DATA PAPER

Knowledge Graph based Mutual Attention for Machine Reading Comprehension over Anti-Terrorism Corpus

Feng Gao[1,2,3], Jin Hou[1,2,3†], Jinguang Gu[1,2,3], Lihua Zhang[4]

[1]School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, Hubei

[2]The Key Laboratory of Rich-Media Knowledge Organization and Service of Digital Publishing Content, Institute of Scientific and Technical Information of China, Beijing 100038, China

[3]Wuhan University of Science and Technology Big Data Science and Engineering Research Institute, Wuhan 430065, Hubei

[4]Eastchina Jiaotong University, Nanchang 330013, Jiangxi

Keywords: Machine reading comprehension; Anti-terrorism domain; Knowledge embedding; Knowledge attention; Mutual attention

Citation: Gao, F., Hou, J., Gu, J.G., Zhang, L.H.: Knowledge Graph based Mutual Attention for Machine Reading Comprehension over Anti-Terrorism Corpus. Data Intelligence 5(3), 685-706 (2023). doi: https://doi.org/10.1162/dint_a_00210
Submitted: May 20, 2022; Received: December, 15, 2022; Accepted: April, 16, 2023

ABSTRACT

Machine reading comprehension has been a research focus in natural language processing and intelligence engineering. However, there is a lack of models and datasets for the MRC tasks in the anti-terrorism domain. Moreover, current research lacks the ability to embed accurate background knowledge and provide precise answers. To address these two problems, this paper first builds a text corpus and testbed that focuses on the anti-terrorism domain in a semi-automatic manner. Then, it proposes a knowledge-based machine reading comprehension model that fuses domain-related triples from a large-scale encyclopedic knowledge base to enhance the semantics of the text. To eliminate knowledge noise that could lead to semantic deviation, this paper uses a mixed mutual attention mechanism among questions, passages, and knowledge triples to select the most relevant triples before embedding their semantics into the sentences. Experiment results indicate that the proposed approach can achieve a 70.70% EM value and an 87.91% F1 score, with a 4.23% and 3.35% improvement over existing methods, respectively.

† Corresponding author: Jin Hou (Email: 1371707917@qq.com).

1. INTRODUCTION

Machine reading comprehension (MRC) makes it possible for machines to "read" textual material and answer questions based on its content [1]. MRC has much potential in education, QA systems, intelligence engineering, etc. A typical MRC architecture is shown in Figure 1. MRC models first conduct contextual semantic analysis on vectorized input text and questions. Secondly, the relevance of different text segments to the question is typically calculated by an attention mechanism algorithm [2]. Finally, the answer to the question is marked as a text sequence from the context with the highest probability of answering the question. In this architecture, knowledge representation methods and annotated domain corpus play important roles [3], as the former allows the machine to "understand" the semantics of the question and its context and calculate its relevance, whereas the latter provides the training data and the testbed for fine-tuning [4, 5].

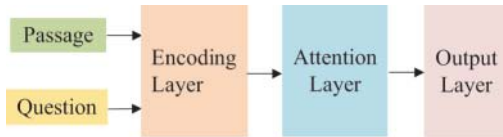

Figure 1. The basic architecture of MRC.

In 2018, the Google team released BERT [6] pre-training language model, arousing enthusiastic responses in the academic community. BERT served as a pre-training based knowledge representation model in 11 NLP tasks such as text classification, named entity recognition, sentiment analysis, and reading comprehension. Many of these tasks achieved state-of-the-art results at the time. Several efforts have been made to improve BERT, and numerous pre-trained models have been proposed since then, including RoBERTa [7], SpanBERT [8], ALBERT [9], etc.

Compared to pure corpus-based language models, knowledge-enhanced pre-training models leverage Knowledge Graphs (KGs) to provide inferencing capability for MRC [10, 11], as shown in Figure 2. However, most of the existing knowledge enhancement models use textual context for knowledge embedding [12] and do not consider the degree of match between questions and knowledge. In MRC, understanding the semantics of the question has an important impact on generating answers [13, 14]. Hence, it is desirable to embed relevant knowledge into both the text and the question without introducing too much knowledge noise, which could lead to semantic distortions of the original sentences and ultimately generate incorrect answers. In addition, most existing MRC datasets do not present sufficient samples and features for the anti-terrorism domain, which has characteristics such as being time-sensitive and containing prominent key entities. As a result, MRC models trained in the general domain cannot perform well on anti-terrorism tasks.

In summary, there are two main problems in enhancing the pre-trained model with external knowledge in the field of anti-terrorism: (1) Lack of datasets from the anti-terrorism domain to train and test anti-terrorism MRC models. (2) Lack of ability to inject accurate knowledge into the model to enrich the semantic information of the sentences.

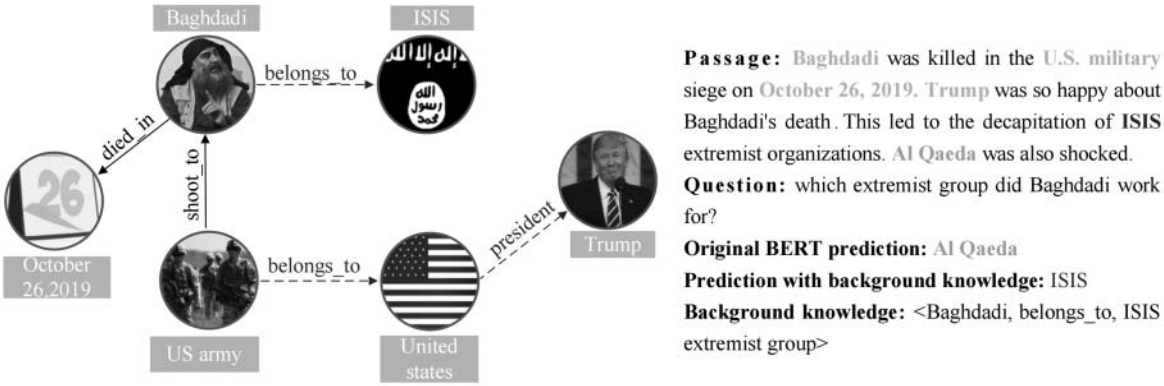

Figure 2. Examples of introducing external knowledge for language comprehension and intellectual reasoning. In the knowledge graphs, the black solid lines present the facts extracted from the sentence, and the red dotted lines present the existing knowledge facts.

