




Incremental BERT with commonsense representations for multi-choice reading comprehension

Ronghan Li¹ · Lifang Wang¹ · Zejun Jiang¹  · Dong Liu¹ · Meng Zhao¹ · Xinyu Lu¹

Received: 14 August 2020 / Revised: 7 January 2021 / Accepted: 28 June 2021 /
Published online: 28 July 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Compared to extractive machine reading comprehension (MRC) limited to text spans, multi-choice MRC is more flexible in evaluating the model's ability to utilize external commonsense knowledge. On the one hand, existing methods leverage transfer learning and complicated matching networks to solve the multi-choice MRC, which lacks interpretability for commonsense questions. On the other hand, although Transformer based pre-trained language models such as BERT have shown powerful performance in MRC, external knowledge such as unspoken commonsense and world knowledge still can not be used explicitly for downstream tasks. In this work, we present three simple yet effective injection methods plugged in BERT's structure to fine-tune the multi-choice MRC tasks with off-the-shelf commonsense representations directly. Moreover, we introduce a mask mechanism for the token-level multi-hop relationship searching to filter external knowledge. Experimental results indicate that the incremental BERT outperforms the baseline by a considerable margin on DREAM and CosmosQA, two knowledge-driven multi-choice datasets. Further analysis shows the robustness of the incremental model in the case of an incomplete training set.

Keywords Machine reading comprehension · BERT · External knowledge · Common sense · Deep learning

1 Introduction

Machine Reading Comprehension (MRC) is a classic task in textual question answering (QA), where models are required to answer a natural language question given the relevant/irrelevant passages. Thanks to the release of large-scale datasets [17, 22, 25, 43], related end-to-end neural networks have achieved promising results in various scenarios [1, 7, 27, 36, 47]. Usually, MRC based question answering (QA) can be divided into three types of tasks: extractive QA, generative QA, and multi-choice QA. Compared to extractive MRC

✉ Zejun Jiang
claud@mail.nwpu.edu.cn

limited to text spans, multi-choice MRC allows more flexible design of multiple types of questions such as summarization, commonsense, logical reasoning, arithmetic, and sentiment analysis. Hence, most commonsense-based QA datasets are designed in a multi-choice form. For example, as shown in Fig. 1, the well-known fact that “*McDonalds*” is a restaurant is useful to find the correct option.

Existing multi-choice QA datasets are small in size, making previous methods focus on transfer learning with out-of-domain datasets and tasks [14, 33] or designing complicated matching networks [35, 47]. Nevertheless, more data and more parameters mean more computing resources are consumed. Besides, the out-of-domain data and the accumulation of model capacity can not solve the fact-based QA task well and explain the commonsense reasoning explicitly.

On the other hand, although pre-trained language models (LMs) such as BERT [5] have shown powerful achievements at downstream tasks, including MRC, in the past year, their pre-training methods ignore the role of factual knowledge. Existing work injects knowledge into LMs by auxiliary knowledge-driven objectives and updating parameters in a multi-task learning manner [24, 48], requiring pre-calculating knowledge representation and even pre-training from scratch. Another solution is to leverage the language model as an encoder, whose outputs are fed into the knowledge-text interaction layer for specific downstream tasks [41], increasing model complexity and computational cost.

To alleviate these problems, we take BERT as a base pre-trained model and incorporate the off-the-shelf commonsense representations for multi-choice MRC. Intuitively, it is easier to get the correct answer by fusing the commonsense relationships between the passage and options into the model for inference. Instead of stacking interaction layers downstream, we introduce three simple yet effective methods plugged in BERT structure, respectively named additive feature-based gating, multi-level linear transformation, and multi-head attentional fusion, to integrate token-level knowledge representations into BERT. Thus, text can be encoded in BERT while considering commonsense information. Different from previous work training the knowledge embedding before/after retrieving relevant entities, we directly leverage pre-computed ConceptNet embeddings [28] as external knowledge representation. Moreover, since not all commonsense concepts are necessary to the token and much external knowledge implicitly exists in conversations, a mask mechanism is introduced for token-level multi-hop relationship searching. Our goal is to enable the self-attention (SA) in BERT to identify the knowledge-aware tokens without additional knowledge-driven objectives or pre-training from scratch.

The remainder of this paper is organized as follows: Section 2 summarizes the main contributions. Section 3 describes the task and related notations, followed by a concise introduction to the baseline BERT. In Section 4, we propose our incremental language

Dialog :

M: Right. Where was it stolen?

W: In the city center, outside *McDonalds*, on Hope Avenue.

Question : Where was the woman's camera stolen?

A: Outside an ice cream place.

B: Outside a *restaurant*. ★

C: Outside her home.

Fig. 1 An example of DREAM dataset. (★: the correct answer)

models with three variants of injection methods. In Section 5, we present our token-level multi-hop relationship filtering mechanism. Section 6 shows the experimental details and results. Section 7 gives further analysis to verify the effectiveness of our methods. Section 8 introduces related work. Section 9 concludes.

2 Contributions

The main contributions of this paper can be summarized as follows:

1. We have proposed three simple yet effective injection methods plugged in BERT to incorporate off-the-shelf commonsense representations for multi-choice MRC;
2. We have introduced a token-level multi-hop mask mechanism to adaptively select relevant external knowledge, emphasizing the knowledge-aware tokens through the self-attention (SA) scores;
3. We have evaluated the incremental BERT on three prevalent multi-choice datasets, DREAM, CosmosQA and RACE. DREAM and CosmosQA contain a higher proportion of commonsense questions while RACE has few commonsense questions. The incremental BERT has obtained considerable improvements on two knowledge-driven datasets and comparable results on DREAM compared with the vanilla system. Further experimental analysis shows the robustness of the incremental model in the case of an incomplete training set.

3 Background

3.1 Task description

Given a passage $C = \{c_1, c_2, \dots, c_s\}$, a question $Q = \{q_1, q_2, \dots, q_m\}$ about this passage, and the answer options $A = \{A_1, A_2, \dots, A_k\}$, the target of multi-choice MRC is to choose the correct one from the candidate answer set A .

3.2 Baseline

BERT is based on Transformer backbone framework. In this paper, we directly use BERT as a baseline, which includes a multi-layer bidirectional Transformer encoder and a linear classifier. Following [23] we concatenate the context C , question Q , and answer option A_i as the input sequence:

$$[\text{CLS}]c_{1..s}[\text{SEP}]q_{1..m}[\text{SEP}]a_{1..n}^i[\text{SEP}]$$

where [SEP] is the separating token, and [CLS] is the token for classification. For each token, the input representation is constructed as:

$$BE_i = e_i^{tok} + e_i^{pos} + e_i^{seg}, i = 1..T$$

where e_i^{tok} , e_i^{pos} , e_i^{seg} , and T are the token embeddings, position embeddings, segment embeddings, and maximum length of sequence respectively. Tokens in C share a same segment embedding p^{seg} , and tokens in Q and A_i a same segment embedding qa^{seg} .

