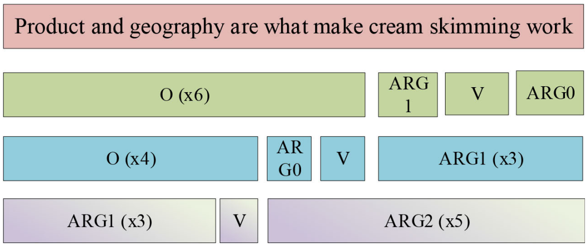


FIGURE 1. Example of the semantic role labeling. Three label sequences are corresponding to three different predicates. (V: Predicate verb; ARG0: Prototypical Agent of the verb; ARG1: Prototypical Patient or Theme; ARG2: scope of the predicate; O: non-argument word).

D. TRANSFER LEARNING FOR MULTIPLE-CHOICE MACHINE READING COMPREHENSION

Transfer learning [33] is an important technique in machine learning, which learns knowledge from one task and then applies the knowledge to another related task. This technique has been widely used in a number of domains including computer vision [34], automatic speech recognition [35], [36], and natural language processing [37]. To the best of our knowledge, Guolub *et al.* [38] first apply transfer learning to machine reading comprehension task through a method of unsupervised transfer learning. The process of transfer learning includes two steps. The first step is to pre-train the model in one source dataset with rich training data. The second step is to fine-tune the model in another limited target dataset [39]. Sun *et al.* [40] exploit this method and achieve state-of-the-art performance in MCRC. Besides, Jin *et al.* [18] extend the transfer learning approach by applying the out-of-domain dataset to coarse-tune the model and then fine-tune the model on the source and target datasets simultaneously via multi-task training. Inspired by these studies, we have designed multi-stage and multi-task strategies to improve the generalization of our model.

III. METHOD

In this section, we first briefly introduce the task definition and describe the framework of our CSFN model (Section III-A). Then, we elaborate on the contextual and semantic fusion layer (Section III-B). Finally, we introduce the combination of the CSFN model with multi-stage and multi-task strategies (Section III-C).

A. TASK DEFINITION AND MODEL ARCHITECTURE

The main layers of our model are the encoder layer, contextual and semantic fusion layer, and prediction layer. The whole framework is shown in Fig. 2.

1) TASK DEFINITION

In the scenario of MCRC, given a passage, a question, and a candidate options set, our model needs to select the correct answer from the candidate options set. Formally, the MCRC datasets can be described as a triple (P, Q, O). $P = \{P_1, P_2, \ldots, P_m\}$ is the passage with m sentences and

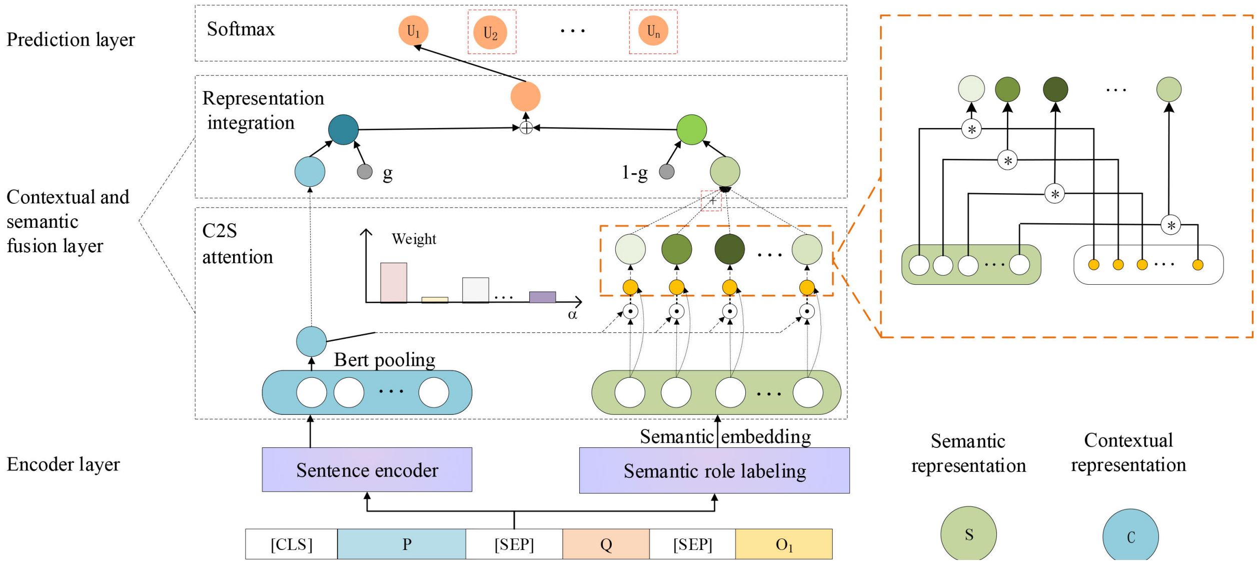


FIGURE 2. **The framework of our Contextual and Semantic Fusion Network (CSFN). P: passage, Q: question, O1: the first option. U2, . . . , Un are obtained in the same procedure as U1. CSFN comprises three parts: encoder layer, contextual and semantic fusion layer, and prediction layer. Among them, the encoder layer includes sentence encoder, semantic role labeling, and semantic embedding. The C2S denotes the contextual-to-semantic attention. The representation integration component will balance the information flow between contextual representation and semantic representation. The circles in yellow represent normalized weight values.**