In order to solve the above two problems, this paper constructs an extractive machine reading comprehension dataset (ATSMRC) in the anti-terrorism field by manual annotation and proposes KG-ATT-MRC model. The model is a fusion of external knowledge and a pre-trained language model, trained and tested on ATSMRC dataset. We use a named entity recognition tool to identify the entities involved in the passage and then link to CN-DBpedia based on these entities to select the knowledge triple associated with the passage. Finally, knowledge fusion is performed by computing the 3-way mutual attention among passages, questions, and knowledge triples.

The contributions of this paper are two-fold:

(1) We get nearly 15 years of anti-terrorism field news on China Daily, Sohu.com, and CCTV.com. Then we construct an extractive machine reading comprehension dataset (ATSMRC) in the anti-terrorism field through three steps: Data collection and data cleaning; text analysis and rule formulation; manual annotation.
(2) We propose KG-ATT-MRC model. It allocates lower scores to irrelevant knowledge triples by calculating the mixed mutual attention among knowledge, questions, and passages so as to embed knowledge.

The remainder of the paper is organized as follows. Section 2 describes the dataset and the work related to the knowledge-enhanced pre-trained models. Section 3 describes the process of ATSMRC dataset construction. Section 4 elaborates on the process of KG-ATT-MRC model construction. Section 5 presents the experimental results and analysis before Section 6 concludes.

2. RELATED WORK

2.1 MRC Dataset

The rapid development of MRC technology is inseparable from the release of large-scale machine reading comprehension datasets. Common MRC datasets include SQuAD[15], DuReader[16], CNN/Daily Mail[17], and cmrc2018[18]. SQuAD is a relatively sizeable English extraction MRC dataset constructed by humans. It contains 536 articles and 107,785 question-answer pairs. CNN/Daily Mail is a complete fill-in-the-blank MRC dataset proposed by Google, with CNN and Daily Mail data. cmrc2018 is a relatively large Chinese extractive MRC dataset with data from Wikipedia. In this paper, we focus on extractive MRC. Therefore, eight more common extractive MRC datasets are summarized in Table 1.

Table 1. Extracted MRC datasets information summary.

Dataset	Content	Language	Number of paragraphs	Number of questions
cmrc2018	Wikipedia	Chinese	-	19071
DROP	Wikipedia	English	7000	96567
HotpotQA	Wikipedia	English	-	112779
NewsQA	CNN	English	12744	119633
QuAC	Wikipedia	English	13594	98407
SearchQA	Wikipedia	English	4500000	140461
SQuAD	Wikipedia	English	536	107785
TrivialQA	Wikipedia	English	662659	95956

As shown in Table 1, most of the current extractive MRC datasets are in English and have generic domain content. There is a lack of datasets for the anti-terrorism domain. Therefore, this paper proposes extractive Chinese MRC datasets on the anti-terrorism domain. It can be used to test more MRC models in China and help Chinese MRC technology grow.

2.2 Knowledge Enhanced Pre-trained Models

Recent research has shown an interest in combining pre-trained models and knowledge graphs. There are four main research strains in fusing external knowledge and pre-trained language models.

1) Feature fused. Feature fused knowledge enhanced pre-trained models (KEPTMs) mainly focuses on entity information. They extract the features required by entities from KG and then embed knowledge. ERNIE 1.0 [12] uses TransE to pre-train the extracted entities to obtain a vector representation of the entities. It then passes the text vector and entity vector together to the multi-headed attention layer for encoding. Then the matrix multiplies the encoded vectors to obtain the fusion vector. However, entities and sentences are not encoded by the same encoder, and semantic space inconsistencies may exist. Its improved version, ERNIE 2.0 [19] proposes a continuous learning framework. Continuous incremental learning of vocabulary, syntax, and semantics by introducing more tasks.

2) Embedding combined. Embedding combined KEPTMs encodes knowledge in advance through knowledge representation methods [20]. The corresponding embeddings are then performed via a variant of the attention mechanism. SyntaxBERT [21] proposes a masking matrix of syntactic relations to model syntactic knowledge. Then the knowledge is integrated into the PTMs through the self-attention mechanism of syntactic relations. But SyntaxBERT does not consider external knowledge. BERT-MK [22] uses a novel transformer-based encoder model for learning knowledge embedding. However, BERT-MK model is similar to ERNIE without considering the heterogeneous space embedding problem. CokeBERT [23] dynamically selects knowledge subgraphs for embedding according to text context. JointLK [24] Given a question and a knowledge subgraph, the representations of the two modalities are obtained through an LM encoder and a GNN encoder, respectively. Finally, the fusion is performed through the attention mechanism.

3) Data Structure Unified. Data structure EMPTM transforms sequence and knowledge into a unified structure. The same encoder is then used for encoding, thus avoiding the heterogeneous spatial embedding problem. K-BERT [25] injects triples into sentences and generates a sequence model. It uses a weak position and a visible matrix to limit the effect of knowledge noise. Although K-BERT model considers the heterogeneous space problem, it is difficult to inject triples into sentences and convert them into sequence models when the knowledge subgraph is large. KMQA [13] fuses the knowledge in KG by obtaining the entity in the question and retrieving the corresponding triples in KG based on this entity and then converting the triples into sentences. However, converting triples into sentences is not a simple matter, and there may be the problem of sentence disfluency. KT-NET [10] selects entity-related knowledge from two knowledge bases: wordNet and NELL. The model uses the attention mechanism to get the most relevant knowledge vector, and then splices the knowledge vector and the text vector. Finally, the representation of BERT and KB is further fused through the self-attention mechanism to achieve the effect of improving the performance of MRC. But KT-NET does not take into account the relevance of knowledge to the question, and there may be some knowledge that is irrelevant to the question. CoLAKE [26] model is a more classical use of neural graph networks to fuse triples of information into a pre-trained model. For a given piece of text, CoLAKE treats it as a word graph formed by connecting multiple words, then extracts the knowledge subgraph from the knowledge graph centered on the entities mentioned in the passage. Finally, it splices the word graph and the knowledge subgraph to form a word-knowledge subgraph, which is then input to the transformer for pre-training.

