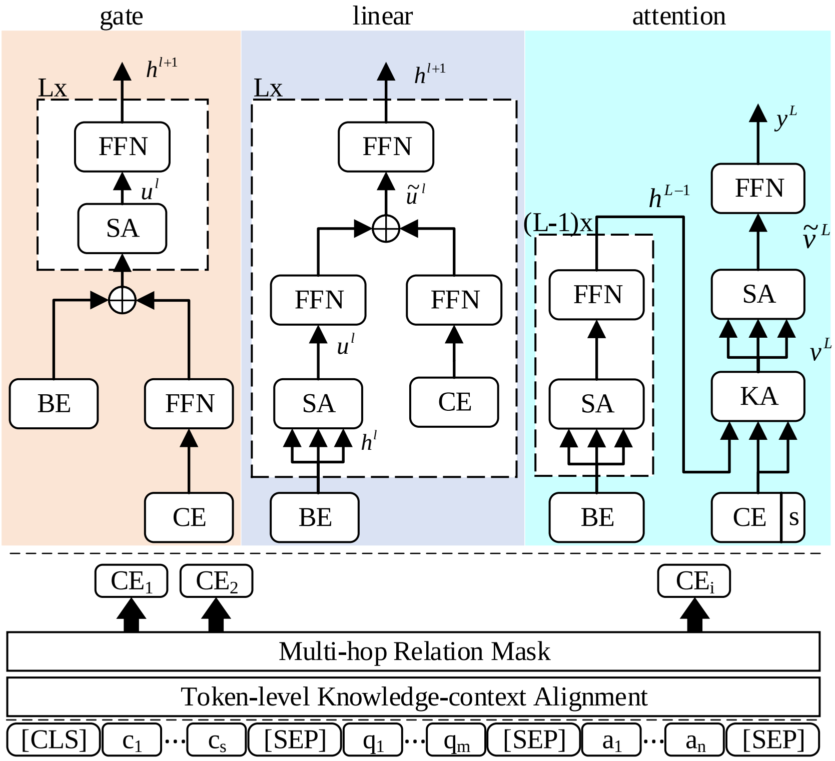
Graphical Abstract

Enhancing Transformer-based language models with Commonsense Representations for Knowledge-driven Machine Comprehension

Ronghan Li, Zejun Jiang, Lifang Wang, Xinyu Lu, Meng Zhao, Daqing Chen



Highlights

Enhancing Transformer-based language models with Commonsense Representations for Knowledge-driven Machine Comprehension

Ronghan Li, Zejun Jiang, Lifang Wang, Xinyu Lu, Meng Zhao, Daqing Chen

- To explicitly augment Transformer-based language models (TrLMs) for knowledge-driven MRC task with fewer model structure changes, we proposed three simple yet effective injection methods integrated into pre-trained TrLMs to incorporate off-the-shelf commonsense representations directly in the fine-tuning stage.

- A token-level multi-hop mask mechanism was introduced to filter irrelevant knowledge and enabled the self-attention (SA) in Transformer to identify the knowledge-aware tokens, which can improve the efficiency of knowledge fusion.

- The performance of the incremental TrLMs on two prevalent knowledge-driven datasets, DREAM and CosmosQA, has been evaluated showing a 1%-4.1% improvement compared with the vanilla TrLMs. Compared with existing methods, our model variants also achieve competitive results with the fewer computational cost. Further experimental analysis has demonstrated the effectiveness of proposed methods and the robustness of the incremental models in the case of an incomplete training set.

Enhancing Transformer-based language models with Commonsense Representations for Knowledge-driven Machine Comprehension

Ronghan Li[a], Zejun Jiang[a], Lifang Wang[a,b,*], Xinyu Lu[a], Meng Zhao[a], Daqing Chen[c]

[a] *School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, PR China, 710072*

[b] *Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce, Luoyang, PR China, 471000*

[c] *Division of Computer Science and Informatics, School of Engineering, London South Bank University, London, UK*

Abstract

Compared to the traditional machine reading comprehension (MRC) with limitation to the information in a passage, knowledge-driven MRC tasks aim to enable models to answer the question according to text and related commonsense knowledge. Although pre-trained Transformer-based language models (TrLMs) such as BERT and Roberta, have shown powerful performance in MRC, external knowledge such as unspoken commonsense and world knowledge still can not be used and explained explicitly. In this work, we present three simple yet effective injection methods integrated into the structure of TrLMs to fine-tune downstream knowledge-driven MRC tasks with off-the-shelf commonsense representations. Moreover, we introduce a mask mechanism for a token-level multi-hop relationship searching to filter external knowledge. Experimental results indicate that the incremental TrLMs have significantly outperformed the baseline systems by 1%-4.1% on DREAM and CosmosQA, two prevalent knowledge-driven datasets. Further analysis shows the effectiveness of the proposed methods and the robustness of the incremental model in the case of an incomplete training set.

Keywords: Machine Reading Comprehension, Transformer, Commonsense

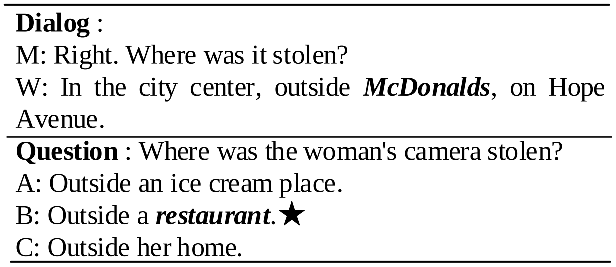
[*] Corresponding author

Email address: wanglf@nwpu.edu.cn (Lifang Wang)

Knowledge, Pretrained Language Model

1. Introduction

Machine Reading Comprehension (MRC) is a challenging task in natural language processing, which requires answering a question by comprehending the relevant passages. Thanks to the release of large-scale datasets [1, 2, 3, 4], related end-to-end neural networks have achieved promising results in various scenarios [5, 6, 7, 8, 9]. Traditional MRC is limited to question answering (QA) for given text content, ignoring the importance of external knowledge for text understanding. Recently knowledge-driven MRC has attracted increasing attention and raised new challenges. Several related datasets such as ARC [10], DREAM [11], OpenBookQA [12], CommonsenseQA [13] and CosmosQA [14] have been proposed, which are designed in a multi-choice form and require models to answer the question by combining related commonsense knowledge. Fig. 1 shows an example on DREAM, where the well-known fact that "*McDonald's*" is a restaurant is useful to find the correct option.



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. M: Man. W: Woman. (★: the correct answer)

On the other hand, although Transformer [15]-based language models (TrLMs) such as BERT [16] and Roberta [17] have shown powerful achievements with downstream tasks including MRC in the past year, their pre-training methods have usually ignored the role of factual knowledge. Existing work can be roughly divided into two groups: global fusion during pre-training and selective fusion in the fine-tuning stage. The first group injects knowledge into TrLMs by auxiliary knowledge-driven objectives and updating parameters in a multi-task learning manner [18, 19], which requires expensive further pre-training even from scratch. Besides, global fusion lacks interpretability for specific downstream tasks; that is, which commonsense is used cannot be explicitly evaluated. The second group leverages the language model as an encoder, whose outputs are fed into the complicated knowledge-text interaction layer for specific downstream tasks [20]. However, it has also inevitably increased model complexity and computational cost.

To alleviate these problems, we consider explicitly incorporating off-the-shelf commonsense representations into TrLMs' internal structure to enhance encoding during the fine-tuning phase for knowledge-driven MRC. Intuitively, it is easier to get the correct answer by fusing any existing commonsense relationships between a passage and options into the TrLMs for inference. Instead of stacking interaction layers downstream, we introduce three simple yet effective plug-in methods, named additive feature-based gating, multi-level linear transformation, and multi-head attentional fusion, respectively, to explicitly integrate token-level knowledge representations into TrLMs. Thus, text can be encoded in TrLMs while considering commonsense information. We directly leverage pre-computed ConceptNet embeddings [21] as external knowledge representation. Moreover, since not all commonsense concepts are necessary to the token and many commonsense relations are indirectly between a passage and candidate answer (e.g., *ConceptNet* `ISA` *knowledge graph* `HAS` *commonsense knowledge*), a mask mechanism is introduced for token-level multi-hop relationship searching. Our goal is to enable the self-attention (SA) in TrLMs to identify the knowledge-aware tokens without additional knowledge-driven objectives or pre-training from scratch.

We have conducted extensive experiments on two prevalent knowledge-driven datasets, DREAM and CosmosQA, to evaluate the proposed models. Our model variants have obtained a 1%-4.1% improvement on average accuracy with fewer computational resources than baseline systems. For a fair comparison, we have also evaluated our models on the RACE [3] dataset that is one of the largest datasets in multi-choice MRC with few commonsense questions. Experimental results show that incremental TrLMs do not lose the textual information after heterogeneous knowledge fusion. The further analysis illustrates that our fusion methods and mask mechanism effectively help TrLMs form commonsense-aware token representations and maintain the robustness of QA in the case of an incomplete training set.