Such input representations are then fed into a stack of Transformer encoder blocks, which contains two sub-layers. The first sub-layer is a multi-head self-attention MHA. Given

a matrix of T query vectors $\mathbf{Q} \in \mathbb{R}^{T \times d_1}$, keys $\mathbf{K} \in \mathbb{R}^{T \times d_1}$ and values $\mathbf{V} \in \mathbb{R}^{T \times d_1}$, MHA($\mathbf{Q}, \mathbf{K}, \mathbf{V}$) is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_1}}\right)\mathbf{V} \tag{1}$$

$$b_j = \text{Attention}(\mathbf{Q}\mathbf{W}_j^Q, \mathbf{K}\mathbf{W}_j^K, \mathbf{V}\mathbf{W}_j^V) \tag{2}$$

$$\mathbf{B} = \text{Concat}(b_1, \dots, b_H) \tag{3}$$

where d_1 is the number of the hidden units, H denotes the number of heads used to focus on different parts of channels of the value vectors, $\mathbf{W}_j^Q \in \mathbb{R}^{T \times d_1/H}$, $\mathbf{W}_j^K \in \mathbb{R}^{T \times d_1/H}$ and $\mathbf{W}_j^V \in \mathbb{R}^{T \times d_1/H}$ are the parameters of linear mapping layer for j -th head. The second sub-layer is a position-wise fully connected feed-forward network (FFN), which consists of two dense linear layers with a GELU activation in between.

$$\mathbf{u}^l = \text{MHA}(\mathbf{h}^l, \mathbf{h}^l, \mathbf{h}^l) \tag{4}$$

$$\mathbf{h}^{l+1} = \text{FFN}(\mathbf{u}^l) \tag{5}$$

$$\text{FFN}(\mathbf{x}) = \mathbf{W}_2 \text{GELU}(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2 \tag{6}$$

where $\mathbf{h}^l \in \mathbb{R}^{T \times d_1}$ denotes the hidden state at the l -th layer. We utilize the input representations \mathbf{BE} as the initial state \mathbf{h}^0 . Note that we omit residual connection and layer normalization used in each sub-layer for simplicity, and refer readers to [31] and [5] for more details.

The final hidden state of the token [CLS], $\mathbf{h}_{[CLS]}^L$, is then projected into a score $p_i \in \mathbb{R}^1$ via a linear layer. For each question, we obtain the logit vector $\mathbf{p} = [p_1, p_2, \dots, p_k]$ for all options. We choose the option with highest score p as the answer.

4 Incremental BERT with commonsense

4.1 Knowledge integration mechanism

There have been many studies proving that large-scale pre-training language models based on Transformer, such as BERT, have a promising ability to represent text. However, they ignore the effective integration of external commonsense and consensus, which plays an important role in conversation comprehension. To this end, we explore three token-level injection methods to extend BERT to allow flexibility in incorporating external knowledge. Specifically, we integrate the commonsense embeddings \mathbf{CE} selected with a multi-hop co-occurrence mask (We will describe the knowledge representations and selection in Section 5) into BERT in three ways: additive feature-based gating, multi-level linear transformation, and multi-head attentional fusion. We denote the three methods as “gate”, “linear”, and “attention”, respectively.

Additive Feature-based Gating As depicted in the upper left part of Fig. 2, the method “gate” tries to add the ConceptNet representation of the selected commonsense associated token to the corresponding hidden state at each layer. To be specific, for each token t_i , we integrate the input representations \mathbf{BE}_i with external knowledge embeddings $\mathbf{CE}_i \in \mathbb{R}^{d_2}$ as:

$$\mathbf{In}_i = \mathbf{BE}_i + \sigma(\mathbf{W}_g \mathbf{CE}_i + \mathbf{b}_g) \tag{7}$$

where σ denotes the sigmoid activation function served as a gate mechanism and $\mathbf{W}_g \in \mathbb{R}^{d_1 \times d_2}$ is a trainable weight parameter. This gating mechanism generates a mask-vector

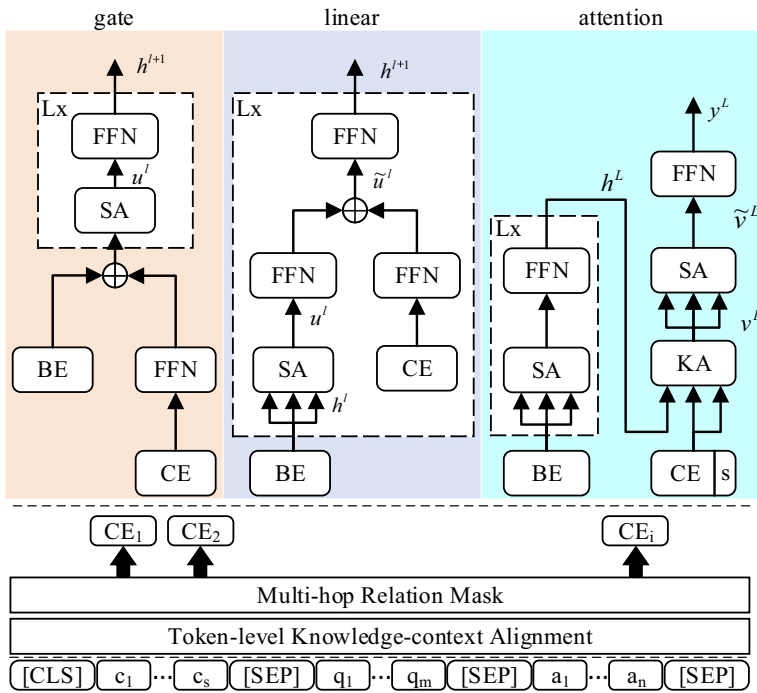


Fig. 2 Overview of the incremental language model. Three proposed fusion methods are abbreviated as “gate”, “linear”, and “attention”, respectively

from each CE_i with values between 0 and 1, incorporating information into salient dimensions of BE_i .

Multi-level Linear Transformation The middle part of Fig. 2 shows the second method “linear” that integrates the external knowledge at each intermediate FFN layer of BERT. For each Transformer encoder block, we replace the second sub-layer with a knowledge fusion layer for the incorporation of the token representations and their corresponding commonsense embeddings, which is computed as:

$$\tilde{u}_i^l = \text{GELU}(W_1^l u_i^l + \tilde{W}_1^l CE_i + b^l) \tag{8}$$

$$h_i^{l+1} = W_2 \tilde{u}_i^l + b_2 \tag{9}$$

where $\tilde{W}_1^l \in \mathbb{R}^{d_1 \times d_2}$ is a trainable weight parameter. Note that this method is in a similar spirit to the work of [48]. However, since our method focuses on the role of commonsense invariance between related tokens in text-based comprehension and their approach focuses on knowledge-driven tasks, we did not apply multi-head self-attention and mutual projection to knowledge embedding encoding. Instead, the knowledge embeddings are fixed for multi-level Transformer encoder blocks, which is simpler and does not require pre-training objective.

Multi-head Attentional Fusion The third method, as depicted in the “attention” part of Fig. 2, is inspired by the work of [18] and applies attention-based integration to the final

hidden states \mathbf{h}^L . Specifically, we add another Transformer encoder block with two multi-head attention sub-layers to the output of the BERT encoder. The first sub-layer is a multi-head knowledge attention (KA) computed as:

$$\mathbf{v}^L = \text{MHA}(\mathbf{h}^L, \tilde{\mathbf{C}}\mathbf{E}, \tilde{\mathbf{C}}\mathbf{E}) \quad (10)$$

where $\tilde{\mathbf{C}}\mathbf{E}$ is a concatenation of $\mathbf{C}\mathbf{E}$ and a knowledge sentinel $\mathbf{s} \in \mathbb{R}^{d_2}$. Considering not all tokens are relevant to the background knowledge, we follow [42] to employ the sentinel vector to control the tradeoff between background knowledge and information from the passage text. Thus, we get the knowledge-aware context representations \mathbf{v}^L and feed them into the second sub-layer, which consists of a multi-head self-attention and a FFN:

$$\tilde{\mathbf{v}}^L = \text{MHA}(\mathbf{v}^L, \mathbf{v}^L, \mathbf{v}^L) \quad (11)$$

$$\mathbf{y}^L = \text{FFN}(\tilde{\mathbf{v}}^L) \quad (12)$$

Note that we also employ residual connection and layer normalization around each attention layer. We replace \mathbf{h}^L with \mathbf{y}^L to predict the correct answer.