4) Knowledge Supervised. Knowledge supervised KEPTMs select keywords as training data under the supervision of KG. Then its semantic representation is learned by original PLMs. KEPLER [27] integrates knowledge information into a pre-trained model while using PLM to generate text-enhanced knowledge representations. However, pretraining from scratch is expensive. KG-BART[28] model is the first to integrate the knowledge graph into PLMs and improve the common sense reasoning ability in the generation process. However, when integrating knowledge graphs into PLMs, issues such as spatial alignment and knowledge noise need to be considered.

Existing pre-trained language models that introduce external knowledge to enhance pre-trained language models utilize the passage context. That is, to retrieve knowledge subgraphs by identifying entities in the

text. However, for MRC tasks, they do not consider that the information in the knowledge subgraph matches the question and the passage. In contrast, KG-ATT-MRC model fully integrates the semantic information among passages, questions, and knowledge by embedding the knowledge into the passage.

3. ATSMRC DATASET

This section explains the construction process of ATSMRC dataset before presenting the summary and examples of the dataset. The capability of the dataset is compared with the Chinese extractive MRC dataset cmrc2018.

3.1 ATSMRC Dataset Construction

3.1.1 Data Collection and Cleaning

We use "terrorist attack", "ramming and crushing", "shooting incident", "suicide attack", "bomb attack", and "extremist group" as keywords to get the news content. Then we clean the news in three steps: (1) Remove text spaces and line breaks. (2) Remove long text (>1000 characters). (3) Remove duplicate texts.

3.1.2 Text Analysis and Rule Formulation

We have summarized the features of anti-terrorism news based on experts' opinions. (1) The news content contains a lot of information about anti-terrorism entities. (2) Anti-terrorism news covers a wide range of topics, such as terrorist attacks, terrorism caused by regional conflicts, terrorism caused by social problems, and official statements.

We developed the annotation rules for extracted question-answer pairs based on anti-terrorism news texts' characteristics. (1) Ask as many questions as possible about the time and place of the terrorist event, the method of attack, the casualties, the terrorist organization, and other relevant circumstances. (2) When asking questions about people, they must be about people related to the terrorist event, e.g., information about the attackers or the victims.

3.1.3 Manual Annotation

We generate Q&A pairs for anti-terrorism news texts based on the annotation rules. There are ten rounds of annotation, and each round has two groups of two annotators. After each round, members of the group check with each other whether the question-answer pairs comply with the annotation rules. Mark the unqualified question-answer pairs, indicate the reasons and hand them over to the next group for re-marking in the next round (the third group will be handed over to the first group for re-marking). At the end of the labeling work, the final round of unqualified question-and-answer pair proofreading is completed. The specific labeling process for each round is shown in Figure 3.

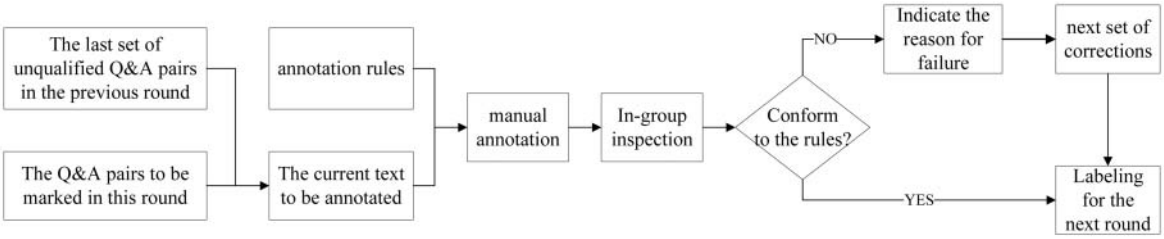

Figure 3. Question-and-answer pair single-round labeling process.

3.1.4 ATSMRC Dataset Summary

The data on ATSMRC dataset comes from China Daily, Sohu.com, and CCTV.com. It has 2,218 new passages and 9900 Q&A pairs. All of the questions and answers are generated by hand. Examples of their datasets are shown in Figure 4.

[Passage]
At 7 p.m. on June 20, a knife attack took place in Reading Faubury Garden, England. At present, 3 people have died, and more than 3 people have been sent to the hospital for treatment. British police declared the attack a terrorist attack. According to "Russia Today" (RT), a few hours before the case occurred, some demonstrators held a "black life is life" protest in the region. However, according to the police, the protest ended about two hours before the case occurred. At present, "there is no sign that this (attack) event is related to the 'black life is also life' protest held in Reading today". According to the report, a 25-year-old local man was arrested at the scene and is now facing murder charges. (715 characters in full)

Q1: What kind of casualties were caused by the knife attack in Reading Faubury Garden, England?
A1: 3 people have died, and more than 3 people have been sent to the hospital for treatment. answer_start: 97
Q2: A few hours before the case, what activities were held by demonstrators in the area?
A2: black life is life answer_start: 340
Q3: A 25-year-old local man was arrested at the scene. What charges are he facing?
A3: murder charges answer_start: 701

Figure 4. Example of ATSMRC dataset. The answer start shows where in the original passage the answer begins.

News on ATSMRC dataset can be divided into three categories by event distribution: terrorist attacks, military conflicts, and non-terrorist attacks. The distribution of specific news text types is shown in Figure 5. As can be seen from Figure 5, news of terrorist attacks and military conflicts accounted for more than three quarters. Non-terrorist attacks include regional conflicts, conflicts caused by social issues, official statements, etc. Therefore, ATSMRC dataset is rich in news topics and diverse in variety, which can provide sufficient downstream task corpora for the extractive model. It avoids the poor generalization ability of the model caused by the single news content.