each sentence $P_m = \{w_t^{P_m}\}_{t=1}^{l_{pm}}$ is composed of a sequence of words $w_t^{P_m}$. $Q = \{w_t^{Q}\}_{t=1}^{l_q}$ is the question which is composed of a sequence of words w_t^{Q}. $O = \{O_1, O_2, \ldots, O_n\}$ is the candidate options set with n options and each option $O_n = \{w_t^{O_n}\}_{t=1}^{l_{on}}$ is composed of a sequence of words $w_t^{O_n}$.

2) ENCODER LAYER

a: SENTENCE ENCODER

We concatenate the passage, question, and one of the options into a long sequence. For each option O_n, we can obtain the corresponding input sequence, which is denoted as (P, Q, O_n). Afterward, the sequence is encoded as follows:

$$C = BERT(P; Q; O_n) \tag{1}$$

where $C \in R^{L\times d}$ is the contextual representation of the input sequence. d and L are the dimension of the hidden state and the sequence length, respectively. (;) denotes the concatenation operation. We use the pre-trained transformer model BERT as the sentence encoder because it is powerful for language representation. More information about BERT is detailed in Devlin *et al.* [6].

b: SEMANTIC ROLE LABELING

Semantic role labeling is used to obtain the semantic labels of input sequences that represent the relation between predicate and argument. We take the PropBank-style semantic frame to annotate each sentence with role labels because it covers every verb in the sentence [17]. Corresponding to different predicates, SRL will produce different predicate-argument sequences. As shown in Fig. 1, there are three predicate-argument sequences corresponding to three predicates: *are*, *make*, and *skimming*, respectively. We conjecture that different predicate-specific argument sequence reflects different sentence meanings from different perspectives.

The input of our model can be further represented as $(P_1, \ldots, P_m, Q, O_n)$, which contains more than one sentences. To get the overall semantic labels of the input sequence, we choose the predicate-specific argument sequence with the least O labels (O represents the non-argument label, which is different from the option O) for each sentence because it covers the most semantic information.

Formally, We define the semantic labels of i_{th} sentence as follows:

$$\forall_k, \quad S_i = min\{(M_o)_i^k\}, \; i \in [1..m+2] \tag{2}$$

where $(M_o)_i^k$ is the number of the non-argument word O in k_{th} semantic structured sequence of i_{th} sentence. Next, we concatenate the $m+2$ semantic label sequences as follows:

$$S = (S_1; S_2, \ldots, S_m; S_Q; S_O) \tag{3}$$

c: SEMANTIC EMBEDDING

Semantic embedding is responsible for mapping each label to a high-dimensional vector space. We map the whole semantic labels to vector embedding $e \in R^{L\times h}$ (h and L are the dimension of embedding and length of input sequence, respectively) by looking up the mapping table. Then, we feed the embedding to a linear layer to obtain the final label representations:

$$E = ReLU(W_1 e + b_1) \tag{4}$$

where $W_1 \in R^{h \times h}$, $b_1 \in R^h$ are learnable weight and bias. *ReLU* is the Rectified Linear Unit.

3) CONTEXTUAL AND SEMANTIC FUSION LAYER

The contextual and semantic fusion layer includes two components: C2S attention and representation integration. Contextual-to-Semantic (C2S) attention is designed for obtaining the enhanced semantic representation which contains the information of context. Next, the representation integration layer is utilized to integrate the contextual and the enhanced semantic representations. More formulation about these two components is detailed in SecIII-B.

4) PREDICTION LAYER

The probability of option O_k to be the correct answer is calculated as follows:

$$P(k|P,Q,A) = \frac{exp(O_k)}{\sum_{i=1}^{n} exp(O_k)} \tag{5}$$

We optimize our model with cross-entropy loss. The loss function is defined as follows:

$$J(\theta) = \frac{1}{N}\sum_{i=1} log(P(k_i|P_i,Q_i,O_i)) + \lambda||\theta||^2 \tag{6}$$

where k_i is the ground truth answer. θ, N, and λ denote all trainable parameters, the number of training examples, and the regularization coefficient, respectively.