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

(1) To explicitly augment Transformer-based language models (TrLMs) for knowledge-driven MRC task with fewer model structure changes, we proposed three simple yet effective injection methods integrated into pre-

trained TrLMs to incorporate off-the-shelf commonsense representations directly in the fine-tuning stage.

(2) A token-level multi-hop mask mechanism was introduced to filter irrelevant knowledge and enabled the self-attention (SA) in Transformer to identify the knowledge-aware tokens, which can improve the efficiency of knowledge fusion.

(3) The performance of the incremental TrLMs on two prevalent knowledge-driven datasets, DREAM and CosmosQA, has been evaluated showing a 1%-4.1% improvement compared with the vanilla TrLMs. Compared with existing methods, our model variants also achieve competitive results with the fewer computational cost. Further experimental analysis has demonstrated the effectiveness of proposed methods and the robustness of the incremental models in the case of an incomplete training set.

The remainder of this paper is organized as follows: Section 2 describes the task and the related notations, followed by a concise introduction to the Transformer-based LM for multi-choice MRC as the baseline system. In Section 3, we propose our incremental language models with three variants of injection methods. In Section 4, we present our token-level multi-hop relationship filtering mechanism. Section 5 shows the experimental details and the results. Section 6 gives further analysis to verify the effectiveness of our methods. Section 7 introduces related work. Section 8 concludes and looks forward to future work.

2. Background

2.1. Task Description

Given a question $Q = \{q_1, q_2, ..., q_m\}$ and optionally a supporting passage $C = \{c_1, c_2, ..., c_s\}$, a knowledge-driven MRC system is expected to select the correct answer from multiple candidate answers $A = \{a^1, a^2, ...a^k\}$ through the supporting passage and the related commonsense knowledge. A summary of key notations is presented in Table 1.

2.2. Baseline

Recently, Transformer has become the backbone framework for most pre-trained language models. In this paper, we directly use BERT and its enhanced variant Roberta as the baseline systems for knowledge-driven MRC

Table 1. A summary of key notations used in this paper.

Notation	Description
Q, q_i	The given question and its i-th token
C, c_i	The given passage and its i-th token
A, a^i, a^i_j	The candidate answer set, the i-th candidate answer and its j-th token
$\boldsymbol{BE}_i$	The pre-trained word embeddings of the i-th token in the input sequence
$\boldsymbol{CE}_i$	The ConceptNet embeddings of the i-th token in the input sequence
`MHA`	The function for computing the multi-head attention score
`Attention`	The function for computing the single-head attention score
`FFN`	The feed-forward network
$\boldsymbol{W}^Q_j, \boldsymbol{W}^K_j, \boldsymbol{W}^V_j$	The query, key and value parameters of linear mapping layer for j-th head
H	The number of heads at each layer of Transformer-based LM
T	The maximum length of the input sequence
L	The number of layer in Transformer-based LM
Φ_{mask}	The mask function for masking out the context token
$\boldsymbol{M}$	The mask vector of length T
$\boldsymbol{h}^l$	The hidden state at the l-th layer
$\boldsymbol{u}^l_i$	The output representation of the i-th token from `MHA` at the l-th layer
$\tilde{\boldsymbol{u}}^l$	The fused output representation of the i-th token at the l-th layer
$\boldsymbol{v}^L$	The fused output representation from `MHA` at the L-th layer
m	The max length of the question Q
s	The max length of the passage C
k	The number of the candidate answers in the set A
d_1	The hidden size for the attention function
d_2	The dimension for the ConceptNet embeddings

tasks, which includes a multi-layer bidirectional Transformer encoder and a linear classifier. Following [22] we concatenate the context C, question Q, and answer option A_i as the input sequence:

$$[\mathrm{CLS}]c_{1..s}[\mathrm{SEP}]q_{1..m}[\mathrm{SEP}]a^i_{1..n}[\mathrm{SEP}]$$

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

$$\boldsymbol{BE}_i = \boldsymbol{e}^{tok}_i + \boldsymbol{e}^{pos}_i + \boldsymbol{e}^{seg}_i, i = 1..T$$

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

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. For the input sequence $\boldsymbol{X} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_T\}$, the self-attention is formally calculated as:

$$w_{i,j} = \frac{\boldsymbol{t}_i^q \boldsymbol{t}_j^{k\top}}{\sqrt{d}} \tag{1}$$

$$\alpha_{i,j} = \frac{\exp(w_{i,j})}{\sum_{k=1}^{T} \exp(w_{i,k})} \tag{2}$$

$$\hat{\boldsymbol{t}}_i = \sum_{j=1}^{T} \alpha_{i,j} \boldsymbol{t}_j^v \tag{3}$$

where d is the hidden size, $\boldsymbol{t}^q$, $\boldsymbol{t}^k$, and $\boldsymbol{t}^v$ are the query, key, and value vector for the token, respectively. In practice, they are obtained by passing the same token representation through three separate linear layers. In this way, each token pays attention to other tokens for global information. For simplicity, we denote the above processing progress (Eq. 1-Eq. 3) as a single function $\texttt{Attention}(\boldsymbol{X}^q, \boldsymbol{X}^k, \boldsymbol{X}^v)$.

Given a matrix of T query vectors $\boldsymbol{Q} \in \mathbb{R}^{T \times d_1}$, keys $\boldsymbol{K} \in \mathbb{R}^{T \times d_1}$ and values $\boldsymbol{V} \in \mathbb{R}^{T \times d_1}$, the multi-head attention is computed as:

$$b_p = \texttt{Attention}_p(\boldsymbol{Q}\boldsymbol{W}_p^Q, \boldsymbol{K}\boldsymbol{W}_p^K, \boldsymbol{V}\boldsymbol{W}_p^V) \tag{4}$$

$$B = Concat(b_1, ..., b_H) \tag{5}$$

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, $\boldsymbol{W}_p^Q \in \mathbb{R}^{T \times d_1/H}$, $\boldsymbol{W}_p^K \in \mathbb{R}^{T \times d_1/H}$ and $\boldsymbol{W}_p^V \in \mathbb{R}^{T \times d_1/H}$ are the parameters of linear mapping layer for p-th head. We denotes the above processing progress (Eq. 4-Eq. 5) as:

$$B = \texttt{MHA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \tag{6}$$

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.

$$\boldsymbol{u}^l = \texttt{MHA}(\boldsymbol{h}^l, \boldsymbol{h}^l, \boldsymbol{h}^l) \tag{7}$$

$$\boldsymbol{h}^{l+1} = \texttt{FFN}(\boldsymbol{u}^l) \tag{8}$$

$$\texttt{FFN}(\boldsymbol{x}) = \boldsymbol{W}_2\text{GELU}(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2 \tag{9}$$

where $\boldsymbol{h}^l \in \mathbb{R}^{T \times d_1}$ denotes the hidden state at the l-th layer. We utilize the input representations $\boldsymbol{BE}$ as the initial state $\boldsymbol{h}^0$. Note that we omit residual connection and layer normalization used in each sub-layer for simplicity, and refer readers to [15] and [16] for more details.

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

3. Incremental TrLM with Plugged-in Knowledge Integration Mechanism

Pre-trained language models based on Transformer backbone network have a powerful ability to represent the context of the given text, while 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 $\boldsymbol{CE}$ selected with a multi-hop co-occurrence mask (We will describe the knowledge representations and selection in §4) 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. The first approach, as illustrated in the "gate" part of Fig. 2, learns a feature mask from the obtained commonsense embeddings, which is applied to each token that satisfies commonsense relationship filtering. To be specific, for each token t_i, we integrate the input representations $\boldsymbol{BE}_i$ with external knowledge embeddings $\boldsymbol{CE}_i \in \mathbb{R}^{d_2}$ as:

$$\boldsymbol{In}_i = \boldsymbol{BE}_i + \sigma(\boldsymbol{W}_g \boldsymbol{CE}_i + \boldsymbol{b}_g) \tag{10}$$

where σ denotes the sigmoid activation function served as a gate mechanism and $\boldsymbol{W}_g \in \mathbb{R}^{d_1 \times d_2}$ is a trainable weight parameter. This gating mechanism generates a mask-vector from each $\boldsymbol{CE}_i$ with values between 0 and 1, incorporating information into salient dimensions of $\boldsymbol{BE}_i$.