5 Commonsense representation and filtering

Existing commonsense libraries are usually presented in structured data. Taking into account the diversity of commonsense and the ready-made vector representation acquisition, we use ConceptNet 5.5,¹ a knowledge graph (KG) including linguistic and world knowledge from many different sources such as WordNet [21] and DBPedia. Commonsense in ConceptNet is represented in the form of a triple (*subject, relation, object*). For example, “a dog has a tail” can be represented as (*dog, HasA, tail*). Additionally, daily lexical knowledge and even emojis can be found in ConceptNet (e.g., (*lol, DerivedFrom, laugh*)). We believe that the graph-structured knowledge can be useful for multi-choice MRC that involves further reasoning with commonsense. Below we first introduce commonsense knowledge representations, and then present a token-level multi-hop knowledge filtering method.

5.1 Knowledge graph embedding

Unlike previous work training the knowledge embedding before/after retrieving relevant entities, we directly leverage off-the-shelf ConceptNet embeddings as external knowledge representation, representing global commonsense relationships. To be specific, we retrieve the tokens from the common vocabulary of BERT and ConceptNet and extract the corresponding KG embeddings. For those BERT tokens that are not found in ConceptNet, we set them to 0. We use three types of representation for common tokens: ConceptNet-PPMI², ConceptNet Numberbatch,³ and Randomly Initialized Embedding.

ConceptNet-PPMI A matrix of word embeddings trained on a sparse, symmetric term-term matrix where each cell contains the sum of the weights of all edges that connect the two corresponding terms. For each term in the ConceptNet graph, its ConceptNet-PPMI

¹<https://github.com/commonsense/conceptnet5/wiki>

²<https://conceptnet.s3.amazonaws.com/precomputed-data/2016/numberbatch/16.09/conceptnet-55-ppmi.h5>

³<https://github.com/commonsense/conceptnet-numberbatch>

representation reflects the context containing the information of other nodes to which it is connected.

ConceptNet Numberbatch A set of semantic vectors built with an ensemble that combines data from ConceptNet, word2vec, GloVe, and OpenSubtitles 2016, using a variation on retrofitting. Word embeddings in ConceptNet Numberbatch can represent both text-based context and structured knowledge.

Randomly Initialized Embedding Since the relations are not scored and represented explicitly, we also use randomly initialized embeddings for tokens to analyze the indirect commonsense relation between words in the passage and the effect of KG embeddings.

5.2 Token-level multi-hop knowledge filtering

Although vectors calculated based on the knowledge graph can represent the commonsense relationships, fusing these embeddings into all tokens of the question-oriented passage is usually invalid or even noisy. Moreover, the model requires commonsense relation not directly stated in the context to reach the correct option. For example, Fig. 3 shows that the model possibly needs multi-hop commonsense to reason about where the conversation takes place. Therefore, to improve the precision of useful information, we design a mask vector M to filter commonsense representations. Specifically, the length of M is the same as the sequence length input to the model and we initialize the mask values of all tokens to 1. For each token $t_1 \in A_i$ that is not a stop word or a padding token, we set $M_{index(t_1)} = 0$ and use it as a subject concept to search for the object concept $t_2 \in C \cup Q$ connected to t_1 in ConceptNet, then set $M_{index(t_2)} = 0$ and continue searching $t_3 \in C$ with t_2 . For concepts consisting of multiple tokens (e.g., *sign_contract*), we mask subtokens in the passage and repeat the above operation. We present this overall procedure in Fig. 4.

Thus, we obtain the mask vector M , which only contains 0 and 1 binary values. We further define the mask operation as follow:

$$\Phi_{mask}(CE_i) = \begin{cases} CE_i, & M_i = 0 \\ 0, & M_i = 1 \end{cases} \quad (13)$$

Dialog:

W: Good morning, can I help you?
 M: Yes, please. I'd like to **cash** two traveler's cheques.
 W: Could you **sign** your name here please?
 M: Sure.
 W: Thank you. How would you like your **money**?
 M: In hundreds and fifties, please.
 W: Ok. It's 1,660 yuan, here you are.
 M: Thanks. May I know the **exchange** rate?
 W: Well, at the moment the **exchange** rate between US **dollars** and RMB is 1:8.3. You give me two \$100 cheques; here is 1,660 yuan. Is that right?
 M: Yes, thanks
Question : Where is the conversation most probably taking place?
 A: In a supermarket.
 B: In a **bank**. ★
 C: In an office.

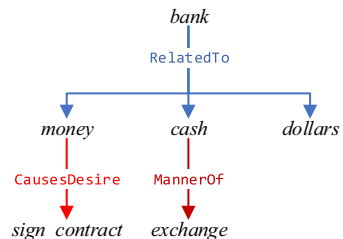


Fig. 3 An example of multi-hop relation searching. In ConceptNet, “bank” is connected to “money”, “cash” and “dollars” through the RelatedTo relationship. Further, “sign contract” and “exchange” can be found