B. CONTEXTUAL AND SEMANTIC FUSION LAYER

1) THE CONTEXTUAL-TO-SEMANTIC ATTENTION

We use C2S attention to exploit semantic information. For convenience, we denote the contextual representation and the semantic label embedding as $C \in R^{L \times d}$ and $E \in R^{L \times h}$, respectively. Here we introduce two methods to compute the weight for every label in the predicate-specific argument sequence.

- **Dot product:** The model computes the weight between contextual representation and every semantic embedding. Then, the enhanced semantic representation is obtained by a weighted sum of the semantic embedding. To form a standard probability distribution for each label, we use a softmax function to normalize the weights.

$$\widehat{C} = W_2 \cdot BertPooling(C) + b_2 \tag{7}$$
$$\alpha = softmax(\widehat{C} \cdot E) \tag{8}$$
$$\dot{E} = \alpha^T \cdot E \tag{9}$$

  where $W_2 \in R^{d \times h}$ and $b_2 \in R^d$ are learnable weight and bias. $\alpha \in R^{L \times h}$ is the attention weight of semantic labels. $\cdot$ is the dot product operation. $\dot{E} \in R^h$ is the enhanced semantic representation.

- **Bilinear attention:** Inspired by Chen *et al.* [2], we compute the weights by the bilinear method:

$$\widehat{C} = W_3 \cdot BertPooling(C) + b_3 \tag{10}$$

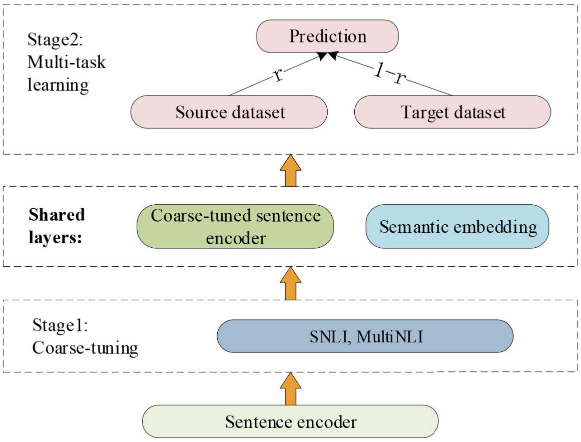


FIGURE 3. Setting of multi-stage and multi-task training. All parameters are shared for source and target datasets including sentence encoder and semantic embedding.

$$\alpha_i = softmax_i(\widehat{C}^T W_4 E_i) \tag{11}$$
$$\dot{E} = \sum_i \alpha_i E_i \tag{12}$$

  where $W_3 \in R^{d \times h}$, $b_3 \in R^d$, and $W_4 \in R^{h \times d}$ are three learnable parameters. α_i is the i_{th} weight of semantic label of the input sequence.

2) REPRESENTATION INTEGRATION

In this section, we integrate semantic representation and contextual representation. Integrating features from different aspects have been proved to be effective in practice. Inspired by Srivastava *et al.* [41], we exploit a gate mechanism to control the information flow between the enhanced semantic representation and the contextual representation. Then, we obtain the ultimate representation $U \in R^1$ which owns abundant contextual and semantic information.

$$g = \sigma(W_5\dot{E} + W_6 C + b) \tag{13}$$
$$U = g * \dot{E} + (1-g) * C \tag{14}$$

where $W_5 \in R^{h \times L}$, $W_6 \in R^{d \times L}$, and $b \in R^L$ are three learnable parameters. $g \in R^L$ is a reset gate.

C. COMBINATION OF CSFN AND MM

In this section, we employ the multi-stage and multi-task strategies. The architecture is shown in Fig. 3 which includes two steps: coarse-tuning and multi-task learning. **In the coarse-tuning stage**, we tune the sentence encoder in the NLI task because it can provide the model with language inference ability [42]. Later, knowledge learned by the model can be transferred from the NLI task to the machine reading comprehension task.

Several MCRC datasets are small-in-size and suffer from data insufficiency, which limits the potency of large pre-trained models. For instance, the performance of the BERT-Base encoder in MC500 is about 8% higher than MC160 because the former size is three times larger than

TABLE 2. Statistics of multiple-choice reading comprehension datasets. These values are from [4], [25], and [26]. (crowd.: crowd-sourcing; *: correct answer options that are not text snippets from reference documents).

	of passages	of questions	of options	construction method	sources of documents	non-extractive answer* (%)
RACE	27,933	97,687	4	exams	general	87.0
DREAM	6,444	10,197	3	exams	dialogues	83.7
MCTest	660	2640	4	crowd.	stories	45.3

the latter. Besides, some datasets are designed for specific passage types which are too monotonous, e.g., MCTest is about children's story whereas RACE is more general and includes a wide variety of topics. To tackle the above problems, the strategy of multi-task learning is proposed. **In the multi-task learning stage**, we use source dataset and target dataset to train our model. It's worth pointing out that the source dataset is RACE which owns sufficient training data and covers plenty of passage types. The knowledge learned from the source dataset can benefit other target datasets via multi-task learning. The parameters of transformer encoder, semantic embedding, and top-level classifier are shared across two different datasets.