Multi-level Linear Transformation. The "linear" part of Fig. 2 shows the second method which integrates the information in every 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

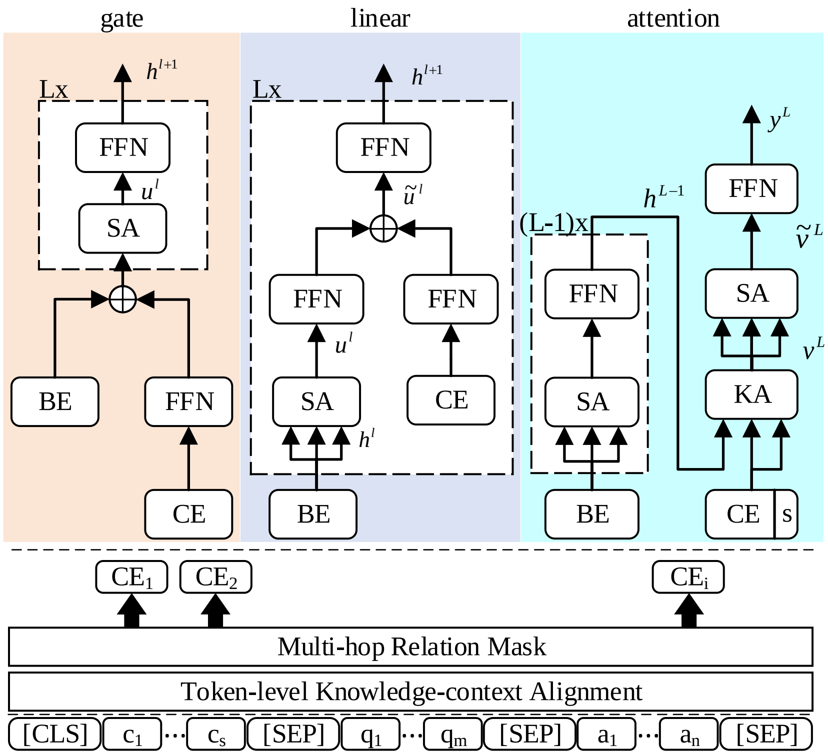


Fig. 2. Overview of the incremental language model. Three proposed fusion methods are abbreviated as "*gate*", "*linear*", and "*attention*", respectively.

representations and their corresponding commonsense embeddings, which is computed as:

$$\tilde{\boldsymbol{u}}_i^l = \mathrm{GELU}(\boldsymbol{W}_1^l \boldsymbol{u}_i^l + \tilde{\boldsymbol{W}}_1^l \boldsymbol{CE}_i + \boldsymbol{b}^l) \tag{11}$$

$$\boldsymbol{h}_i^{l+1} = \mathtt{FFN}(\tilde{\boldsymbol{u}}_i^l) \tag{12}$$

where $\tilde{\boldsymbol{W}}_1^l \in \mathbb{R}^{d_1 \times d_2}$ is a trainable weight parameter. Note that this method is in a similar spirit to the work of [18]. 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 do 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 attention calculation in TrLMs during pre-training contributes to multiple downstream tasks. Also, there are also efforts proving that the attention mechanism also plays an important role in multi-task learning [23]. We investigate using the attention mechanism to integrate external knowledge representations during fine-tuning. The third method, as depicted in the "attention" part of Fig. 2, is inspired by the work of [24] and applies attention-based integration to the final hidden states $\boldsymbol{h}^L$. Specifically, we add another multi-head attention sub-layers to the output of Transformer encoder block of a certain layer. In this paper, the sub-layer is a multi-head knowledge attention (KA) and placed on the last layer, which is computed as:

$$\boldsymbol{v}^L = \texttt{MHA}(\boldsymbol{h}^{L-1}, \tilde{\boldsymbol{CE}}, \tilde{\boldsymbol{CE}}) \tag{13}$$

where $\tilde{\boldsymbol{CE}}$ is a concatenation of $\boldsymbol{CE}$ and a knowledge sentinel $\boldsymbol{s} \in \mathbb{R}^{d_2}$. Considering not all tokens are relevant to the background knowledge, we follow [25] 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 $\boldsymbol{v}^L$ and feed them into the second sub-layer, which consists of a multi-head self-attention and a `FFN`:

$$\tilde{\boldsymbol{v}}^L = \texttt{MHA}(\boldsymbol{v}^L, \boldsymbol{v}^L, \boldsymbol{v}^L) \tag{14}$$

$$\boldsymbol{y}^L = \texttt{FFN}(\tilde{\boldsymbol{v}}^L) \tag{15}$$

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

4. Commonsense Representation and Filtering

In this work, we leverage off-the-shelf commonsense knowledge from ConceptNet 5.5[1], which is a knowledge graph (KG) including lexical and world knowledge from many different sources such as WordNet [26] and DBPedia [27]. Commonsense in ConceptNet is represented in the form of a triple (*subject*, *relation*, *object*). Below we first introduce the representations of commonsense knowledge, and then present a token-level multi-hop knowledge filtering method.

[1] `https://github.com/commonsense/conceptnet5/wiki`

4.1. Knowledge Graph Embedding

Intuitively, the fusion method should be independent of how the knowledge representation is obtained. Previous methods usually use their own commonsense embeddings (such as TransE) pre-trained on a large-scale knowledge graph, which increases computational consumption and uncertainty. Instead, we directly leverage off-the-shelf ConceptNet embeddings as external knowledge representation that can represent global commonsense relationships. In addition, this can also provide better scalability and a convenient interface for integrating other external knowledge representations. To be specific, we retrieve 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-PPMI [2], ConceptNet Numberbatch [3], and Randomly Initialized Embedding. We will discuss the impact of knowledge representation on model performance in Section 6.1.

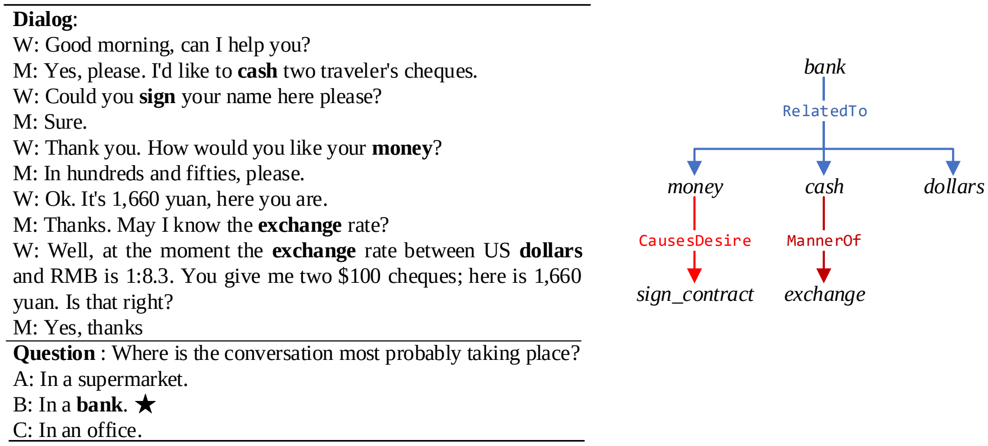
ConceptNet-PPMI. A matrix of word embeddings trained on a sparse, symmetric term-term matrix where each cell contains the sum of the weights of all edges that connect the two corresponding terms. For each term in the ConceptNet graph, its ConceptNet-PPMI representation reflects the context containing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

[2] `https://conceptnet.s3.amazonaws.com/precomputed-data/2016/numberbatch/16.09/conceptnet-55-ppmi.h5`

[3] `https://github.com/commonsense/conceptnet-numberbatch`



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.

4.2. Token-level Multi-hop Knowledge Filtering

These extracted and generated embeddings represent relevant and useful information between tokens. Nevertheless, background knowledge also needs to be used effectively, otherwise, it will become unnecessary noise. 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 $\boldsymbol{M}$ to filter commonsense representations. Specifically, the length of $\boldsymbol{M}$ is the same as the sequence length T and we initialize the mask values of all the tokens to 1. For each token $t_1 \in A_i$ that is neither a stop word nor a padding token, we 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 set $\boldsymbol{M}_{index(t_1)} = 0$ and $\boldsymbol{M}_{index(t_2)} = 0$ 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 Algorithm 1.