The organizations that appear more frequently in the Q&A pairs of the dataset are the Islamic State, Kurdish forces, Al-Qaeda, and the Taliban. The distribution of Q&A pairs for the above four organizations is shown in Figure 6. The volume count has shown that there are a lot of Q&A pairs about ISIS and the Taliban, which helps the operations staff figure out how dangerous these groups are. There are 5,483

question-answer pairs about these four major terrorist organizations, accounting for more than half of the total number of question-answer pairs. This shows that the terrorist organizations in Figure 6 are highly dangerous and rich in information, which can provide the basis for a risk or danger level as-sessment for anti-terrorism experts.

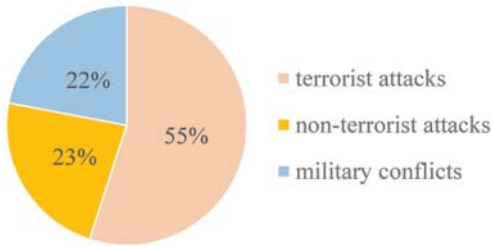

Figure 5. Distribution of ATSMRC news types.

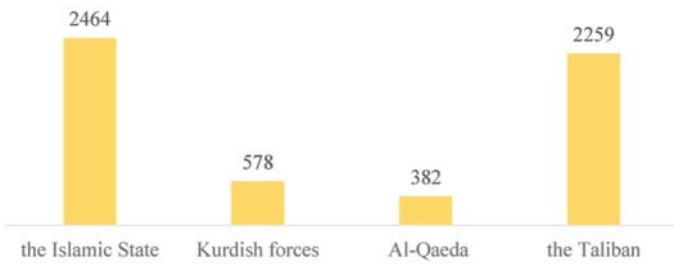

Figure 6. Distribution of Q&A Pairs for Major Terrorist Groups.

We compared the number of paragraphs and Q&A pairs on ATSMRC dataset and cmrc2018 dataset, as shown in Table 2.

Table 2. Comparison of cmrc2018 and ATSMRC paragraph and Q&A pair counts.

Dataset	Dataset type	Number of pararaphs	Number of Q&A pairs
cmrc2018	train set	2410	10132
	test set	848	3219
ATSMRC	train set	1818	8136
	dev set	300	1312
	test set	100	452

Since cmrc2018 validation set is not publicly available, it is not listed in table 3. ATSMRC dataset has fewer paragraphs and Q&A pairs than cmrc2018 dataset, but they are much more focused on domain-specific tasks.

4. KG-ATT-MRC MODEL

The overall framework of the proposed KG-ATT-MRC model is aligned with the architecture in Figure 1, and the detailed structure of the model is shown in Figure 7. It consists of four main parts: BERT Encoding layer, Knowledge Selection layer, Attention layer, and Output layer. The BERT encoding layer has the same encoding structure as BERT[6]. It takes question-answer pairs as input to the model and finally computes a context-aware vector representation. The knowledge selection layer selects relevant knowledge triples through entity links to the Chinese knowledge graph. And select the appropriate number of triples for knowledge injection through experiments. The Attention layer takes questions, passages, and knowledge vectors as input and calculates the mixed mutual attention of the three through the attention mechanism. The Output layer predicts the start and end positions of answers and outputs them. These four parts are described in detail below.