The dataset for the current training step is selected according to the preordained probability. Formally, we let X and Y denote the random event of choosing the source dataset and target dataset, respectively. Both of them have two possible results. $X = 1$ and $Y = 1$ mean random event happens. $X = 0$ and $Y = 0$ mean the opposite one. Because choosing source dataset and target dataset are two mutually antagonistic events, the probability of the event X and Y can be defined as follows:

$$P_r = \begin{cases} r\% & X = 1 \text{ or } Y = 0 \\ 1 - r\% & X = 0 \text{ or } Y = 1 \end{cases} \tag{15}$$

where r represents the occurrence probability of event X (nonoccurrence probability of event Y).

Next, we randomly select a batch of data from the selected dataset to train our model. The process is repeated until the maximum (t) of the training step is met and it can be regarded as the Binomial Distribution. We let Z represents the occurrence frequency of event X, and the possible value of Z is $(0, 1, \ldots, t)$. The implicate meaning of event $\{Z = k\}$ is that there are k batches of data from the source dataset and $t-k$ batches of data from the target dataset. Model parameters are updated by these selected batches. The probability of event $\{Z = k\}$ is calculated as follows:

$$P_r\{Z = k\} = C_t^k r^k (1 - r)^{t-k} \tag{16}$$

where $C_t^k = \frac{t!}{k!(t-k)!}$ is the Combination Number Formula.

IV. EXPERIMENTS

A. DATASET

We use three MCRC datasets for our experiment: RACE, DREAM, and MCTest. Statistics of these datasets are summarized in Table 2.

- **RACE** [4] is a large-scale reading comprehension dataset with 97,687 questions for 27,933 passages. This dataset is collected from the English exams for students in middle and high schools, and it covers a variety of topics for carefully evaluating the reading comprehension and reasoning ability of students. Besides, the proportion of questions that requires reasoning is higher than other benchmark datasets. So this dataset is recognized as the most typical dataset in MCRC.
- **DREAM** [25] is a dialogue-based reading comprehension dataset with 10,197 questions for 6,444 dialogues. This dataset is collected from the English-as-a-foreign-language examinations designed by human experts. DREAM is extremely challenging as 85% of questions require reasoning with two or more sentences and 34% of questions require common knowledge.
- **MCTest** [26] is a crowd produced reading comprehension dataset with 2,640 questions for 660 stories. This dataset is gathered from fictional stories that can be easily understood by a young child at the age of 7. MCTest is also challenging as over 50% of questions require reasoning with two or more sentences. MCTest contains two variants: MC160 and MC500. MC500 is considered more difficult than MC160.

B. IMPLEMENTATION DETAILS

Our model is based on the pre-trained model BERT-Base [6]. BERT-Base model has 110M parameters with 12-layer transformer blocks, 12 self-attention heads, and 768 hidden-size. In our experiment, the max length of the input sequence is set to 512. The dropout rate is set to 0.1 for each BERT layer and the optimizer is BertAdam.

For the hidden size of the semantic label sequence, we follow the setting of Zhang *et al.* [12]. The learning rate, number of training epochs, and batch size are different in different datasets. For the RACE dataset, the learning rate, the epoch, and the batch size are set to 2e-5, 5, and 16, respectively. For the DREAM dataset, the learning rate, the epoch, and the batch size are set to 2e-5, 8, and 16, respectively. For the MCTest dataset, the learning rate, the epoch, and the batch size are set to 1e-5, 8, and 16, respectively. We use vanilla stochastic gradient descent (SGD) to train our model. For all datasets, the model is trained on two 2080Ti GPUs.

C. EVALUATION ON RACE

In Table 3, we first report the results of the state-of-the-art models in the leaderboard. Then, we report the performance

TABLE 3. Accuracy on the RACE dataset. All the results are from the single model. (*: our implementation; MM [18]: multi-stage and multi-task strategies).

Model	DEVELOPMENT	TEST
HCM [29]	-	50.4
MMN [43]	57.4	54.7
RSM [40]	-	63.8
BERT-Base [6]	66.9	65.0
DCMN+ [30]	67.4	67.0
OCN [31]	-	66.8
MM*	69.1	67.6
MMM [18]	-	68.0
Our Models		
CSFN	67.6	67.7
CSFN + MM	**69.7**	**68.3**

of the CSFN model and the combination of CSFN with multi-stage and multi-task strategies (CSFN + MM). The results are shown on the RACE dataset.

The comparison indicates that our method obtains significant improvement over pre-trained language model BERT (65.0% vs. 67.7% on BERT-Base). Our CSFN model also outperforms the models DCMN+ and OCN by 0.7% and 0.9%, respectively. It's worth noting that both the models (DCMN+ and OCN) have used sophisticated architecture to model the relationship among passage, question, and option. As CSFN brings a more remarkable improvement, we can reasonably conclude that semantic information can facilitate sentence representation and enhance machine reading comprehension.