Thus, we obtain the mask vector $\boldsymbol{M}$, which only contains binary values. We further define the mask operation as follow:

$$\Phi_{mask}(\boldsymbol{CE_i}) = \begin{cases} \boldsymbol{CE_i} & , \boldsymbol{M}_i = 0 \\ 0 & , \boldsymbol{M}_i = 1 \end{cases} \tag{16}$$

Algorithm 1 Procedure of the token-level multi-hop knowledge filtering mechanism

```
Input: The mask vector $\boldsymbol{M} \in \mathbb{R}^T$, the context $C$, the question $Q$, the $i$-th
  candidate answer $a^i$, the ConceptNet triples $(S, R, O)$, the number of hop $K$
Output: The filtered mask vector $\boldsymbol{M}$
  Initialize the values of $\boldsymbol{M}$ to 1
  for $k$ in $K$ do
    $[] \Leftarrow List_k$
  end for
  $k \Leftarrow 0$
  $List_k$.add$(a_1^i, a_2^i, ..., a_m^i)$
  for $l_i^k$ in $List_k$ do
    if $l_i^k$ is neither a stop word nor a padding token and $l_i^k$ in $S$ then
      for $t_k$ in $C \cup Q$ do
        if $t_k$ in $O' \subseteq O \xleftarrow{R} S(l_i^k)$ then
          $0 \Leftarrow \boldsymbol{M}_{index(t^k)}$
          $0 \Leftarrow \boldsymbol{M}_{index(l_i^k)}$
          $List_{k+1}$.add$(t^k)$
        end if
      end for
    end if
    $k \Leftarrow k + 1$
    if $k \leqslant K$ then
      Continue
    end if
  end for
```

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

$$\boldsymbol{CE_i} = \sum_{k=1}^{K} \alpha_k * \boldsymbol{c}_{i,k} \tag{17}$$

$$\alpha_k = \frac{e^{\boldsymbol{w}\boldsymbol{c}_{i,k}}}{\sum_{k=1}^{K} e^{\boldsymbol{w}\boldsymbol{c}_{i,k}}} \tag{18}$$

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

The filtered commonsense embeddings $\boldsymbol{CE}$ will be taken as input to the three fusion methods, as 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.

5. Experiments

5.1. Dataset and Evaluation Metric

We report results on two well-known knowledge-driven datasets, CosmosQA [14] and DREAM [11]. For a fair comparison, we also evaluate our models on the RACE [3] dataset that is one of the largest datasets in multi-choice MRC with few commonsense questions. The statistics of these three datasets are summarized in Table 2, and a brief introduction to these datasets is given below.

CosmosQA[4] is a large-scale dataset that requires commonsense-based reading comprehension, formulated as multiple-choice questions. In contrast to the 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[5] 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 a model to answer these questions not only by advanced reading skills but also with rich background knowledge.

RACE[6] 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.

[4] Leaderboard: `https://leaderboard.allenai.org/cosmosqa/submissions/public`

[5] Leaderboard: `https://dataset.org/dream/`

[6] Leaderboard: `http://www.qizhexie.com/data/RACE_leaderboard.html`

Table 2. 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	8.8*

Table 3. The best hyperparameters on different datasets (BERT-base/BERT-large/Roberta-large). T denotes the max sequence length.

Dataset	lr	epoch	batch size	T
CosmosQA	$2e^{-5}/2e^{-5}/1.5e^{-5}$	10/8/4	32/32/32	256
DREAM	$2e^{-5}/2e^{-5}/1.5e^{-5}$	8/8/4	24/12/12	512
RACE	$3e^{-5}/2e^{-5}/2e^{-5}$	3/3/3	16/8/8	512

5.2. Implementation Details

We implemented our experiments using Huggingface[7]. We used BERT-base, BERT-large and Roberta-large as baseline systems. To keep the order of magnitude close, we used L2 normalization to preprocess ConceptNet-PPMI. We experimented with commonsense relation searching of up to three hops. We set $K = 3$. The embeddings of commonsense were fixed during the fine-tuning process and the parameters of TrLMs were trainable and initialized from Huggingface checkpoint. For all fine-tuning experiments, we used BertAdam as the optimizer. We experimented with 10 different random seeds and computed the average results, whether it was base or large. For the test set and the submitted version, we used the best performing model on the dev set to predict.

For training, we run all experiments on two Nvidia Titan RTX 24GB GPUs. 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 took about 8 hours to get the best result. For DREAM

[7] `https://github.com/huggingface/transformers`

dataset, we run experiments for 8 epochs, set the max sequence length to be 512, and selected the hyperparameters from batch size: {8, 12, 24, 36}, learning rate: {2e-5, 1e-5, 8e-6}. It took 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 selected the hyperparameters from batch size: {8, 16, 32}, learning rate: {3e-5, 2e-5, 1e-5}. It took about 12 hours to get the best result. In Table 3, we present the best hyperparameters on the development set and used them to verify on the test set.

5.3. Results

In addition to the three vanilla TrLMs on the leaderboards, we have also compared our methods with the following published models.

(1) **Fine-tuned GPT** conducts generative pre-training of the Transformer decoder on a diverse corpus of unlabeled text, followed by discriminative fine-tuning on each specific task. It makes use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture.

(2) **OCN** compares options at word-level to better identify their correlations to help reasoning. Each option is encoded into a vector sequence using a skimmer to retain fine-grained information as much as possible. An attention mechanism is leveraged to compare these sequences vector-by-vector to identify more subtle correlations between options.

(3) **MMM** [28] is a multi-stage multi-task learning framework for multi-choice reading comprehension. It mainly involves two sequential stages: coarse-tuning stage using natural language inference (NLI) task datasets and multitask learning stage using a larger in-domain dataset to help model generalize better with limited data.

(4) **DUMA** [29] uses a dual multi-head co-attention to bidirectionally capture relationships among passage, question and answer options for multi-choice MRC. The question-answer-aware passage representation simulates human re-reading details in the passage with impression of question and answer, and the passage-aware question-answer representation simulates re-considering the question-answer with deeper understanding of the passage.

(5) **DCMN+** [9] models the relationship among passage, question and answer options bidirectionally. Besides, inspired by how humans solve multi-choice questions, it integrates two reading strategies: (i) passage sentence selection, (ii) answer option interaction.

(6) **Multiway** [14] is the official baseline model of CosmosQA, which computes interactive attention to capture the relationship among the passage, question and answer. Taking the context as an example, it computes three types of attention weights to capture the passage's correlation to the question, the answer, and both the question and answer.

ConceptNet Numberbatch was used as commonsense representation (We will discuss the role of knowledge embedding in §6), and we applied a two-hop commonsense relationship to filter knowledge. The overall results are shown in Table 4.

Internal Comparison. From Table 4, we can observe that our plug-in methods of incorporating commonsense have significantly improved performance over the vanilla TrLMs, and in particular, (1) Multi-level linear transformation achieves the best results on CosmosQA (69.2% vs. 66.8% with BERT-large and 80.1% vs. 81.9% with Roberta-large) and DREAM (65.3% vs. 62.8% with BERT-base, 69.3% vs.66.6% with BERT-large and 87.0% vs. 84.7% with Roberta-large). Compared with the other two methods, additive feature-based gating has improved less on CosmosQA and DREAM, while multi-head attentional fusion has shown a negative impact on RACE; (2) As the scale of the pre-training model increases, the effect of integrating external commonsense decreases. According to [32], we can hypothesize that TrLMs with more sufficient pre-training (more data and longer pre-training time) may contain more implicit commonsense relationships; (3) In knowledge-driven tasks, variants of the incremental model have obtained 1%-4.1% improvement in average accuracy over the baselines of directly fine-tuned TrLMs. In contrast, our increment TrLMs have also achieved comparable results on RACE. It is reasonable and expected for our models to achieve limited improvement on RACE. Compared to DREAM and CosmosQA, which contain a higher proportion of commonsense questions, it is the experimental results on RACE that show that language models enhanced by commonsense representation can be directly improved in the fine-tuning stage. On the other hand, it illustrates our methods do not lose the textual information after heterogeneous knowledge fusion.

Table 4. Accuracy (%) on the test 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, "-L" means the large model and "-R" means the Roberta-large model. ‡ indicates that there are better results than this on the leaderboard, and we extract the result with public hyperparameters similar to ours. Due to the submission limit of CosmosQA, we only evaluate the incremental BERT-large and Roberta-large models and publish the best results.