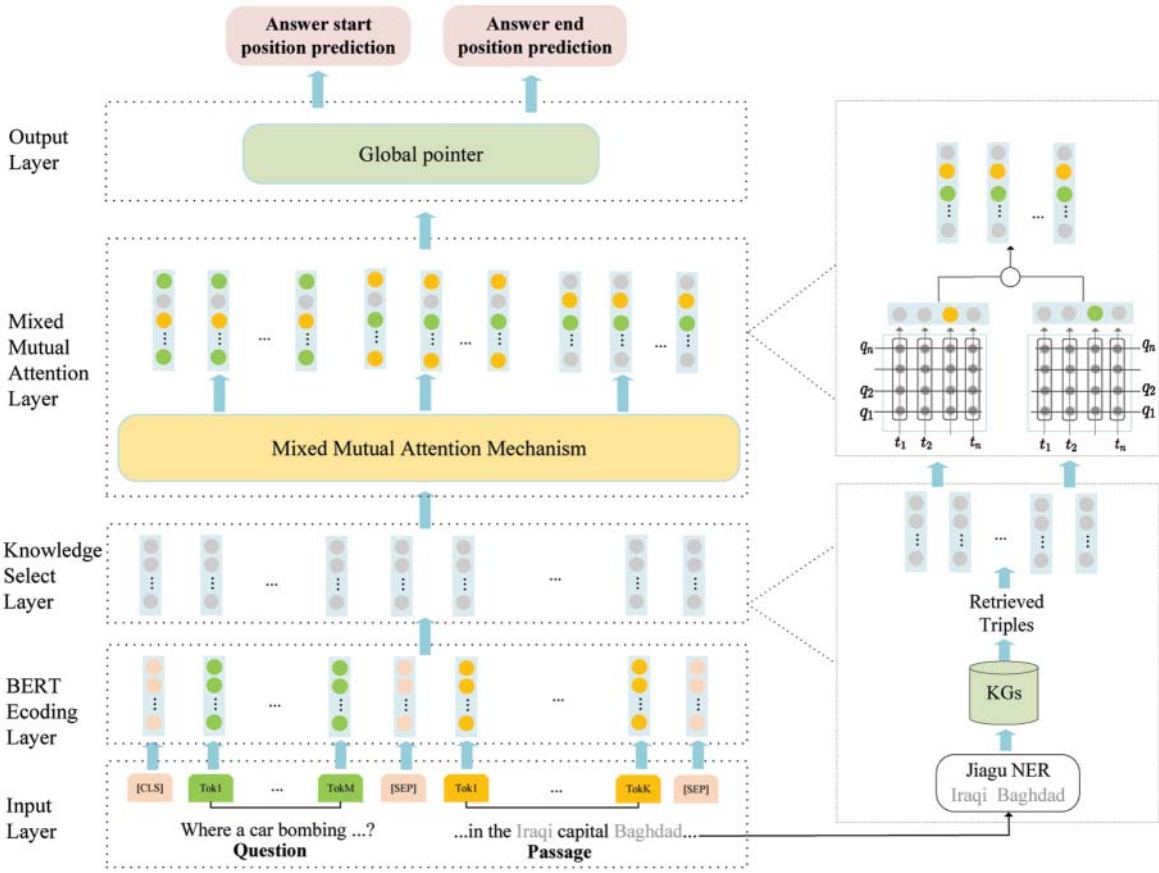

Figure 7. The overall architecture of KG-ATT-MRC.

4.1 BERT Encoding Layer

This layer uses BERT base encoder to model the passages and the questions. It takes the passage p and the question q as input and computes a context-aware representation for each token. where h_i is the sum of the token vector representation, the segment vector representation, and the position vector representation for each word.

$$\boldsymbol{h}_i^l = Transformer(h_i^{l-1}), l = 1, 2, \cdots, L \tag{1}$$

We use the final hidden layer states $\{\boldsymbol{h}_i^L\}_{i=1}^{k+m+3}$ as the output of BERT Encoding layer.

We denote the set of injected triples T_s, $t_i \in \mathrm{T}_s$ is a knowledge triple with head(s_i), tail(o_i) and relation(p_i), we splice all $t_i \in \mathrm{T}_s$ to create a knowledge sequence T. The format of T is described as follows:

$$T = [< CLS >, t_1, < SEP >, t_2, < SEP >, \ldots, t_i, < SEP >], i = 4 \tag{2}$$

where <*CLS*> is a specific classifier token and <*SEP*> is a sentence separator which are defined in BERT. Here we use <*SEP*> to split knowledge triples. The length of the sequence is determined by the number of injected triples (i=4 in experiments in Section 5.3.2).

Here is a example for the knowledge sequence T: [<*CLS*>, Baghdadi, belongs to, ISIS extremist group, <*SEP*>, US army shoot to Baghdadi, ..., <*SEP*>].

The knowledge sequence T is encoded by BERT-Encoder to obtain its knowledge representation vector matrix **T**, as follows:

$$\boldsymbol{T} = (\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n) = BERT - Encoder(T) \tag{3}$$

Question $q = (q_1, q_2, \cdots, q_m)$, $\{q_j\}_{j=1}^m$ represents each token in question q. BERT-Encoder encodes to obtain its question representation vector-matrix **Q**, as follows:

$$\boldsymbol{Q} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_m) = BERT - Encoder(q) \tag{4}$$

Passage $p = (p_1, p_2, \cdots, p_k)$, $\{p_i\}_{i=1}^k$ represents each token in passage p, which is encoded by BERT-Encoder to obtain its text representation vector matrix **Q**, as follows:

$$\boldsymbol{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_k) = BERT - Encoder(p) \tag{5}$$

4.2 Knowledge Selection and Mixed Mutual Attention

The Knowledge Selection layer and Mixed Mutual Attention are the core layers of KG-ATT-MRC model. It focuses on the extraction of knowledge triple and the implementation of the attention mechanism. As shown in Figure 8, this paper computes the 3-way mutual attention among passages, questions, and knowledge triples. The mutual attention between passage and question has been done by Minjoon Seo [29] et al., and will not be elaborated in this paper.

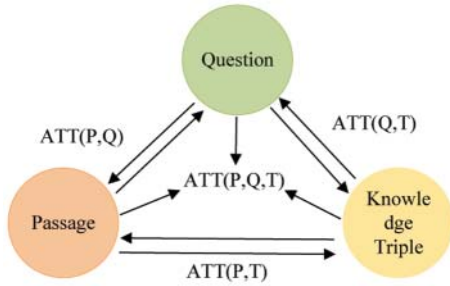

Figure 8. 3-way mutual attention among passages, questions, and knowledge triples. ATT (P, Q) indicates the mutual attention between passages and questions, and others are similar. ATT (P, Q, C) represents the mixed attention among passages, questions, and knowledge.

We extract the more useful entity information from the anti-terrorism news passages. Here we compare several domestic and foreign deep learning natural language processing tools, combine the typical characteristics of the entities to be extracted, and choose the jiagu[①] as the named entity recognition tool for the anti-terrorism domain text. For each news passage, we extract 3 to 5 entities of information. Then, according to the API interface provided by CN-DBpedia, we can get the knowledge triples returned by each entity. Considering that too much knowledge fusion may change the meaning of the original news text and degrade the model performance. While less knowledge fusion may not achieve the purpose of knowledge enhancement. Therefore, for each entity, the number interval of selected triples is located in [1, 7]. Then, experiments are conducted on ATSMRC dataset to select the number of triples when KG-ATT-MRC model performs best in section 5.

Mutual Attention Between Questions and Knowledge Triples. We compute attention in two directions: the attention of the knowledge triples to the questions (abbreviated as T2Q) and the attention of the questions to the knowledge triples (Q2T). T2Q and Q2T share a relevance matrix [29] $\mathbf{S}$, $\mathbf{S} \in R^{n*m}$. It represents the similarity between the nth character in the triple and the mth character in the question.

Knowledge Triples' Attention to Questions (T2Q). T2Q is used to compute which words in the question are most relevant for each word in the knowledge triple.

Incoming messages from the knowledge triples are aggregated via attention mechanism [2], the question representation vector that incorporates the knowledge information is:

$$\tilde{\boldsymbol{Q}} = \sum_{m} soft\max(\boldsymbol{S}_{nm})\boldsymbol{Q}_{,m} \tag{6}$$

$\tilde{\boldsymbol{Q}}$ is the question vector that contains the knowledge triple.