It should be noted that MMM [18] shown in the table contains MM strategies and a multi-step attention network (MAN). MAN is used to model the relationship among passage, question, and option. Here we combine the CSFN model with MM strategies. In the first stage, we directly use the coarse-tuned model with the NLI task, as NLI has been proved to be beneficial for text understanding. In the second stage, we simultaneously train the largest source dataset RACE and limited target dataset DREAM by sharing parameters of the coarse-tuned sentence encoder and semantic embedding. For a more direct comparison, we report the result of the MM strategy (our implementation). From Table 3, we can see that our method (CSFN + MM) outperforms MM by 0.7%. The improvement is more remarkable than MAN (0.7% vs. 0.4%), which indicates that introducing explicit semantic information can enhance machine reading comprehension in the scenario of transfer learning.

D. EVALUATION ON OTHER MULTIPLE-CHOICE DATASETS

The results of our model on three small datasets (MC160, MC500, DREAM) are shown in Table 4. In the same way, we first report the results of previous state-of-the-art models for comparison. Here we give the results of directly fine-tuning the BERT-Base encoder on these datasets. The accuracy are 58.8%, 67.2%, and 63.2%, respectively. It should be noted that the batch size is set to 16 because of limited computing resource (two 2080Ti GPUs), which leads to some decreases (58.8% vs. 63.8% on MC160 and 67.2% vs. 71.3% on MC500) compared to the results reported in Jin *et al.* [18].

TABLE 4. Accuracy on other MCRC datasets. All the results are from the single model. (*: our implementation; MM [18]: multi-stage and multi-task strategies).

Model	MC160	MC500	DREAM	**Average**
BERT-Base*	58.8	67.2	63.2	63.1
RSM [40]	80.0	78.7	-	-
BERT-Base (Jin *et al.* [18])	63.8	71.3	63.2	66.1
QACNN [39]	76.4	72.3	-	-
FSR [13]	**86.1**	**84.2**	-	-
MM*	82.9	81.0	70.8	78.2
MMM [18]	85.4	82.7	72.2	80.1
Our Models				
CSFN	65.8	68.0	64.0	65.9
CSFN + MM	85.8	82.7	**72.5**	**80.3**

TABLE 5. Ablation study on the datasets: RACE, MC160, MC500, and DREAM. The number in the parenthesis is the increases in accuracy from using the previous component to the current one. Accuracy is on the development set.

	RACE	MC160	MC500	DREAM
baseline	66.6	55.0	67.0	62.7
(a) + *sem*	67.4 (**+0.8**)	57.5 (+2.5)	69.5 (**+2.5**)	63.1 (+0.4)
(b) + *sem* + C2S	**67.6** (+0.2)	**63.3 (+5.8)**	**72.0 (+2.5)**	**64.1 (+1.0)**
(c) + *sem* + C2S + MM	69.7 (+2.1)	83.3 (+20)	84.5 (+12.5)	72.9(+8.8)

From the comparison, we can see that our model gets a 2.8% improvement in average accuracy over the baseline of directly fine-tuning BERT-Base (65.9% vs. 63.1%), which shows that employing explicit semantic information can improve the performance of pre-trained models (which only utilize contextual representation). The models (RSM, QACNN, and MMM) are all based on transfer learning. To compare with them in the same condition, we combine CSFN with MM strategy. The result shows that our model (CSFN + MM) exceeds the previous best model (MMM) by 0.3% on average. The best results on MC160 and MC500 are achieved by the FSR model based on BERT-Large, which is not comparable with our model (based on BERT-Base). Despite this, the result of our model is very close to theirs (85.8% vs. 86.1%) on the MC160 dataset.

E. ABLATION STUDY

To further analyze the performance, we conduct an ablation study to observe the contribution of main components (semantic information (+*sem*), C2S attention, and MM). In Table 5, +*sem* means that we leave out the procedure of C2S attention in Fig. 2 and the separate semantic representation is obtained by summing up all semantic embedding of the input sequence.

From the baseline and Row (a), we can see that introducing explicit semantic information improves the model performance significantly.

TABLE 6. The comparison of different attention computing methods. Accuracy is on the MC160 development set.

Weight	Dot-C2S	Bilinear-C2S
0.1	44.1	52.5
0.2	55.8	57.5
0.3	60.0	55.8
0.4	53.3	55.0
0.5	55.8	58.3
0.6	57.5	60.0
0.7	55.0	60.0
0.8	57.5	58.3
0.9	55.8	57.5

From row (a) and row (b), we have the following observations. For RACE, the main contribution is from the semantic information +*sem*. For MC160 and DREAM, the main contribution is from C2S attention. For MC500, +*sem* and C2S attention promote the model performance equally. It is obvious that +*sem* gives the main contribution to the source dataset (RACE) and C2S attention gives the main contribution to target datasets (DREAM, MC160, and MC500). The results indicate that the semantic representation trained only by target datasets is incomplete due to insufficient data and can be further enhanced by incorporating contextual information (C2S).