Model	Knowledge-driven MRC		Few-commonsense MRC
	CosmosQA	DREAM	RACE
Leaderboard			
BERT-B	62.9	63.2	65.0
BERT-L	–	66.8	72.0
BERT-B+WAE	–	64.7	–
BERT-L+WAE	–	69.0	–
Roberta-L	79.2‡	–	83.2
Publication			
Fine-tuned GPT-B [30]	–	–	59.0
OCN-B [31]	–	–	66.8
OCN-L [31]	–	–	71.7
MMM-B [28]	–	72.2	68.0
MMM-L [28]	–	76.0	72.5
MMM-R [28]	–	**88.9**	**85.0**
DUMA-B [29]	–	64.0	–
DCMN-B [9]	–	–	67.0
DCMN+-L [9]	–	–	75.8
Multiway-L [14]	68.4	–	–
Ours (Concept Numb.+2hop)			
BERT-B	–	62.8	65.0
BERT-B$_{gate}$	–	64.8 (+2.0)	64.9 (-0.1)
BERT-B$_{linear}$	–	65.3 (+2.5)	65.3 (+0.3)
BERT-B$_{attention}$	–	64.5 (+1.7)	64.5 (-0.5)
BERT-L	66.8	66.6	72.1
BERT-L$_{gate}$	67.9 (+1.1)	67.8 (+1.2)	72.4 (+0.3)
BERT-L$_{linear}$	69.2 (+2.4)	69.3 (+2.7)	72.6 (+0.5)
BERT-L$_{attention}$	68.6 (+1.8)	68.3 (+1.7)	72.0 (-0.1)
Roberta-L	80.1	84.7	83.3
Roberta-L$_{gate}$	80.9 (+0.8)	85.7 (+1.0)	83.6 (+0.3)
Roberta-L$_{linear}$	**81.9** (+1.8)	87.0 (+2.3)	83.7 (+0.4)
Roberta-L$_{attention}$	81.5 (+1.4)	86.6 (+1.9)	83.5 (+0.2)

Comparison with Public Models. Compared to these public models, our incremental TrLMs have also achieved competitive performance but need less computational cost. Table 5 summarizes the computational resources of several published models. **MMM** achieved the best results on DREAM by first fine-tuning the sentence encoder with NLI task datasets MultiNLI [33] and SNLI [34], and then fine-tuning on RACE together through multi-task learning, which required more data and computing resources. **DCMN+** and **DUMA** designed complicated matching network patterns among the passage, the question and the candidate answer, increasing the scale of the model and the number of parameters. Instead, the proposed methods have two advantages: 1) the incremental TrLMs have improved the performance requiring no expensive further pretraining on out-of-domain datasets; 2) only a few mapping parameters and a single layer of parallel attention calculation are added to fuse commonsense into TrLMs. In addition, prediction results involving commonsense questions can not be clearly explained. On the contrary, we directly incorporate off-the-shelf commonsense representations into the internal structure of Transformer through token-level pre-matching to achieve the purpose of explicit use of external knowledge, obtaining interpretable performance improvement with fewer parameters (See Section 6.3). There also have been better models submitted on the leaderboard, they did not publish their methods.

6. Discussion

6.1. Effectiveness of Knowledge Embedding

Table 6 shows the results of our incremental TrLMs obtained by adding initialization with different commonsense representations. From this table, we can see that the integration of Concept Numbatch can generally improve the performance, which shows that the structured representations including the multi-source semantics can effectively improve the performance of the pre-training language models. Adding Concept-PPMI globally has a negative impact on the performance of BERT and Roberta while fusing it according to multi-hop commonsense relation improves the results. A reason for this could be 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 can help to effectively utilize the structured information, which is also demonstrated in the experiment with random initialization. Moreover, the incremental

Table 5. The computation cost comparison between our proposed methods and published methods. Target refers to the target datasets including DREAM and CosmosQA. BERT-base is taken as the baseline. $*$ means the results are extracted from [29], where albert-base is used as the baseline. The hidden size of BERT-base is the same as that of albert-base.

Model	Pre-training	Data	Matching Network	#Parameter (M)

MMM	✓	{MultiNLI, SNLI}
		→ {RACE, Target}
			GRU(C, QA)	-

DCMN+	✗	Target	Attention(A_i, A_j) BiAttention(C, Q) BiAttention(C, A) BiAttention(Q, A)	19.4*
DUMA	✗	Target	MHA(C, QA, QA) MHA(QA, C, C)	13.5*
gate	✗	Target	✗	0.23
linear	✗	Target	✗	2.76
attetnion	✗	Target	✗	1.05

model using random initialization commonsense has performed better than using Concept-PPMI in global fusion, which means heterogeneous information could be difficult to integrate directly without prior filtering since the pre-training procedure for language representation is quite different from the knowledge representation procedure.

We also experimented with other token attributes such as Part-Of-Speech (POS) and entity type to further analyze the role of commonsense representation. For POS, we used Stanford CoreNLP toolkit[8] to tag the input sequence, obtaining *noun*, *verb*, *adjective*, etc. Besides, we used spaCy[9] to recognize the entity types including *PERSON*, *ORG*, *LOC*, and *DATE*. We set 50-dimensional vectors for POS and entity types. For multiple features fusion, we replaced $\boldsymbol{CE}$ with the concatenation of heterogeneous representations. Since POS and entity type do not need to be determined by commonsense relationships, we investigate the global integration. Table 7 shows the results From the table, we can observe that: (1) Compared to ConceptNet Numberbatch, POS and entity type have little performance gains for TrLMs on

[8] `https://stanfordnlp.github.io/CoreNLP/`

[9] `https://spacy.io/`

Table 6. Performance in accuracy (%) with different knowledge representation. We use DREAM development set for analysis. "global" means common- sense representations are integrated into all tokens.

KG Embeddings	global	one-hop	two-hop	three-hop
BERT-B_{linear}				
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**
Roberta-L_{gate}				
Random	84.5	84.6	85.0	84.7
Concept-PPMI	84.4	84.9	85.7	85.1
Concept Numb.	**85.6**	**85.2**	**86.2**	**85.7**

knowledge-driven tasks, while they have a more significant impact on RACE. We hypothesize that it is difficult to distinguish the correct answer to the commonsense question based on linguistic and entity information alone; (2) Entity type has a more significant impact on TrLMs performance for QA than POS. A possible reason is that entity information can help TrLMs clarify reasonable answer types (e.g., *where* → *LOC*).

6.2. Effectiveness of Multi-hop Commonsense Selection

Table 8 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 8). From this table, we can see that: (1) All three methods have achieved their own best results in the two-hop commonsense relation search, which means that the indirect commonsense concept may not always work; (2) Multi-head attentional fusion is less sensitive to the hop number of commonsense relationship search than the other two methods (±0.6 points for *attention* vs. ±1.5 points for *gate* vs. ±1.1 points for *linear*), which is probably attributed to the speculation that the attention mechanism can effectively distinguish excessive noise while the other two methods are "hard fusion"; (3) Interestingly, additive feature-based gating with global commonsense has performed better than itself with one-hop commonsense on DREAM and CosmosQA. We can hypothesize that the ConceptNet Numberbatch contains text-based lexicon information since it is obtained by jointly retrofitting from word2vec and GloVe; (4) Compared with the two datasets, the performance of the three methods on RACE did

Table 7. Performance in accuracy (%) with different token attributes. We use global BERT-base$_{gate}$ and BERT-base$_{attention}$ as the baseline systems. “C”, “P”, and “T” denote the ConceptNet Numberbatch, POS, and Entity type representations, respectively.

Methods	C	P	T	CosmosQA	DREAM	RACE
gate	✓	✗	✗	**64.1**	64.5	64.5
	✗	✓	✗	62.9	63.1	64.1
	✗	✗	✓	63.2	63.3	64.7
	✓	✓	✗	63.6	64.5	64.6
	✓	✓	✓	63.9	**64.7**	**65.1**
attention	✓	✗	✗	64.3	**64.6**	64.0
	✗	✓	✗	62.6	63.1	64.0
	✗	✗	✓	62.9	63.4	64.3
	✓	✓	✗	63.8	64.3	64.2
	✓	✓	✓	**64.4**	64.5	**64.6**

not fluctuate much with the number of hops, which once again shows that RACE hardly requires external commonsense to identify correct options.

6.3. Case Study for Self-attention

To verify our goal to enable the self-attention in TrLMs to identify the knowledge-aware tokens, we consider the case depicted in Fig. 3. In this case, the vanilla BERT has chosen the wrong candidate option (A) and our models can 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. 4(a) and Fig. 4(b), obtained from the penultimate self-attention layer of BERT and two-hop BERT-base$_{linear}$, respectively.[10] For BERT, the token “*bank*” has a low degree of similarity to all the tokens $t_i \in C$ except “*traveler*” and “*cheques*”, and the focus of almost all the tokens in the dialog is quite discrete. Moreover, part of the 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 to

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

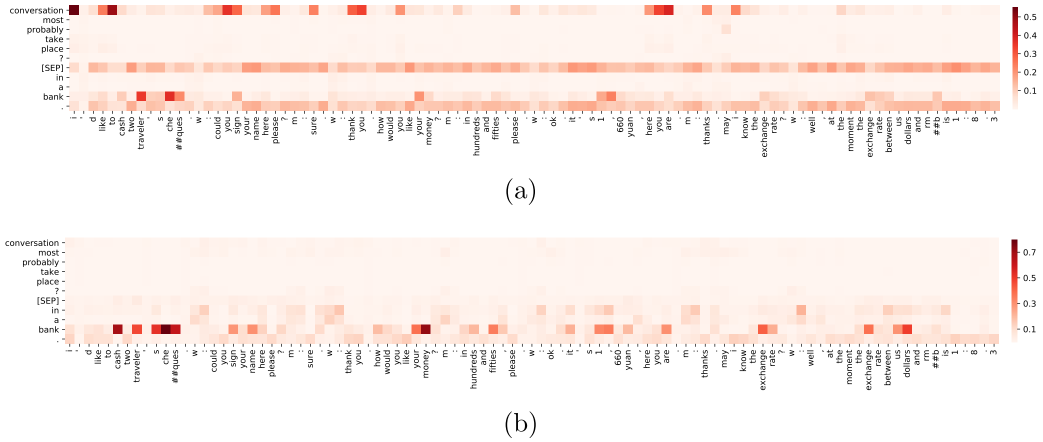
Table 8. Accuracy (%) on the CosmosQA, DREAM and RACE dataset based on the different number of hop commonsense relation searching, where "global" means commonsense representations are integrated into all tokens.