① https://github.com/ownthink/Jiagu

Questions' Attention to Knowledge Triples (Q2T). Q2T is used to calculate which words in the knowledge triple are most similar to the question. We perform a softmax to determine the attention weight of each word in the knowledge triple. And then a weighted average of the vectors for each column in the similarity matrix. Which is calculated as:

$$\boldsymbol{b} = \frac{\sum_{i=1}^{m} softmax(\boldsymbol{S}_i)}{m} \tag{7}$$

The knowledge triple vector can then be calculated as:

$$\tilde{\boldsymbol{T}} = \sum_{n} \boldsymbol{b}_n \boldsymbol{T}_{:n} \tag{8}$$

After obtaining the attention in both directions, we obtain the fused matrix-vector **G** by matrix multiplication of the question vector and the knowledge triple vector, where ⊙ stands for matrix multiplication.

$$\boldsymbol{G} = [\tilde{\boldsymbol{Q}} \odot \tilde{\boldsymbol{T}}] \tag{9}$$

Mutual Attention Between Passages and Knowledge Triples. The mutual attention between passages and knowledge triples is calculated in the same way as the mutual attention between questions and knowledge triples. We first calculate the correlation matrix of the passage feature representation vector and the knowledge triple feature representation vector. Then calculate the attention of passage to knowledge triple and the attention of knowledge triple to passage. Lastly, the attention of both is multiplied by a matrix to get the combined matrix-vector **G**.

Mixed Mutual Attention Among Passages, Questions, and Knowledge Triples. This paper also computes the mixed attention among passages, questions, and knowledge triples to embed knowledge triples and fully integrate the semantic information between them. First, we compute the semantic vector of passage that incorporates both question and knowledge triple information. As shown in Figure 9, we compute attention in two directions: triple-to-passage attention and question-to-passage attention.

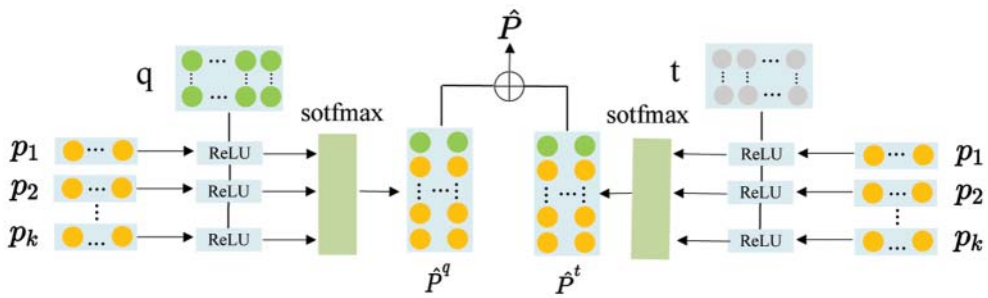

Figure 9. Triple-to-passage attention and question-to-passage attention.

We use the ReLU function to calculate the relevance score $\mathbf{S}_{ij}$ of the ith vector in the news passage to each vector in the knowledge as:

$$\boldsymbol{S}_{ij} = (ReLU(W_1 \boldsymbol{P}_{:i}))^T ReLU(W_2 \boldsymbol{T}_{:j}) \tag{10}$$

Where W_1 and W_2 are trainable parameters. The information from the knowledge triple is

$$\alpha_{ij} = \frac{\exp(\boldsymbol{S}_{ij})}{\sum_j \exp(\boldsymbol{S}_{ij})} \tag{11}$$

$$\boldsymbol{P}_i = \sum_j \alpha_{ij} \boldsymbol{T}_{:j} \tag{12}$$

The α_{ij} in the above equation is the attention weight of each word in the knowledge triple to the news passage, and $\mathbf{P}_i^t$ is the vector of the ith word in the passage incorporating the knowledge triple information, and the vector of news passage incorporating the knowledge triple can be obtained by concatenating $\mathbf{P}_i^t$ as:

$$\hat{\boldsymbol{P}}^t = [\boldsymbol{P}_1^t, \boldsymbol{P}_2^t, \ldots, \boldsymbol{P}_k^t] \tag{13}$$

A passage vector representation incorporating question information can be computed using the same method as above:

$$\hat{\boldsymbol{P}}^q = [\boldsymbol{P}_1^q, \boldsymbol{P}_2^q, \ldots, \boldsymbol{P}_k^q] \tag{14}$$

Matrix summing of $\hat{\boldsymbol{P}}^t$ and $\hat{\boldsymbol{P}}^q$ yields the final passage vector representation $\hat{\boldsymbol{P}}$, where $\oplus$ stands for vector summation.

$$\hat{\boldsymbol{P}} = [\boldsymbol{P}^t \oplus \boldsymbol{P}^q] \tag{15}$$

Using the same method, we can compute the knowledge triple vector representation $\hat{\boldsymbol{T}}$ that incorporates the question and passage information, and the question vector representation $\hat{\boldsymbol{Q}}$ incorporates the passage and knowledge triple information. The three vectors are matrix multiplied to obtain the fused matrix-vector $\mathbf{G}$.

$$\boldsymbol{G} = [\hat{\boldsymbol{P}} \odot \hat{\boldsymbol{Q}} \odot \boldsymbol{T}] \tag{16}$$

4.3 Output

The output layer predicts the start and end positions of the answer interval. It is a stitching of the matrix-vector $\mathbf{G}$ obtained from the previous layer, the fusion vector representation of text and question $\{\mathbf{h}_i^L\}_{i=1}^{k+m+3}$ obtained from the first layer. Then obtain the fusion vector $\mathbf{U}$ of text, question, and knowledge as follows:

$$\boldsymbol{U} = [\boldsymbol{G}; \{\boldsymbol{h}_i^L\}_{i=1}^{k+m+3}] \tag{17}$$

The output layer in this paper uses GlobalPointer[2] as the output structure. The loss function it uses is shown below.

$$p(i,j)=\frac{e^{s(i,j)}}{\sum_{i\le j}e^{s(i,j)}} \tag{18}$$

$$-\sum_{i\le j}p(i,j)f_1(i,j)+\lambda\sum_{i\le j}p(i,j)\log p(i,j) \tag{19}$$

where $f_1(i, j)$ is the F1 similarity between the predicted fragment and the standard answer, and λ is the hyperparameter. $s(i, j)$ is the scoring as a continuous fragment formula from i to j as the answer.

5. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we compare KG-ATT-MRC (CQT) model[3], which uses mixed mutual attention among passages, questions, and knowledge triples, with the better performing MRC models. At the same time, to verify the generalization ability of KG-ATT-MRC model, this paper also conducts training and testing on cmrc2018, yiqing, and webqa datasets. Finally, KG-ATT-MRC model is subjected to ablation experiments on ATSMRC dataset[4] and cmrc2018 dataset.

5.1 Experimental Settings

The dataset used in this paper is our own constructed ATSMRC dataset in section 3. It is randomly divided into a train set, validation set, and test set in the ratio of 6:2:2. The model parameter settings are shown in Table 3.

Table 3. Parameter settings.

Parameter	Value
Batch size	16
Max sentence length	512
Learning rate	0.0005
Epochs	20
Dropout	0.1
Vocab size	21128
Hidden size	768
Number of triples	4

[2] https://kexue.fm/archives/8373

[3] Our code will be available at https://github.com/houjin0803/KG-ATT-MRC

[4] The dataset will be available at https://github.com/houjin0803/ATSMRC

5.2 Experimental Results and Analysis

In this subsection, comparative and ablation experiments are conducted for KG-ATT-MRC model to verify its performance.

Comparison Experiments. We fine-tune KG-ATT-MRC model proposed in this paper with various effective PLMs on extractive reading comprehension, including both vanilla PLMs and knowledge-enhanced PLMs. At the same time, F1 and EM indicators are used to evaluate the model performance comprehensively. F1 metric is the summed average of word accuracy and recall, i.e., $F1 = 2 \times \frac{precision \times recall}{precision + recall}$, precision stands for accuracy, which refers to the percentage of words in the standard answers given in the model; recall refers to the percentage of words in the standard answers given in the model. EM stands for an exact match, if the answer given by the model is exactly the same as the standard answer, the score is 1, otherwise, it is 0. The comparative experimental results are shown in Table 4.

Table 4. Comparison of experimental results.

Model Types	Models/Datasets	ATSMRC		cmrc2018		yiqing		webqa	
		EM	F1	EM	F1	EM	F1	EM	F1
Pre-Trained Language Models	UER[30]	67.72	87.61	59.27	82.84	26.87	66.34	71.14	81.15
	CMRC2018[18]	66.47	84.56	51.66	75.49	17.98	40.89	71.27	81.34
	macbert[31]	66.78	86.24	51.73	75.78	20.18	44.76	72.17	82.69
	KT-NET[10]	67.07	85.63	59.97	82.22	34.37	67.01	70.54	80.42
Knowledge Enhanced Pre-Trained Language Models	Dual BERT[32]	68.92	86.47	63.29	84.32	35.68	67.57	71.53	81.76
	KMQA[13]	68.73	85.97	63.86	84.19	35.49	66.58	71.36	81.49
	CoLake[26]	69.38	86.61	63.94	84.59	35.76	67.32	72.25	83.04
	JointLK[24]	70.59	87.78	**64.57**	**84.98**	35.94	67.21	72.31	82.35
	ERNIE[12]	69.89	86.88	63.45	84.65	35.67	67.08	71.04	81.23
	CokeBERT[23]	70.68	87.58	63.74	84.91	34.35	67.98	**73.54**	**83.11**
	KG-ATT-MRC (CQT)	**70.70**	**87.91**	64.51	84.97	**36.46**	**68.26**	71.24	81.34

As shown in Table 4, the performance of KG-ATT-MRC (CQT) models is consistently better than the corresponding PLMs such as UER, CMRC2018, and Macbert on four extractive datasets. This shows that introducing external knowledge can enhance the pre-trained model.

For knowledge enhanced PLMs, KG-ATT-MRC (CQT) outperforms those PLMs on ATSMRC and yiqing datasets. On ATSMRC dataset, the EM metrics increased by 0.02%-3.63%, and the F1 metrics increased by 0.13%–2.28%. On yiqing dataset, our model improves 0.52%–2.09% and 0.28%–1.68%. It directly demonstrates that the mixed mutual attention among passages, questions, and knowledge triples proposed in this paper can improve the model's comprehension of questions and passages, especially on vertical domain datasets.

For knowledge enhanced PLMs, KG-ATT-MRC (CQT) models have only comparable results with JointLK and CokeBERT on cmrc2018 and webqa datasets. This is because both JointLK and CokeBERT use graph embeddings, which contain more knowledge. And cmrc2018 and webqa belong to the general field of datasets, the characteristics of the data are not obvious, so more benefits can be obtained from the KEPLMs model embedded in the graph. On the contrary, ATSMRC and yiqing belong to the vertical field of datasets, where the data characteristics are more obvious, and more knowledge noise may be introduced after using graph embedding.

Overall, our model improves significantly on vertical datasets. In follow-up work, our model can be improved by using graph embedding.

5.3 Ablation Study

5.3.1 Effects of Mutual Attention

To explore the effect of mutual attention on model performance, we evaluate our model by removing the attention of different modules. The experimental results are shown in Table 5.

Table 5. Results of ablation study.