Row (c) shows that the combination of MM strategies and other components (+*sem*, C2S attention) greatly improves the performance of target datasets, which indicates that the semantic representation can be further enhanced based on transfer learning.

F. RESULTS OF DIFFERENT FEATURE WEIGHTS

We regard semantic and contextual representation as two important features for MCRC. By adjusting the value of g (which represents the proportion of semantic representation), we explore the influence of these features on the development results of MC160. The final vector used for model prediction is defined as (14).

Table 6 shows the performance comparison with different attention computing methods. We observe that the performance achieved by the bilinear method is better than the dot method in almost all groups of the value of g. The possible reason is that the bilinear method can generate a more relevant C2S matrix when computing contextual-to-semantic attention.

Fig. 4 shows the results of different value of g. To form a direct contrast, we also give the results without C2S attention, where the semantic representation is obtained by summing up all semantic embedding (+*sem*). We observe that the semantic information (+*sem*) brings a meaningful improvement compared to the baseline (only uses contextual representation). This indicates that the semantic representation provides some useful information for model prediction. Besides, the bilinear-C2S method performs better than +*sem* in most value of g, which indicates that bilinear-C2S can help the model obtain a more powerful semantic representation.

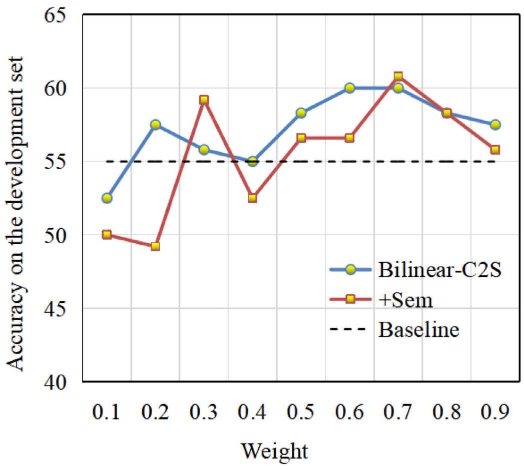
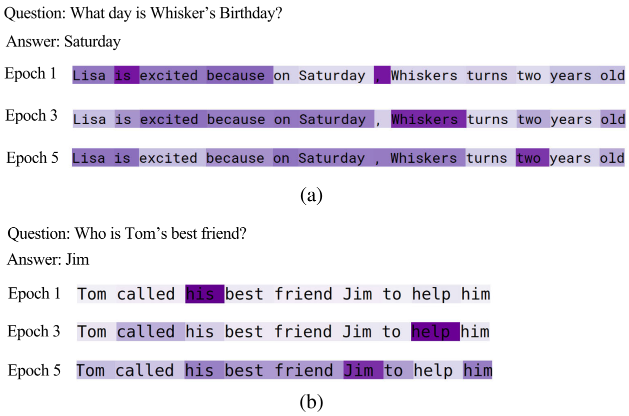


FIGURE 4. Results of different value of *g*. The x-axis is the value of *g*, and the y-axis is the corresponding accuracy on the MC160 development set. +*sem* means directly integrating contextual representation and semantic representation without C2S attention. The baseline represents only using contextual representation.



Question: What day is Whisker's Birthday?

Answer: Saturday

Epoch 1 Lisa is excited because on Saturday , Whiskers turns two years old

Epoch 3 Lisa is excited because on Saturday , Whiskers turns two years old

Epoch 5 Lisa is excited because on Saturday , Whiskers turns two years old

(a)

Question: Who is Tom's best friend?

Answer: Jim

Epoch 1 Tom called his best friend Jim to help him

Epoch 3 Tom called his best friend Jim to help him

Epoch 5 Tom called his best friend Jim to help him

(b)

FIGURE 5. Visualization of changes in C2S attention weights during Epoch 1, 3, and 5. The darker of the word, the more C2S attention views the word as a key feature. The input questions and answers are from MC160.

G. ATTENTION MAPS VISUALIZATION

To better understand how CSFN helps the keywords locating from the passage, we visualize the weights changes of C2S attention during Epoch 1, 3, and 5. In Fig. 5, the darker color shows the higher weights. The examples are from the MC160 dataset. The sentence fragments come from relevant passages of about 500 words. Here we leave aside the other redundant information in the context and analyze this sentence fragment in detail. As we can see, with the training epochs increase, the C2S focuses more on the words related to the correct answer. For instance, the ground truth answer in Fig. 5(a) is "Jim". In Epoch 1, the attention weight is randomly generated and the model does not focus on the right phrase "Jim", but the model gradually attends to the correct phrase in Epoch 3 and 5.