Model	CosmosQA				DREAM				RACE			
	global	1h	2h	3h	global	1h	2h	3h	global	1h	2h	3h
BERT-B$_{gate}$	64.1	63.7	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$_{attention}$	64.3	64.5	64.9	64.6	64.6	64.5	64.8	64.5	64.0	64.4	64.4	64.1
Roberta-L$_{gate}$	80.5	80.4	81.1	80.8	85.6	85.2	86.2	85.7	83.4	83.4	83.6	83.1
Roberta-L$_{linear}$	80.8	81.5	**81.8**	81.4	86.3	86.5	**87.3**	86.2	83.7	83.8	**84.1**	83.6
Roberta-L$_{attention}$	80.9	81.1	81.5	81.4	86.2	86.4	86.8	86.3	83.7	83.6	83.9	83.7

infer the correct answer. From Fig. 4(b), 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 vanilla similarity between "*bank*" and "*cheques*" has also been retained or even strengthened. It illustrates that the commonsense fusion method preserves textual information while effectively utilizing heterogeneous knowledge.

6.4. Robustness for Incomplete Training Set

TrLMs pre-trained on large-scale texts have appeared 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 LM. We show the results of three methods in different incomplete training set settings in Fig. 5, using BERT-base and Roberta-large. We can see that the performance of all models has demonstrated a similar trend with the decrease in training set size. Larger pre-trained models were more stable as the training set decreased. Compared to the vanilla TrLMs, our incremental models have maintained 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 set size drops to 60% and 40%. We argue that external commonsense would be more needed when the scale of text training set decreases to a certain extent. Augmenting TrLMs with external knowledge incorporation results in more robust performance in the settings with an incomplete training set.



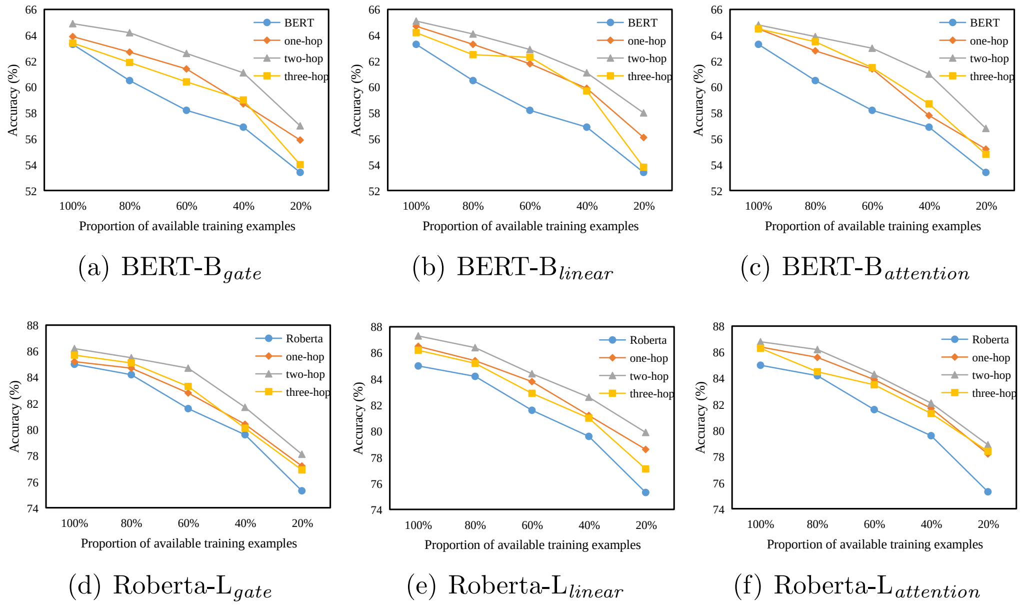
(a)

(b)

Fig. 4. 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.

6.5. Relation Ablations

We have also examined the effect of different relationships for commonsense filtering. As shown in Table 9, we use two-hop BERT-B$_{linear}$ and select 6 commonsense relations in ConceptNet and conduct an ablation study on two knowledge-driven datasets. For CosmosQA, the relations `RelatedTo`, `IsA` and `HasContext` have played a relatively important role in the performance of the model, which is reasonable since CosmosQA includes quite a proportion of assumption and pre-/post-condition type of questions ("*What may (or may not) be the plausible reason for...*" or "*What may (or may not) happen before (or after, or during)...*"). Correspondingly, the relations `FormOf`, `Synonym` and `AtLocation` have less influence because CosmosQA pays less attention to lexical commonsense and "*where*" question design. For DREAM, it is worth mentioning that the relation `AtLocation` has a greater impact on model performance although it occupies a small proportion. This is because DREAM designs considerable questions about the conversation scene ("*Where might this conversation happen?*"). Additionally, since there are more temporal and spoken words involved in the conversations, the relations `FormOf` and `HasContext` have more influence in DREAM than CosmosQA. Note that these relationships are not mutually exclusive, hence, only keeping/excluding a relationship does not affect performance linearly.



(a) BERT-B$_{gate}$

(b) BERT-B$_{linear}$

(c) BERT-B$_{attention}$

(d) Roberta-L$_{gate}$

(e) Roberta-L$_{linear}$

(f) Roberta-L$_{attention}$

Fig. 5. Accuracy on DREAM development set with the decrease in training set size. BERT-base and Roberta-large are used for comparison.

6.6. Investigation for Integration Layer

We have further investigated which layers the fusion methods are applied to are more effective. To this end, we have conducted experiments on two methods, multi-level linear transformation and multi-head attentional fusion. Taking the BERT-base as the benchmark, we examined the effect of multi-level linear transformation on the bottom six layers and the top six layers, and the performance of multi-head attentional fusion on the top three layers, respectively. Table 10 shows the accuracy on DREAM. First, applying multi-level linear transformation to only the top six layers has achieved better results than applying only to the bottom six layers. This indicates that commonsense knowledge is easier to capture and use explicitly on higher layers during fine-tuning, which is consistent with the conclusion of Liu et al. [35]. Second, multi-head attentional fusion has given the best performance when the output on the top layer was used for prediction, and the accuracy generally has decreased as the fusion layer moved to the bottom. This demonstrates that the self-attention in the bottom Transformer block is more inclined to focus on lexical and syntactic information, while the explicit

Table 9. Relation ablation study on the development sets of CosmosQA and DREAM. The column of "Proportion" reports the percentage of commonsense relation types in ConceptNet. Two-hop BERT-B$_{linear}$ is used for analysis.

Relation	Proportion (%)	CosmosQA		DREAM	
		Only_Keep	Remove	Only_Keep	Remove
RelatedTo	52.3	64.0 (-1.3)	63.5 (-1.8)	64.0 (-1.1)	63.6 (-1.5)
FormOf	9.7	62.9 (-2.4)	65.5 (+0.2)	63.3 (-1.8)	64.7 (-0.4)
IsA	8.6	63.0 (-2.3)	64.8 (-0.5)	63.2 (-1.9)	64.5 (-0.6)
Synonym	8.5	63.3 (-2.0)	65.0 (-0.3)	62.9 (-2.2)	64.5 (-0.6)
HasContext	6.3	63.4 (-1.9)	64.2 (-1.1)	63.6 (-1.5)	63.9 (-1.2)
AtLocation	1.0	63.0 (-2.3)	65.1 (-0.2)	63.6 (-1.5)	64.2 (-0.9)

Table 10. Accuracy (%) on the DREAM dataset with variants of different commonsense integration layer.

Model	Integration Layer		
	Layer 0-5	Layer 6-11	All layers
2-hop BERT-B$_{linear}$	64.6 (-0.5)	63.5 (-1.6)	65.1
	The 9th layer	The 10th layer	The 11th layer
2-hop BERT-B$_{attention}$	63.7 (-1.1)	64.4 (-0.4)	64.8

integration of external commonsense knowledge could play a greater role on the higher layer.