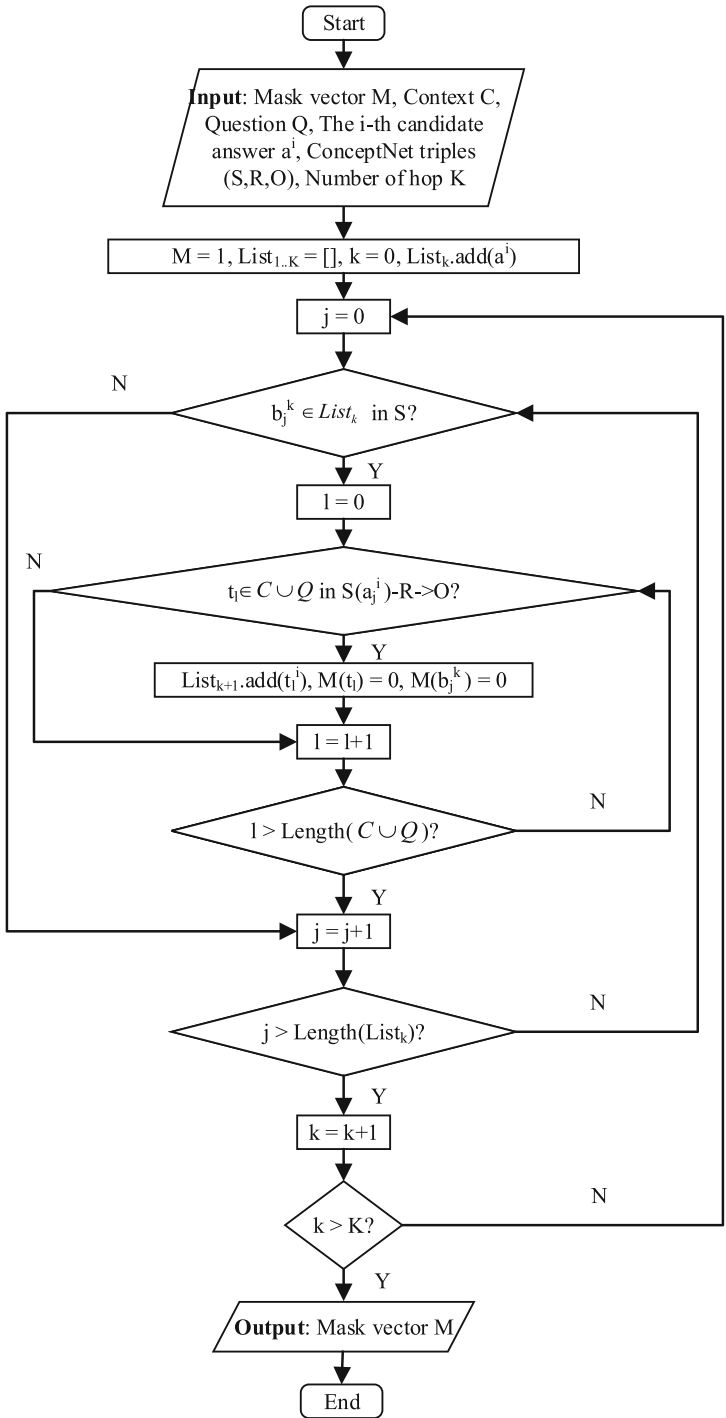


Fig. 4 Procedure of the token-level multi-hop knowledge filtering mechanism

For tokens corresponding to multiple concepts in multi-hop alignment, we use a single-layer feedforward network for weighted integration:

$$CE_i = \sum_{k=1}^K \alpha_k * c_{i,k} \quad (14)$$

$$\alpha_k = \frac{e^{w c_{i,k}}}{\sum_{k=1}^K e^{w c_{i,k}}} \quad (15)$$

where $w \in \mathbb{R}^{d_2}$ is a trainable weight parameter and K is the number of concepts containing the token in multi-hop alignment.

The filtered commonsense embeddings CE will be taken as input to the three fusion methods, depicted in Fig. 2. It is obvious that the commonsense filtering mechanism essentially improves the prediction of commonsense questions by integrating effective representations to change the token-level attention weights within the language model.

6 Experiments

6.1 Dataset and evaluation metric

We report results on three well-known multi-choice datasets, CosmosQA [13], DREAM [29], and RACE [17], which are summarized in Table 1. Specifically, we introduce the datasets:

CosmosQA is a large-scale dataset that requires commonsense-based reading comprehension, formulated as multiple-choice questions. In contrast to most existing MRC datasets where the questions focus on a factual and literal understanding of the context paragraph, CosmosQA focuses on reading between the lines over a diverse collection of people’s everyday narratives.

DREAM is collected from text material of listening comprehension examinations designed for evaluating the dialog understanding level of Chinese learners of English. DREAM contains 34% questions with unspoken commonsense, which requires the model to answer these questions not only by advanced reading skills but also with rich background knowledge.

RACE consists of two subsets: RACE-M and RACE-H respectively corresponding to the English exams for middle and high school Chinese students, which is recognized as one of the largest and most difficult datasets in multi-choice reading comprehension.

For all datasets, we use the official train/dev/test splits. For multi-choice MRC task, the evaluation metric is accuracy calculated as $acc = N^+/N$, where N^+ denotes the number of examples the model selects the correct answer, and N denotes the total number of evaluation examples.

Table 1 Statistics of multi-choice machine reading comprehension datasets. * denotes the numbers are based on 500 samples

	CosmosQA	DREAM	RACE
# paragraphs	21,866	6,444	27,933
# questions	35,588	10,197	97,687
# options	4	3	4
Ave. # paragraph	70.3	85.9	321.9
Need commonsense (%)	93.8	33.7	11.0*

Table 2 The best hyperparameters on different datasets (BERT-base/BERT-large). T denotes the max sequence length

Dataset	lr	epoch	Batch size	T
CosmosQA	$2e^{-5}/2e^{-5}$	10/8	32/32	256
DREAM	$2e^{-5}/2e^{-5}$	8/8	24/12	512
RACE	$3e^{-5}/2e^{-5}$	3/3	16/8	512

6.2 Implementation details

We implement our experiments using Huggingface⁴. We use BERT-base and BERT-large as baseline systems. To keep the order of magnitude close, we use L2 normalization to preprocess ConceptNet-PPMI. We experiment with commonsense relation searching of up to three hops. We set $K = 3$. The embeddings of commonsense are fixed during the fine-tuning process, and the parameters of BERT are trainable and initialized from the Huggingface checkpoint. For all fine-tuning experiments, we use BertAdam as the optimizer. We employ early stopping and predict the test set using the best model on the development set.

For training, we run all experiments on two 16G Quadro P5000. For CosmosQA, we set the max sequence length T to be 256 and select the hyperparameters from batch size: {16, 32, 64}, learning rate: {5e-5, 2e-5, 1e-5, 8e-6}. It takes about 8 hours to get the best result. For DREAM dataset, we run experiments for 8 epochs, set the max sequence length to be 512, and select the hyperparameters from batch size: {8, 12, 24, 36}, learning rate: {2e-5, 1e-5, 8e-6}. It takes about 4 hours to get the best result. For RACE dataset, we run experiments for 3 epochs, set the max sequence length to be 512, and select the hyperparameters from batch size: {8, 16, 32}, learning rate: {3e-5, 2e-5, 1e-5}. It takes about 12 hours to get the best result. In Table 2, we present the best hyperparameters on the development set and use them to verify on the test set.

6.3 Results

We compare the performance of the three proposed fusion methods with the two baselines in Table 3, where models on the leaderboards and publications are also shown.

- (1) **BERT+WAE:** To mimic the human exclusion strategy, authors train their model with the wrong answer loss and correct answer loss to generalize the features of their model, and exclude likely but wrong options.
- (2) **MMM:** It involves two sequential stages: coarse-tuning stage using out-of-domain datasets and multitask learning stage using a larger in-domain dataset to help model generalize better with limited data. Furthermore, the authors propose a novel multi-step attention network (MAN) as the top-level classifier for this task.
- (3) **DUMA:** It proposes a novel going-back-to-the-basic solution that straightforwardly models the MRC relationship as attention mechanism inside the network.
- (4) **DCMN:** It proposes a dual co-matching network (DCMN) which models the relationship among passage, question and answer options bidirectionally. Besides, it integrates two reading strategies including passage sentence selection and answer option interaction.

⁴<https://github.com/huggingface/transformers>

Table 3 Accuracy (%) on the multi-choice datasets including CosmosQA, DREAM and RACE. ConceptNet Numberbatch is used as commonsense representation and two-hop relation searching is applied. “-B” means the base model and “-L” means the large model. Due to the submission limit of CosmosQA, we only evaluate the incremental BERT-large model and publish the best result

Model	CosmosQA	DREAM	RACE
<i>Leaderboard</i>			
BERT-B	62.9	63.2	65.0
BERT-L	–	66.8	72.0
BERT-B+WAE	–	64.7	–
BERT-L+WAE	–	69.0	–
<i>Publication</i>			
MMM-B [14]	–	72.2	68.0
MMM-L [14]	–	76.0	72.5
DUMA-B [50]	–	62.3	–
DCMN-B [47]	–	–	67.0
DCMN+L [47]	–	–	75.8
Multiway-L [13]	68.4	–	–
<i>Ours (Concept Numb.+2hop)</i>			
BERT-B	–	62.8	65.0
BERT-B _{gate}	–	64.8	64.9
BERT-B _{linear}	–	65.3	65.3
BERT-B _{attention}	–	63.5	64.5
BERT-L	66.8	66.6	72.1
BERT-L _{gate}	67.9	67.8	72.4
BERT-L _{linear}	69.2	69.3	72.6
BERT-L _{attention}	67.6	67.3	72.0

- (5) **Multiway:** It performs multiway attention over BERT encoding output. Specifically, for the passage, question and option, the mutual attention will be calculated separately and pooled into the final representation.