V. CONCLUSION AND FUTURE WORK

We propose a contextual and semantic fusion network (CSFN) to introduce explicit semantic information for multiple-choice reading comprehension (MCRC). Specially, our CSFN gets a separate semantic representation that can reflect the central meaning of a sentence. By combining the contextual and semantic representation, the sentence representation is effectively facilitated. Finally, to enhance the generalization of our model on limited datasets, we also combine it with multi-stage and multi-task (MM) strategies. Experimental results in several datasets demonstrate the effectiveness of our CSFN model.

This work indicates that explicit semantic knowledge can be effectively integrated into the pre-trained language model to enhance machine reading comprehension. Besides, our approach can be generalized to other semantic frames because the incorporation of semantic knowledge in our work is achieved by getting a separate semantic representation.

One limitation of this work is the heavy time cost of the semantic annotation. In future work, we will be committed to applying our model to other machine reading comprehension tasks and other pre-trained language models.

REFERENCES

[1] D. Khashabi, S. Min, T. Khot, A. Sabharwal, O. Tafjord, P. Clark, and H. Hajishirzi, ‘‘UNIFIEDQA: Crossing format boundaries with a single QA system,’’ in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2020, pp. 1896–1907.

[2] D. Chen, J. Bolton, and C. D. Manning, ‘‘A thorough examination of the CNN/daily mail reading comprehension task,’’ in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2016, pp. 2358–2367.

[3] Y. Cui, T. Liu, Z. Chen, S. Wang, and G. Hu, ‘‘Consensus attention-based neural networks for Chinese reading comprehension,’’ in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers COLING*, 2016, pp. 1777–1786.

[4] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, ‘‘RACE: Large-scale ReAding comprehension dataset from examinations,’’ in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 785–794.

[5] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, ‘‘Deep contextualized word representations,’’ in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., (Long Papers)*, vol. 1, 2018, pp. 2227–2237.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘‘BERT: Pre-training of deep bidirectional transformers for language understanding,’’ in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[7] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, ‘‘XLNet: Generalized autoregressive pretraining for language understanding,’’ in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5753–5763.

[8] Z. Zhang, Y. Wu, J. Zhou, S. Duan, H. Zhao, and R. Wang, ‘‘SG-Net: Syntax-guided machine reading comprehension,’’ in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, 2020, pp. 9636–9643.

[9] P. Zhu, H. Zhao, and X. Li, ‘‘DUMA: Reading comprehension with transposition thinking,’’ 2020, *arXiv:2001.09415*. [Online]. Available: http://arxiv.org/abs/2001.09415

[10] V. Ingale and P. Singh, ‘‘GenNet : Reading comprehension with multiple choice questions using generation and selection model,’’ 2020, *arXiv:2003.04360*. [Online]. Available: http://arxiv.org/abs/2003.04360

[11] Z. Zhang, Y. Wu, Z. Li, S. He, and H. Zhao, ‘‘I know what you want: Semantic learning for text comprehension,’’ 2018, *arXiv:1809.02794*. [Online]. Available: https://arxiv.org/abs/1809.02794

[12] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou, ‘‘Semantics-aware bert for language understanding,’’ in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, 2020, pp. 9628–9635.

[13] S. Guo, R. Li, H. Tan, X. Li, Y. Guan, H. Zhao, and Y. Zhang, ‘‘A frame-based sentence representation for machine reading comprehension,’’ in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 891–896.

[14] T. Mihaylov and A. Frank, ‘‘Discourse relation sense classification using cross-argument semantic similarity based on word embeddings,’’ in *Proc. CoNLL-16 Shared Task*, 2016, pp. 100–107.

[15] C. Shi, S. Liu, S. Ren, S. Feng, M. Li, M. Zhou, X. Sun, and H. Wang, ‘‘Knowledge-based semantic embedding for machine translation,’’ in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2016, pp. 2245–2254.

[16] C. F. Baker, C. J. Fillmore, and J. B. Lowe, ‘‘The berkeley FrameNet project,’’ in *Proc. 36th Annu. Meeting Assoc. Comput. Linguistics*, 1998, pp. 86–90.

[17] M. Palmer, D. Gildea, and P. Kingsbury, ‘‘The proposition bank: An annotated corpus of semantic roles,’’ *Comput. Linguistics*, vol. 31, no. 1, pp. 71–106, Mar. 2005.

[18] D. Jin, S. Gao, J.-Y. Kao, T. Chung, and D. Hakkani-Tur, ‘‘Mmm: Multi-stage multi-task learning for multi-choice reading comprehension,’’ in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, 2020, pp. 8010–8017.

[19] K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, ‘‘Teaching machines to read and comprehend,’’ in *Proc. 28th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 1, 2015, pp. 1693–1701.

[20] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, ‘‘SQuAD: 100,000+ questions for machine comprehension of text,’’ in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 2383–2392.