6.7. Error Analysis

We have conducted the following error analysis to investigate problems that our models are short of the ability to address. We randomly extracted 200 samples from the development set of DREAM, and then classified them into several question types according to the annotation criterion consistent with [11]. We compared two-hop BERT-base$_{linear}$ with BERT-base on these categories, as shown in Table 11. Both models performed 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 the vanilla BERT on the implicit questions (*e.g.*, under the categories *logic* and *commonsense*) which require external knowledge, our incremental model was less capable of an-

Table 11. 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

swering these questions under the category *summary*. We can hypothesize that integrating token-level commonsense may interfere with the reasoning requiring the aggregation of information from multiple sentences.

7. Related Work

Machine Reading Comprehension. In recent years, many MRC datasets have been released to solve different task scenarios, e.g., cloze-style [36, 37], extractive/abstractive answer [1, 2, 38, 39, 40], multi-choice [3], conversational QA [41, 42], multi-hop [4, 43], and whether external knowledge is needed [44, 10, 12, 13, 14]. Most MRC datasets that require external knowledge are designed in a multi-choice form. In this paper, we focus on the multi-choice knowledge-driven MRC task. For knowledge-driven datasets, SemEval-2018 Task 11 [45] and ROCStories [46] are also challenging datasets for knowledge-driven tasks. However, each question of both is only related to two candidate answers, potentially reducing the difficulty of multi-choice MRC. For few-commonsense multi-choice datasets, MCTest was released earlier and on a smaller scale compared with RACE. Hence, considering timeliness and novelty, we choose CosmosQA, DREAM, and RACE in the experiments. For multi-choice MRC, existing methods include designing the interaction among the passage, question and option [47, 31, 48, 9, 29], or transfer learning through data augmentation [28]. 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.

Yang and Mitchell [25] incorporated retrieved knowledge into LSTM by employing an attention mechanism with a sentinel. Bauer et al. [7] selected grounded multi-hop relational commonsense information from ConceptNet via pointwise mutual information and term-frequency based scoring function and used a selectively gated attention mechanism to fuse the knowledge. Mihaylov and Frank [49] introduced a mixed attention to external knowledge for cloze-style reading comprehension. Wang et al. [50] and Zhong et al. [51] explored the effect of semantic relations from KGs such as WordNet and ConceptNet on MRC. Sheng et al. [52] presented a sentence-level knowledge interaction module to integrate commonsense knowledge with corresponding sentence rather than the whole MRC instance. Wang and Jiang [53] proposed 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. [54] retrieved the corresponding entities and relation from text to aggregate answer evidence from an incomplete KB. Yang et al. [20] took BERT as encoder and employ an attention mechanism similar to [25] 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 Language Model. More recently, pre-trained Transformer-based language models such as BERT have shown powerful achievements in downstream tasks including MRC. The injection of external knowledge to TrLMs 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 [18, 19], which requires pre-training the TrLMs even from scratch. The second group is to pre-train external modules to assist TrLMs [51, 55]. In contrast, our fusion methods are to directly fine-tune on the target MRC datasets. A similar work at the same time is [56], which also focuses on injecting task-specific concept embeddings to BERT during fine-tuning. However, it used aligned entity vectors by string concatenation and replacement. Different from it, our alignment algorithm is based on multi-hop token search and fusion methods are based on the vectors' transformation inside the TrLMs.

8. Conclusion

This paper has introduced increment Transformer-based language models with three plug-in fusion methods, which can enhance vanilla TrLMs with commonsense representations from ConceptNet. We have used off-the-shelf ConceptNet embeddings as external knowledge representation and introduce a mask mechanism for token-level multi-hop relationship searching to filter external knowledge, so as to enable the self-attention in TrLMs to identify the knowledge-aware tokens effectively. Our models have achieved competitive improvements over baselines on two knowledge-driven machine reading comprehension datasets with a fewer computational cost. Extensive experiments have shown the effectiveness of the proposed methods and the robustness of the incremental models in the case of an incomplete training set. This work has demonstrated the role of commonsense incorporation into TrLMs in knowledge-driven reading comprehension. As for future work, in addition to addressing the problems mentioned in error analysis, we can start with more granular relationships and more commonsense sources to integrate external knowledge.

Acknowledgments

This research is supported by the Open Project Foundation of Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce [2020-KF-10].

References

[1] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, 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, 2016, pp. 2383–2392.
URL `http://aclweb.org/anthology/D/D16/D16-1264.pdf`

[2] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, 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., 2016.
URL `http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf`

[3] G. Lai, Q. Xie, H. Liu, Y. Yang, E. H. Hovy, 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, 2017, pp. 785–794.
URL `https://aclanthology.info/papers/D17-1082/d17-1082`

[4] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, C. D. Manning, 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, 2018, pp. 2369–2380.
URL `https://aclanthology.info/papers/D18-1259/d18-1259`

[5] M. J. Seo, A. Kembhavi, A. Farhadi, H. Hajishirzi, Bidirectional attention flow for machine comprehension, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, 2017.
URL `https://openreview.net/forum?id=HJ0UKP9ge`

[6] W. Wang, N. Yang, F. Wei, B. Chang, M. Zhou, 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, 2017, pp. 189–198. `doi:10.18653/v1/P17-1018`.
URL `https://doi.org/10.18653/v1/P17-1018`

[7] L. Bauer, Y. Wang, M. Bansal, 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, 2018, pp. 4220–4230.
URL `https://aclanthology.info/papers/D18-1454/d18-1454`

[8] M. Ding, C. Zhou, Q. Chen, H. Yang, J. Tang, 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, 2019, pp. 2694–2703.
URL `https://www.aclweb.org/anthology/P19-1259/`

[9] S. Zhang, H. Zhao, Y. Wu, Z. Zhang, X. Zhou, X. Zhou, 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, 2020, pp. 9563–9570.
URL `https://aaai.org/ojs/index.php/AAAI/article/view/6502`

[10] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, O. Tafjord, Think you have solved question answering? try arc, the AI2 reasoning challenge, CoRR abs/1803.05457 (2018). `arXiv:1803.05457`.
URL `http://arxiv.org/abs/1803.05457`

[11] K. Sun, D. Yu, J. Chen, D. Yu, Y. Choi, C. Cardie, DREAM: A challenge dataset and models for dialogue-based reading comprehension, TACL 7 (2019) 217–231.
URL `https://transacl.org/ojs/index.php/tacl/article/view/1534`

[12] T. Mihaylov, P. Clark, T. Khot, A. Sabharwal, 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, 2018, pp. 2381–2391.
URL `https://www.aclweb.org/anthology/D18-1260/`

[13] A. Talmor, J. Herzig, N. Lourie, J. Berant, 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), 2019, pp. 4149–4158.
URL `https://www.aclweb.org/anthology/N19-1421/`

[14] L. Huang, R. L. Bras, C. Bhagavatula, Y. Choi, Cosmos QA: machine reading comprehension with contextual commonsense reasoning, in: K. Inui, J. Jiang, V. Ng, X. Wan (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, 2019, pp. 2391–2401. doi:10.18653/v1/D19-1243.
URL https://doi.org/10.18653/v1/D19-1243

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, 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, 2017, pp. 5998–6008.
URL http://papers.nips.cc/paper/7181-attention-is-all-you-need

[16] J. Devlin, M. Chang, K. Lee, K. Toutanova, 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), 2019, pp. 4171–4186.
URL https://aclweb.org/anthology/papers/N/N19/N19-1423/

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). arXiv:1907.11692.
URL http://arxiv.org/abs/1907.11692

[18] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, 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, 2019, pp. 1441–1451.
URL https://www.aclweb.org/anthology/P19-1139/

[19] M. E. Peters, M. Neumann, R. L. L. IV, R. Schwartz, V. Joshi, S. Singh, N. A. Smith, Knowledge enhanced contextual word representations, in: K. Inui, J. Jiang, V. Ng, X. Wan (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, 2019, pp. 43–54.

doi:10.18653/v1/D19-1005.
URL https://doi.org/10.18653/v1/D19-1005

[20] A. Yang, Q. Wang, J. Liu, K. Liu, Y. Lyu, H. Wu, Q. She, S. Li, 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, 2019, pp. 2346–2357.
URL https://www.aclweb.org/anthology/P19-1226/

[21] R. Speer, J. Chin, C. Havasi, 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., 2017, pp. 4444–4451.
URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972

[22] X. Pan, K. Sun, D. Yu, J. Chen, H. Ji, C. Cardie, D. Yu, Improving question answering with external knowledge, in: A. Fisch, A. Talmor, R. Jia, M. Seo, E. Choi, D. Chen (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, 2019, pp. 27–37. doi:10.18653/v1/D19-5804.
URL https://doi.org/10.18653/v1/D19-5804