ConceptNet Numberbatch is used as commonsense representation (We will discuss the role of knowledge embedding in Section 7), and we apply a two-hop commonsense relationship to filter knowledge.

From the results, we observe that our plug-in methods of incorporating commonsense can improve performance over the vanilla BERT on DREAM and CosmosQA. Specifically, multi-level linear transformation achieves the best results on CosmosQA (69.2% vs. 66.8% with BERT-large) and DREAM (65.3% vs. 62.8% with BERT-base and 69.3% vs. 66.6% with BERT-large). Compared with the other two methods, multi-head attentional fusion improves less on CosmosQA and DREAM, and decreases performance on RACE. In knowledge-driven multi-choice tasks, the incremental model variants obtain 0.7%–2.7% considerable improvement in average accuracy over the baseline of directly fine-tuned BERT. In contrast, our increment models have achieved comparable results on RACE. On the one hand, it means RACE requires little external knowledge for reading comprehension. On the other hand, it illustrates our methods do not lose the textual information after heterogeneous knowledge fusion. Compared to these public models, although the performance is

slightly worse on DREAM and RACE, the proposed methods have two advantages: 1) Different from DUMA and DCMN+, which are designed to be complex interactive matching networks, only a few mapping parameters and a single layer of parallel attention calculation are added to fuse commonsense into BERT; 2) Different from MMM using data from out-of-domain tasks for transfer learning, the incremental BERT has significantly improved the performance by direct fine-tuning. In addition, prediction results involving commonsense questions are difficult to explain in the existing methods clearly. On the contrary, we directly incorporate off-the-shelf commonsense representations into BERT's internal structure through token-level pre-matching to achieve the purpose of explicit use of external knowledge, obtaining interpretable performance improvement.

7 Discussion

7.1 Knowledge embedding

Table 4 shows the results of our incremental BERT-base_{linear} model obtained by adding initialization with different commonsense representations. From this table, we see that adding Concept-PPMI globally has a negative impact on the performance of BERT, while fusing it according to multi-hop commonsense relation improves the results. A possible reason is that Concept-PPMI only contains structured information based on the knowledge graph, providing a lot of noise when integrated indiscriminately. Hence, leveraging the multi-hop commonsense filtering algorithm helps BERT effectively utilize the structured information, which is also demonstrated in the experiment with random initialization. Moreover, the incremental model using random initialization commonsense performs better than using Concept-PPMI in global fusion, which means heterogeneous information is difficult to integrate directly without prior filtering since the pre-training procedure for language representation is quite different from the knowledge representation procedure.

7.2 Multi-hop commonsense selection

Table 5 illustrates the role of filtering commonsense, where we also integrate commonsense representations for each token in C and A_i for multi-hop analysis (*global* in Table 5). We can see that: (1) All three methods achieve their own best results in the two-hop commonsense relation search, which means that the indirect commonsense concept does not always work; (2) Multi-head attentional fusion performs better only in no more than two-hop commonsense relation, which is probably due to the knowledge-context attention mechanism is not sensitive to excessive noise fusion. Interestingly, additive feature-based gating with global commonsense performs better than itself with one-hop commonsense on DREAM

Table 4 Performance in accuracy (%) with different knowledge representation. We use BERT-base_{linear} and DREAM development set for analysis

KG Embeddings	Global	One-hop	Two-hop	Three-hop
Random	62.9	63.5	63.6	63.0
Concept-PPMI	62.3	63.9	64.2	63.9
Concept Numb.	64.4	64.7	65.1	64.2

The best performance of each variant is illustrated in bold

Table 5 Accuracy (%) on the CosmosQA, DREAM and RACE development dataset based on the different number of hop commonsense relation searching, where “global” means commonsense representations are integrated into all tokens

Model	CosmosQA				DREAM				RACE			
	global	1h	2h	3h	global	1h	2h	3h	global	1h	2h	3h
BERT-B _{gate}	63.7	64.1	64.4	64.3	64.5	63.9	64.9	63.4	64.5	64.5	64.9	64.6
BERT-B _{linear}	64.9	64.6	65.3	64.8	64.4	64.7	65.1	64.2	64.8	65.1	65.4	64.6
BERT-B _{attention}	62.9	63.4	63.8	63.3	62.6	63.2	63.5	62.2	64.0	64.4	64.4	64.1

Bold entries demonstrate that indirect commonsense concept does not always work

and CosmosQA. We hypothesize that the ConceptNet Numberbatch contains text-based lexicon information since it is obtained by jointly retrofitting from word2vec and GloVe.

7.3 Self-attention

To verify our goal to enable the self-attention in BERT to identify the knowledge-aware tokens, we consider the case depicted in Fig. 3. In this case, the BERT chooses the wrong candidate option (A) and our models make the right choice (B). We capture the correlation between tokens in the BERT and two-hop BERT-base_{linear} respectively, which are visualized in Fig. 5a and b, obtained from the penultimate self-attention layer of BERT and two-hop BERT-base_{linear}, respectively.⁵ For BERT, the token “bank” has a low degree of similarity to all tokens $t_i \in C$ except “traveler” and “cheques”, and the focus of almost all tokens in the dialog is quite discrete. Moreover, part of tokens has a relatively high degree of similarity to “conversation” and the segment token, which is not enough to support the model to choose the correct conversation place. By contrast, our incremental model can learn more accurate representations to understand the commonsense relation between the passage and the candidate option, and infer the correct answer. From Fig. 5b, we can observe that “bank” has a high degree of relevance with “cash”, “sign”, “money”, “exchange” and “dollars”, which perfectly reflects their commonsense relationships shown in Fig. 3. In addition, the original similarity between “bank” and “cheques” is also retained or even strengthened. It illustrates that the commonsense fusion method preserves textual information while effectively utilizing heterogeneous knowledge.

7.4 Incomplete training set

BERT pre-trained on large-scale texts is still deficient in explicitly representing the relationship between commonsense concepts. The smaller the text training set in the downstream knowledge-driven task, the higher the requirement for the commonsense understanding ability of the model. We show the results of different incomplete training set settings in Fig. 6, using BERT-base as the baseline. We can see that the performance of all models shows a similar trend with the decrease in training set size. Compared to the vanilla BERT, our incremental models maintain better robustness. It is worth mentioning that the performance of the three-hop models have decreased more slowly than the one-hop models when the training

⁵During visualization, we use a row-wise softmax operation to normalize similarity scores over all sequence tokens.