[21] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman, ‘‘NewsQA: A machine comprehension dataset,’’ in *Proc. 2nd Workshop Represent. Learn. NLP*, 2017, pp. 191–200.

[22] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang, ‘‘MS MARCO: A human generated MAchine reading COmprehension dataset,’’ 2016, *arXiv:1611.09268*. [Online]. Available: http://arxiv.org/abs/1611.09268

[23] E. Choi, H. He, M. Iyyer, M. Yatskar, W.-T. Yih, Y. Choi, P. Liang, and L. Zettlemoyer, ‘‘QuAC: Question answering in context,’’ in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2174–2184.

[24] S. Reddy, D. Chen, and C. D. Manning, ‘‘CoQA: A conversational question answering challenge,’’ *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 249–266, Nov. 2019.

[25] K. Sun, D. Yu, J. Chen, D. Yu, Y. Choi, and C. Cardie, ‘‘DREAM: A challenge data set and models for dialogue-based reading comprehension,’’ *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 217–231, Nov. 2019.

[26] M. Richardson, C. J. Burges, and E. Renshaw, ‘‘MCTest: A challenge dataset for the open-domain machine comprehension of text,’’ in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 193–203.

[27] D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay, and D. Roth, ‘‘Looking beyond the surface: A challenge set for reading comprehension over multiple sentences,’’ in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., (Long Papers)*, vol. 1, 2018, pp. 252–262.

[28] W. Yin, S. Ebert, and H. Schütze, ‘‘Attention-based convolutional neural network for machine comprehension,’’ in *Proc. Workshop Hum.-Comput. Question Answering*, 2016, pp. 15–21.

[29] S. Wang, M. Yu, J. Jiang, and S. Chang, ‘‘A co-matching model for multi-choice reading comprehension,’’ in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Short Papers)*, vol. 2, 2018, pp. 746–751.

[30] S. Zhang, H. Zhao, Y. Wu, Z. Zhang, X. Zhou, and X. Zhou, ‘‘Dcmn+: Dual co-matching network for multi-choice reading comprehension,’’ in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, 2020, pp. 9563–9570.

[31] Q. Ran, P. Li, W. Hu, and J. Zhou, ‘‘Option comparison network for multiple-choice reading comprehension,’’ 2019, *arXiv:1903.03033*. [Online]. Available: http://arxiv.org/abs/1903.03033

[32] D. Dowty, ‘‘Thematic proto-roles and argument selection,’’ *Language*, vol. 67, no. 3, pp. 547–619, 1991.

[33] S. Jialin Pan and Q. Yang, ‘‘A survey on transfer learning,’’ *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[34] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, ‘‘CNN features off-the-shelf: An astounding baseline for recognition,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 512–519.

[35] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, ‘‘Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,’’ in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7304–7308.

[36] M. Doulaty, O. Saz, and T. Hain, ‘‘Data-selective transfer learning for multi-domain speech recognition,’’ in *Proc. Interspeech*, Sep. 2015, pp. 2897–2901.

[37] Y. Zhang, R. Barzilay, and T. Jaakkola, ‘‘Aspect-augmented adversarial networks for domain adaptation,’’ *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 515–528, Dec. 2017.

[38] D. Golub, P.-S. Huang, X. He, and L. Deng, "Two-stage synthesis networks for transfer learning in machine comprehension," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 835–844.

[39] Y.-A. Chung, H.-Y. Lee, and J. Glass, "Supervised and unsupervised transfer learning for question answering," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., (Long Papers)*, vol. 1, 2018, pp. 1585–1594.

[40] K. Sun, D. Yu, D. Yu, and C. Cardie, "Improving machine reading comprehension with general reading strategies," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., (Long Short Papers)*, vol. 1, 2019, pp. 2633–2643.

[41] R. Kumar Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," 2015, *arXiv:1505.00387*. [Online]. Available: http://arxiv.org/abs/1505.00387

[42] S. Welleck, J. Weston, A. Szlam, and K. Cho, "Dialogue natural language inference," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3731–3741.

[43] M. Tang, J. Cai, and H. H. Zhuo, "Multi-matching network for multiple choice reading comprehension," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 7088–7095.

QIANWEI DUAN received the bachelor's degree in computer science from Beijing Normal University, China, in 2018. She is currently pursuing the master's degree with the University of Chinese Academy of Sciences. Her research interests include natural language processing, question answering, and machine reading comprehension.

JUN HUANG received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2011. He is currently an Associate Professor with the Shanghai Advanced Research Institute, Chinese Academy of Sciences. His research interests include natural language processing, pattern recognition, and media analysis.

HUIYAN WU received the bachelor's degree in information engineering from Fujian Normal University, China, in 2019. She is currently pursuing the master's degree with the University of Chinese Academy of Sciences. Her research interests include natural language processing, natural language inference, and machine reading comprehension.