[23] G. Liang, H. Mo, Y. Qiao, C. Wang, J. Wang, Paying deep attention to both neighbors and multiple tasks, in: D. Huang, V. Bevilacqua, A. Hussain (Eds.), Intelligent Computing Theories and Application - 16th International Conference, ICIC 2020, Bari, Italy, October 2-5, 2020, Proceedings, Part I, Vol. 12463 of Lecture Notes in Computer Science, Springer, 2020, pp. 140–149. doi:10.1007/978-3-030-60799-9_12.
URL https://doi.org/10.1007/978-3-030-60799-9_12

[24] Z. Li, C. Niu, F. Meng, Y. Feng, Q. Li, J. Zhou, 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, 2019, pp. 12–21.
URL https://www.aclweb.org/anthology/P19-1002/

[25] B. Yang, T. M. Mitchell, 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, 2017, pp. 1436–1446. doi:10.18653/v1/P17-1132.
URL https://doi.org/10.18653/v1/P17-1132

[26] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller, Introduction to wordnet: An on-line lexical database, International journal of lexicography 3 (4) (1990) 235–244.
URL https://doi.org/10.1093/ijl/3.4.235

[27] M. Morsey, J. Lehmann, S. Auer, C. Stadler, S. Hellmann, Dbpedia and the live extraction of structured data from wikipedia, Program 46 (2) (2012) 157–181. doi:10.1108/00330331211221828.
URL https://doi.org/10.1108/00330331211221828

[28] D. Jin, S. Gao, J. Kao, T. Chung, D. Hakkani-Tür, 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, 2020, pp. 8010–8017.
URL https://aaai.org/ojs/index.php/AAAI/article/view/6310

[29] P. Zhu, H. Zhao, X. Li, Dual multi-head co-attention for multi-choice reading comprehension, CoRR abs/2001.09415 (2020). arXiv:2001.09415.
URL https://arxiv.org/abs/2001.09415

[30] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training.

[31] Q. Ran, P. Li, W. Hu, J. Zhou, Option comparison network for multiple-choice reading comprehension, CoRR abs/1903.03033 (2019). arXiv:1903.03033.
URL http://arxiv.org/abs/1903.03033

[32] L. Cui, S. Cheng, Y. Wu, Y. Zhang, Does BERT solve commonsense task via commonsense knowledge?, CoRR abs/2008.03945 (2020). arXiv:2008.03945.
URL https://arxiv.org/abs/2008.03945

[33] A. Williams, N. Nangia, S. R. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: M. A. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 1112–1122. doi:10.18653/v1/n18-1101.
URL https://doi.org/10.18653/v1/n18-1101

[34] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Trans. Assoc. Comput. Linguistics 2 (2014) 67–78.
URL https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/229

[35] N. F. Liu, M. Gardner, Y. Belinkov, M. E. Peters, N. A. Smith, Linguistic knowledge and transferability of contextual representations, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, 2019, pp. 1073–1094. doi:10.18653/v1/n19-1112.
URL https://doi.org/10.18653/v1/n19-1112

[36] K. M. Hermann, T. Kociský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, 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, 2015, pp. 1693–1701.
URL http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend

[37] F. Hill, A. Bordes, S. Chopra, J. Weston, 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, 2016.
URL `http://arxiv.org/abs/1511.02301`

[38] M. Joshi, E. Choi, D. S. Weld, L. Zettlemoyer, 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, 2017, pp. 1601–1611. `doi:10.18653/v1/P17-1147`.
URL `https://doi.org/10.18653/v1/P17-1147`

[39] B. Dhingra, K. Mazaitis, W. W. Cohen, Quasar: Datasets for question answering by search and reading, CoRR abs/1707.03904 (2017). `arXiv:1707.03904`.
URL `http://arxiv.org/abs/1707.03904`

[40] T. Kociský, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, E. Grefenstette, The narrativeqa reading comprehension challenge, TACL 6 (2018) 317–328.
URL `https://transacl.org/ojs/index.php/tacl/article/view/1197`

[41] E. Choi, H. He, M. Iyyer, M. Yatskar, W. Yih, Y. Choi, P. Liang, L. Zettlemoyer, Quac: Question answering in context, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics, 2018, pp. 2174–2184. `doi:10.18653/v1/d18-1241`.
URL `https://doi.org/10.18653/v1/d18-1241`

[42] S. Reddy, D. Chen, C. D. Manning, Coqa: A conversational question answering challenge, TACL 7 (2019) 249–266.
URL `https://transacl.org/ojs/index.php/tacl/article/view/1572`

[43] J. Welbl, P. Stenetorp, S. Riedel, Constructing datasets for multi-hop reading comprehension across documents, TACL 6 (2018) 287–302.
URL `https://transacl.org/ojs/index.php/tacl/article/view/1325`

[44] S. Zhang, X. Liu, J. Liu, J. Gao, K. Duh, B. V. Durme, Record: Bridging the gap between human and machine commonsense reading comprehension, CoRR abs/1810.12885 (2018). `arXiv:1810.12885`.
URL `http://arxiv.org/abs/1810.12885`

[45] S. Ostermann, M. Roth, A. Modi, S. Thater, M. Pinkal, Semeval-2018 task 11: Machine comprehension using commonsense knowledge, in: M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S. Bethard, M. Carpuat (Eds.), Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018, Association for Computational Linguistics, 2018, pp. 747–757. `doi:10.18653/v1/s18-1119`.
URL `https://doi.org/10.18653/v1/s18-1119`

[46] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, J. F. Allen, A corpus and cloze evaluation for deeper understanding of commonsense stories, in: K. Knight, A. Nenkova, O. Rambow (Eds.), NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, The Association for Computational Linguistics, 2016, pp. 839–849. `doi:10.18653/v1/n16-1098`.
URL `https://doi.org/10.18653/v1/n16-1098`

[47] S. Wang, M. Yu, J. Jiang, S. Chang, A co-matching model for multi-choice reading comprehension, in: I. Gurevych, Y. Miyao (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, 2018, pp. 746–751. `doi:10.18653/v1/P18-2118`.
URL `https://www.aclweb.org/anthology/P18-2118/`

[48] H. Miao, R. Liu, S. Gao, A multiple granularity co-reasoning model for multi-choice reading comprehension, in: International Joint Conference

on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019, IEEE, 2019, pp. 1–7. doi:10.1109/IJCNN.2019.8852176.
URL https://doi.org/10.1109/IJCNN.2019.8852176

[49] T. Mihaylov, A. Frank, 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, 2018, pp. 821–832. doi:10.18653/v1/P18-1076.
URL https://www.aclweb.org/anthology/P18-1076/

[50] L. Wang, M. Sun, W. Zhao, K. Shen, J. Liu, 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, 2018, pp. 758–762.
URL https://aclanthology.info/papers/S18-1120/s18-1120

[51] W. Zhong, D. Tang, N. Duan, M. Zhou, J. Wang, J. Yin, Improving question answering by commonsense-based pre-training, in: J. Tang, M. Kan, D. Zhao, S. Li, H. Zan (Eds.), Natural Language Processing and Chinese Computing - 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9-14, 2019, Proceedings, Part I, Vol. 11838 of Lecture Notes in Computer Science, Springer, 2019, pp. 16–28. doi:10.1007/978-3-030-32233-5_2.
URL https://doi.org/10.1007/978-3-030-32233-5_2

[52] Y. Sheng, M. Lan, Residual connection-based multi-step reasoning via commonsense knowledge for multiple choice machine reading comprehension, in: T. Gedeon, K. W. Wong, M. Lee (Eds.), Neural Information Processing - 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12-15, 2019, Proceedings, Part III, Vol. 11955 of Lecture Notes in Computer Science, Springer, 2019, pp. 340–352. doi:10.1007/978-3-030-36718-3_29.
URL https://doi.org/10.1007/978-3-030-36718-3_29

[53] C. Wang, H. Jiang, 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, 2019, pp. 2263–2272.
URL `https://www.aclweb.org/anthology/P19-1219/`

[54] W. Xiong, M. Yu, S. Chang, X. Guo, W. Y. Wang, 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, 2019, pp. 4258–4264.
URL `https://www.aclweb.org/anthology/P19-1417/`

[55] R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, J. Ji, G. Cao, D. Jiang, M. Zhou, K-adapter: Infusing knowledge into pre-trained models with adapters, CoRR abs/2002.01808 (2020). `arXiv:2002.01808`.
URL `https://arxiv.org/abs/2002.01808`

[56] N. Pörner, U. Waltinger, H. Schütze, E-BERT: efficient-yet-effective entity embeddings for BERT, in: T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020, Association for Computational Linguistics, 2020, pp. 803–818. `doi:10.18653/v1/2020.findings-emnlp.71`.
URL `https://doi.org/10.18653/v1/2020.findings-emnlp.71`