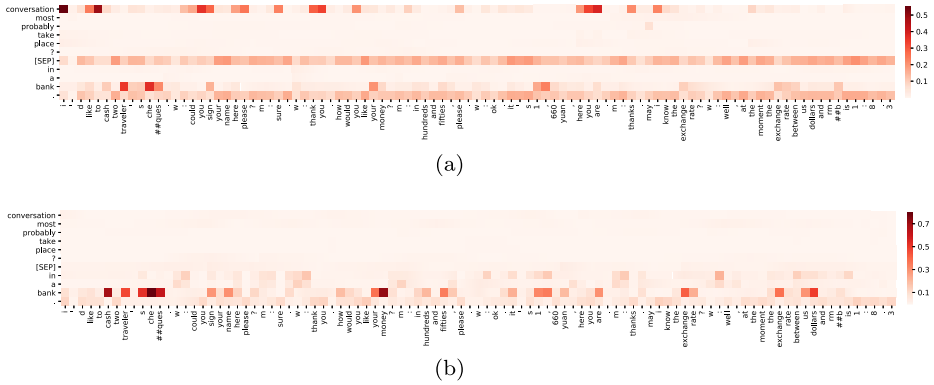


Fig. 5 Case study. In this case, the BERT (a) chooses the wrong candidate option and our models make the right choice. Two-hop BERT-base_{linear} (b) is used for comparison. Heat maps present similarities between correct answer (row) and dialog (column) tokens

set size drops to 60% and 40%. The three methods have different performances for different number of hops. We argue that commonsense would be more needed when the scale of text training set decreases to a certain extent. Augmenting BERT with external knowledge incorporation results in significant improvements in the settings with incomplete training set.

7.5 Computational costs

We present the computing resources used in our experiments. Each component’s parameters and the running time for each variant (1-hop/2-hop/3-hop) are summarized in Table 6. Since proposed methods add few parameters, each variant took the same time as the BERT-large baseline. The computation bottleneck is mainly from BERT and multi-hop token alignment. Considering the performance improvement of the two-hop relation search, the increase in overall running time is acceptable. However, the huge amount of parameters and long running time mean that there is still much work to deploy the model as a practical question answering system. Interestingly, we have found that the running time on RACE did not increase significantly as the number of search hops increased, which further reflects that RACE contains few commonsense questions.

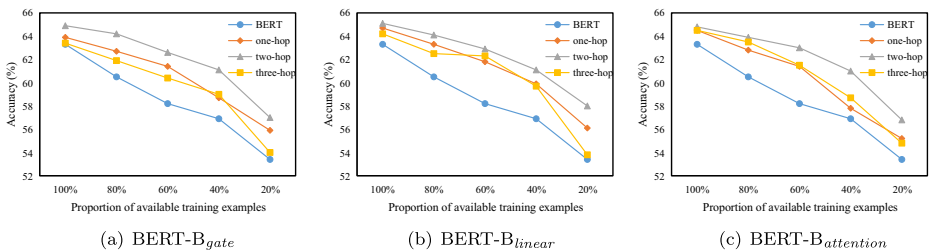


Fig. 6 Accuracy on DREAM development set with the decrease in training set size. BERT-base is used for comparison

Table 6 Computational costs for each variant of proposed methods. BERT-large is taken as the baseline. Running time is the sum of relation search time (1-hop/2-hop/3-hop) and model training time

Components	#Parameters (M)	Running Time (h)		
		DREAM	CosmosQA	RACE
BERT-large	340	6.5	8	12
<i>gate</i>	0.31	6.8/7.4/9	8.7/10/12.6	13.8/14.5/14.9
<i>linear</i>	3.7			
<i>attention</i>	3.15			

7.6 Error analysis

We conduct the following error analysis to investigate problems that our model is short of the ability to address. We randomly extract 200 samples from the development set of DREAM, and then classify them into several question types according to the annotation criterion consistent with [29]. We compare two-hop BERT-base_{linear} with BERT-base on these categories, as shown in Table 7. Both models perform worse than random guessing (33.3%) on math problems since the Conceptnet does not contain the commonsense of mathematical computing, especially time and currency, which can be future work. Although superior to BERT on the implicit questions (e.g., under the categories *logic* and *commonsense*) which require external knowledge, our incremental model is less capable of answering these questions under the category *summary*. We hypothesize that integrating token-level commonsense may interfere with the reasoning requiring the aggregation of information from multiple sentences.

8 Related work

Machine Reading Comprehension In recent years, many MRC datasets have been released to solve different task scenarios, e.g., cloze-style [8, 9], extractive/abstractive answer [6, 15, 16, 22, 25], multi-choice [17], conversational QA [3, 26], multi-hop [38, 43], and whether external knowledge is needed [4, 13, 19, 30, 46]. Most MRC datasets that require external knowledge such as ARC, DREAM, OpenBookQA, CommonsenseQA and CosmosQA are designed in a multi-choice form. In this paper, we focus on the multi-choice MRC task. Hence, we choose CosmosQA, DREAM and RACE in the experiments.

Table 7 Error analysis on DREAM. The column of “Proportion” reports the percentage of question types among 200 samples that are from the development set of DREAM dataset

Question type	BERT-B	BERT-B _{linear}	Proportion
Matching	65.1	65.4	12.2
Reasoning	62.9	64.9	87.8
Summary	78.1	77.7	8.6
Logic	59.3	62.1	76.1
Arithmetic	31.7	32.3	2.5
Commonsense	57.9	62.2	32.5

For multi-choice MRC, existing methods include designing the interaction among the passage, question and option [35, 47, 50], or transfer learning through data augmentation [14]. Nevertheless, these methods do not rely on commonsense knowledge for logical reasoning.

Integrating External Knowledge for MRC Existing work has utilized structured knowledge from KBs/KGs to improve performance on MRC and QA. Existing work has utilized structured knowledge from KBs/KGs to improve performance on MRC and QA. Yang et al. [42] incorporate retrieved knowledge into LSTM by employing an attention mechanism with a sentinel. Bauer et al. [1] select grounded multi-hop relational commonsense information from ConceptNet via pointwise mutual information and term-frequency based scoring function and use a selectively gated attention mechanism to fuse the knowledge. Mihaylov et al. [20] introduce a mixed attention to external knowledge for cloze-style reading comprehension. Chen et al. [2], Wang et al. [33] and Zhong et al. [49] explore the effect of semantic relations from KGs such as ConceptNet on MRC. Wang et al. [32] propose a data enrichment method, which uses WordNet to extract inter-word semantic connections as general knowledge from each given passage-question pair. Xiong et al. [40] retrieve the corresponding entities and relation from text to aggregate answer evidence from an incomplete KB. Yang et al. [41] take BERT as encoder and employ an attention mechanism similar to Yang et al. [42] to fuse globally pre-trained knowledge downstream. Compared to these methods, we mainly focus on plug-in fusion methods and explore token-level multi-hop commonsense representation integration instead of relation embeddings.

Injecting knowledge into LMs Neural networks and deep learning have been widely used in many fields such as computer vision and image processing [10–12, 44, 45]. Recently, pre-trained deep language models such as BERT have shown powerful achievements in downstream NLP tasks including MRC. The injection of external knowledge to LMs can be generally divided into two groups. Methods in the first group design auxiliary knowledge-driven objectives and updating parameters in a multi-task learning manner [24, 37, 39, 48], which requires pre-calculating knowledge representation and even pre-training BERT from scratch. The second group is to pre-train external modules to assist LMs [34, 49]. In contrast, our fusion methods are to directly fine-tune on the target MRC datasets.

9 Conclusion

This paper introduces increment BERT with three plug-in fusion methods, which enhances the vanilla BERT with commonsense representations from ConceptNet. We have used pre-computed ConceptNet embeddings as external knowledge representation and introduced a mask mechanism for token-level multi-hop relationship searching to filter external knowledge, so as to enable the self-attention in BERT to identify the knowledge-aware tokens effectively. Our variants of proposed methods have achieved significant improvements over baseline on two knowledge-driven multi-choice datasets. Experiments on few-commonsense dataset RACE shows that the introduction of external knowledge will not cause loss to the original text information understanding. Future work can start with more granular relationships to integrate external knowledge and how to design an effective yet efficient model architecture for practical deployment.

Acknowledgements We thank the funding 2020-KF-10 supported by Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce.

References

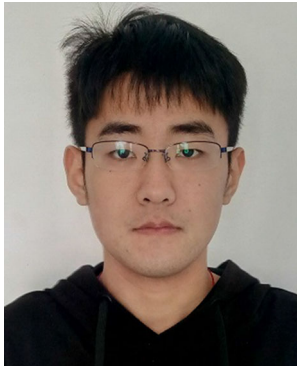
1. Bauer L, Wang Y, Bansal M (2018) Commonsense for generative multi-hop question answering tasks. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pp 4220–4230. <https://aclanthology.info/papers/D18-1454/d18-1454>
2. Chen Q, Zhu X, Ling Z, Inkpen D, Wei S (2018) Neural natural language inference models enhanced with external knowledge. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, Volume 1: Long Papers, pp 2406–2417. <https://doi.org/10.18653/v1/P18-1224>. <https://www.aclweb.org/anthology/P18-1224/>
3. Choi E, He H, Iyyer M, Yatskar M, Yih Wt, Choi Y, Liang P, Zettlemoyer L (2018) Quac: Question answering in context. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp 2174–2184. <http://aclweb.org/anthology/D18-1241>
4. Clark P, Cowhey I, Etzioni O, Khot T, Sabharwal A, Schoenick C, Tafjord O (2018) Think you have solved question answering? try arc, the AI2 reasoning challenge. CoRR arXiv:1803.05457
5. Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers), pp 4171–4186. <https://aclweb.org/anthology/papers/N/N19/N19-1423/>
6. Dhingra B, Mazaitis K, Cohen WW (2017) Quasar: Datasets for question answering by search and reading. CoRR arXiv:1707.03904
7. Ding M, Zhou C, Chen Q, Yang H, Tang J (2019) Cognitive graph for multi-hop reading comprehension at scale. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28– August 2, 2019, Volume 1: Long Papers, pp 2694–2703. <https://www.aclweb.org/anthology/P19-1259/>
8. Hermann KM, Kociský T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P (2015) Teaching machines to read and comprehend. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7–12, 2015, Montreal, Quebec, Canada, pp 1693–1701. <http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend>
9. Hill F, Bordes A, Chopra S, Weston J (2016) The goldilocks principle: Reading children’s books with explicit memory representations. In: 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings. arXiv:1511.02301
10. Hong C, Yu J, Tao D, Wang M (2014) Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval. *IEEE Trans Ind Electron* 62(6):3742–3751
11. Hong C, Yu J, Wan J, Tao D, Wang M (2015) Multimodal deep autoencoder for human pose recovery. *IEEE Trans Image Process* 24(12):5659–5670. <https://doi.org/10.1109/TIP.2015.2487860>
12. Hong C, Yu J, Zhang J, Jin X, Lee K (2019) Multimodal face-pose estimation with multitask manifold deep learning. *IEEE Trans Ind Informatics* 15(7):3952–3961. <https://doi.org/10.1109/TII.2018.2884211>
13. Huang L, Bras RL, Bhagavatula C, Choi Y (2019) Cosmos QA: machine reading comprehension with contextual commonsense reasoning. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019, Association for Computational Linguistics, pp 2391–2401. <https://doi.org/10.18653/v1/D19-1243>
14. Jin D, Gao S, Kao J, Chung T, Hakkani-Tür D (2020) MMM: multi-stage multi-task learning for multi-choice reading comprehension. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020, AAAI Press, pp 8010–8017. <https://aaai.org/ojs/index.php/AAAI/article/view/6310>
15. Joshi M, Choi E, Weld DS, Zettlemoyer L (2017) Triviaqa: a large scale distantly supervised challenge dataset for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pp 1601–1611. <https://doi.org/10.18653/v1/P17-1147>
16. Kociský T, Schwarz J, Blunsom P, Dyer C, Hermann KM, Melis G, Grefenstette E (2018) The narrativeqa reading comprehension challenge. *TACL* 6:317–328. <https://transacl.org/ojs/index.php/tacl/article/view/1197>

17. Lai G, Xie Q, Liu H, Yang Y, Hovy EH (2017) RACE: large-scale reading comprehension dataset from examinations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017, pp 785–794. <https://aclanthology.info/papers/D17-1082/d17-1082>
18. Li Z, Niu C, Meng F, Feng Y, Li Q, Zhou J (2019) Incremental transformer with deliberation decoder for document grounded conversations. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28– August 2, 2019, Volume 1: Long Papers, pp 12–21. <https://www.aclweb.org/anthology/P19-1002/>
19. Mihaylov T, Clark P, Khot T, Sabharwal A (2018) Can a suit of armor conduct electricity? A new dataset for open book question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pp 2381–2391. <https://www.aclweb.org/anthology/D18-1260/>
20. Mihaylov T, Frank A (2018) Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, Volume 1: Long Papers, pp 821–832. <https://doi.org/10.18653/v1/P18-1076>. <https://www.aclweb.org/anthology/P18-1076/>
21. Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ (1990) Introduction to wordnet: An on-line lexical database. *Int J Lexicography* 3(4):235–244. <https://doi.org/10.1093/ijl/3.4.235>
22. Nguyen T, Rosenberg M, Song X, Gao J, Tiwary S, Majumder R, Deng L (2016) MS MARCO: A human generated machine reading comprehension dataset. In: Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016. http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf
23. Pan X, Sun K, Yu D, Chen J, Ji H, Cardie C, Yu D (2019) Improving question answering with external knowledge. In: Fisch A, Talmor A, Jia R, Seo M, Choi E, Chen D (eds) Proceedings of the 2nd Workshop on Machine Reading for Question Answering, MRQA@EMNLP 2019, Hong Kong, China, November 4, 2019, Association for Computational Linguistics, pp 27–37. <https://doi.org/10.18653/v1/D19-5804>
24. Peters ME, Neumann M, Iyyer R, Schwartz R, Joshi V, Singh S, Smith NA (2019) Knowledge enhanced contextual word representations. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019, Association for Computational Linguistics, pp 43–54. <https://doi.org/10.18653/v1/D19-1005>
25. Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) Squad: 100, 000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1–4, 2016, pp 2383–2392. <http://aclweb.org/anthology/D/D16/D16-1264.pdf>
26. Reddy S, Chen D, Manning CD (2019) Coqa: A conversational question answering challenge. *TACL* 7:249–266. <https://transacl.org/ojs/index.php/tacl/article/view/1572>
27. Seo MJ, Kembhavi A, Farhadi A, Hajishirzi H (2017) Bidirectional attention flow for machine comprehension. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings. <https://openreview.net/forum?id=HJ0UKP9ge>
28. Speer R, Chin J, Havasi C (2017) Conceptnet 5.5: An open multilingual graph of general knowledge. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4–9, 2017, San Francisco, California, USA., pp 4444–4451. <http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972>
29. Sun K, Yu D, Chen J, Yu D, Choi Y, Cardie C (2019) DREAM: A challenge dataset and models for dialogue-based reading comprehension. *TACL* 7:217–231. <https://transacl.org/ojs/index.php/tacl/article/view/1534>
30. Talmor A, Herzig J, Lourie N, Berant J (2019) Commonsenseqa: A question answering challenge targeting commonsense knowledge. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers), pp 4149–4158. <https://www.aclweb.org/anthology/N19-1421/>
31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA, pp 5998–6008. <http://papers.nips.cc/paper/7181-attention-is-all-you-need>

32. Wang C, Jiang H (2019) Explicit utilization of general knowledge in machine reading comprehension. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28– August 2, 2019, Volume 1: Long Papers, pp 2263–2272. <https://www.aclweb.org/anthology/P19-1219/>
33. Wang L, Sun M, Zhao W, Shen K, Liu J (2018) Yuanfudao at semeval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension. In: Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018, pp 758–762. <https://aclanthology.info/papers/S18-1120/s18-1120>
34. Wang R, Tang D, Duan N, Wei Z, Huang X, Ji J, Cao G, Jiang D, Zhou M (2020) K-adapter: Infusing knowledge into pre-trained models with adapters. CoRR arXiv:[abs/2002.01808](https://arxiv.org/abs/2002.01808)
35. Wang S, Yu M, Jiang J, Chang S (2018) A co-matching model for multi-choice reading comprehension. In: Gurevych I, Miyao Y (eds) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers, Association for Computational Linguistics, pp 746–751. <https://doi.org/10.18653/v1/P18-2118>. <https://www.aclweb.org/anthology/P18-2118/>
36. Wang W, Yang N, Wei F, Chang B, Zhou M (2017) Gated self-matching networks for reading comprehension and question answering. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pp 189–198. <https://doi.org/10.18653/v1/P17-1018>
37. Wang X, Gao T, Zhu Z, Liu Z, Li J, Tang J (2019) KEPLER: A unified model for knowledge embedding and pre-trained language representation. CoRR arXiv:[abs/1911.06136](https://arxiv.org/abs/1911.06136)
38. Welbl J, Stenetorp P, Riedel S (2018) Constructing datasets for multi-hop reading comprehension across documents. *TACL* 6:287–302. <https://transacl.org/ojs/index.php/tacl/article/view/1325>
39. Xiong W, Du J, Wang WY, Stoyanov V (2020) Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net. <https://openreview.net/forum?id=BJlzm64tDH>
40. Xiong W, Yu M, Chang S, Guo X, Wang WY (2019) Improving question answering over incomplete kbs with knowledge-aware reader. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp 4258–4264. <https://www.aclweb.org/anthology/P19-1417/>
41. Yang A, Wang Q, Liu J, Liu K, Lyu Y, Wu H, She Q, Li S (2019) Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers, pp 2346–2357. <https://www.aclweb.org/anthology/P19-1226/>
42. Yang B, Mitchell TM (2017) Leveraging knowledge bases in lstms for improving machine reading. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pp 1436–1446. <https://doi.org/10.18653/v1/P17-1132>
43. Yang Z, Qi P, Zhang S, Bengio Y, Cohen WW, Salakhutdinov R, Manning CD (2018) Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pp 2369–2380. <https://aclanthology.info/papers/D18-1259/d18-1259>
44. Yu J, Tan M, Zhang H, Tao D, Rui Y (2019) Hierarchical deep click feature prediction for fine-grained image recognition. *IEEE transactions on pattern analysis and machine intelligence*
45. Yu J, Tao D, Wang M, Rui Y (2015) Learning to rank using user clicks and visual features for image retrieval. *IEEE Trans Cybern* 45(4):767–779. <https://doi.org/10.1109/TCYB.2014.2336697>
46. Zhang S, Liu X, Liu J, Gao J, Duh K, Durme BV (2018) Record: Bridging the gap between human and machine commonsense reading comprehension. CoRR arXiv:[abs/1810.12885](https://arxiv.org/abs/1810.12885)
47. Zhang S, Zhao H, Wu Y, Zhang Z, Zhou X, Zhou X (2020) DCMN+: dual co-matching network for multi-choice reading comprehension. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, pp 9563–9570. <https://aaai.org/ojs/index.php/AAAI/article/view/6502>
48. Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q (2019) ERNIE: enhanced language representation with informative entities. In: Proceedings of the 57th Conference of the Association for Computational

- Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp 1441–1451. <https://www.aclweb.org/anthology/P19-1139/>
49. Zhong W, Tang D, Duan N, Zhou M, Wang J, Yin J (2019) Improving question answering by commonsense-based pre-training. In: Tang J, Kan M, Zhao D, Li S, Zan H (eds) Natural Language Processing and Chinese Computing - 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9-14, 2019, Proceedings, Part I, Lecture Notes in Computer Science, vol 11838. Springer, Berlin, pp 16–28. https://doi.org/10.1007/978-3-030-32233-5_2
50. Zhu P, Zhao H, Li X (2020) Dual multi-head co-attention for multi-choice reading comprehension. CoRR arXiv:2001.09415

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Ronghan Li received his M.E. in Computer Science from Northwestern Polytechnical University (NPU), Xi'an, China, in 2015. He is currently studying for Ph.D. at the School of Computer Science and Engineering, NPU. His research areas include neuro-linguistic calculation, natural language processing, and machine learning.



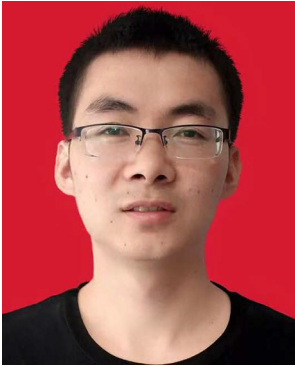
Lifang Wang is a professor at the School of Computer Science, at Northwestern Polytechnical University in Xi'an, China. She received her Ph.D. in Computer Science and Technology from Northwestern Polytechnical University. Her research interests include electronic commerce technology, cloud computing, cloud storage, machine learning, deep learning, and natural language processing.



Zejun Jiang is a professor at the School of Computer Science, at Northwestern Polytechnical University in Xi'an, China. He was born in Hefei, China. He received his B.S degrees and M.S degrees in Computer Science and Technology in 1985, 1988 respectively from Northwestern Polytechnical University. His research interests include deduplication in distributed systems, cloud storage, machine learning, dialog system, and natural language processing.



Dong Liu is currently studying for a master's degree at Northwestern Polytechnical University. His research interests are natural language processing and machine learning.



Meng Zhao a Ph.D. candidate at Northwestern Polytechnical University in Xi'an, China. He was born in Henan, China. He received his B.S degrees and M.S degrees in Computer Science and Technology in 2014, 2017 from Henan Normal University, Henan University of Technology respectively. His research interests include deep learning, question answering system.



Xinyu Lu received an M.S degree in Electronics and Communication Engineering in 2017 from Ningxia University. He is going on pursuing a Ph.D. in Computer Science and Technology from Northwestern Polytechnical University. His research interests include Natural Language Processing, Question Answering system, and Knowledge Representation Learning.

Affiliations

Ronghan Li¹ · Lifang Wang¹ · Zejun Jiang¹  · Dong Liu¹ · Meng Zhao¹ · Xinyu Lu¹

Ronghan Li
lrh000@mail.nwpu.edu.cn

Lifang Wang
wanglf@nwpu.edu.cn

Dong Liu
liudong2018@mail.nwpu.edu.cn

Meng Zhao
zmsmartboy@mail.nwpu.edu.cn

Xinyu Lu
luxy@mail.nwpu.edu.cn

¹ School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, 710072, People's Republic of China