Models/Datasets	ATSMRC		cmrc2018	
	EM	F1	EM	F1
KG-ATT-MRC (CQT)	**70.70**	**87.91**	**64.51**	**84.97**
KG-ATT-MRC (QT)	69.83	87.31	56.52	80.93
KG-ATT-MRC (CT)	68.63	87.19	58.18	82.76
KG-ATT-MRC (w/o attention)	68.83	86.92	60.80	82.36

KG-ATT-MRC (QT) indicates that the third layer of the model is implemented by computing the attention between questions and knowledge triples. KG-ATT-MRC (CT) indicates that the third layer of the model is implemented by computing the attention between passages and knowledge triples. From Table 5, KG-ATT-MRC (CQT) outperforms KG-ATT-MRC (QT) on EM and F1 average metrics by 0.74%–5.93%. It proves that the mutual attention between passages and questions and the mutual attention betweem passages and knowledge play an important role for the model to understand the semantics of the passage. KG-ATT-MRC (CQT) outperforms KG-ATT-MRC (CT) on EM and F1 average metrics by 1.4%–3.91%. However, compared to KG-ATT-MRC(QT) model, its impact on the model performance is greater on ATSMRC dataset. This demonstrates that computing the mutual attention between passages and knowledge improves the model's vertical domain performance more.

In KG-ATT-MRC (CQT), there is an essential mechanism: attention. It takes responsibility for weighing how much knowledge matches the passage and question. To further demonstrate the effect of the attention mechanism, we remove it. From Table 5, we can find that KG-ATT-MRC (CQT) outperforms KG-ATT-MRC (w/o attention), indicating the effectiveness of our attention mechanism.

5.3.2 Effects of Knowledge

In order to select the appropriate number of triples for knowledge fusion and reflect the influence of attention on eliminating knowledge noise, this paper conducts experiments on ATSMRC dataset by selecting one to seven triples for each entity. The results are shown in Figure 10 and Figure 11. w/ATT represents adding attention, and w/o ATT represents removing attention.

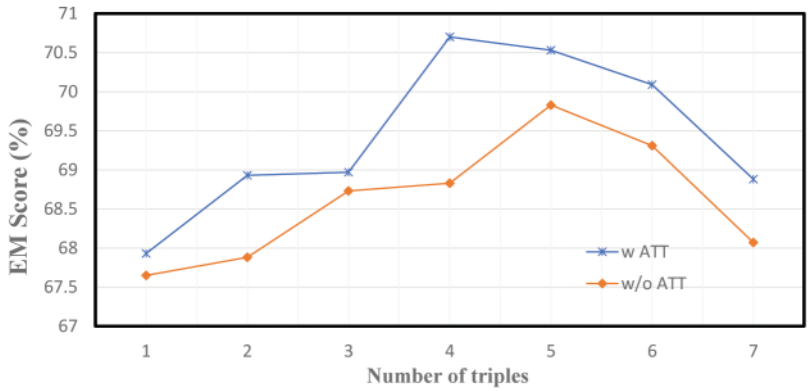

Figure 10. EM score with the number of knowledge triples on ATSMRC dataset.

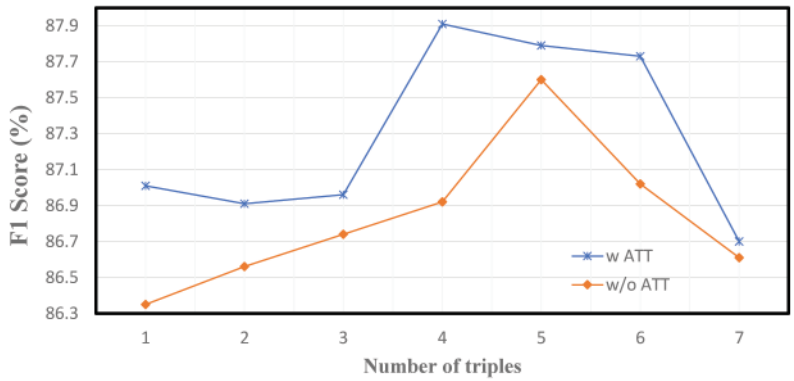

Figure 11. F1 score with the number of knowledge triples on ATSMRC dataset.

The results in Figure 10 and Figure 11 show that KG-ATT-MRC(w AAT) model performs best when the number of injected triples is equal to 4. Therefore, for each entity, four triples are selected for fusion in this paper. When the number of triples is less than 5, the performance of KG-ATT-MRC (w/o ATT) model increases gradually with the increase in the number of triples. This indicates that the injection of knowledge can enrich the semantic information of the sentences and improve the performance of the model. However, when the number of triples is greater than 5, the performance of the model gradually decreases. This indicates that too much knowledge injection can lead to an increase in useless information, which distorts the original semantic information of the sentences and appears as knowledge noise. Compared with KG-ATT-MRC (w ATT), its model performs better overall. This shows that the attention mechanism can make

the model focus more on useful knowledge and reduce knowledge noise. However, when the number of injected triples is 7, the model performance degrades faster compared to KG-ATT-MRC (w/o ATT). The possible reason is that the attention mechanism makes the model focus too much on the parts it considers important and ignore the other information.

6. CONCLUSION

In this paper, we construct the Anti-Terrorism Domain Dataset (ATSMRC) for the MRC tasks on the anti-terrorism domain. Then, this paper proposes the knowledge-based KG-ATT-MRC model, which incorporates domain-related triples from a large-scale encyclopedic knowledge base to enhance the semantic information of sentences on the anti-terrorism corpus. This paper computes the mixed mutual attention among passages, questions, and knowledge triples to reduce the knowledge noise problem caused by the introduction of knowledge. KG-ATT-MRC (CQT) model improved EM metrics by 1.32%–4.23% and F1 metrics by 1.3%–3.35% on ATSMRC dataset. EM metrics improved by 0.57%–12.85%, and F1 metrics improved by 0.38%–9.48% on cmrc2018 dataset, which is a very significant improvement. By comparing the experimental results, we obtained that KG-ATT-MRC model has the following characteristics:

(1) KG-ATT-MRC (CQT) model performance on the vertical domain dataset is better than that on the general domain dataset. The possible reason for this is that the data features of the vertical domain dataset are more pronounced, and our model can learn such features faster. Hence, the model performs better when the features of the general domain dataset are more fragmented, and the model cannot learn well. This is where the model proposed in this paper needs to be improved.
(2) KG-ATT-MRC (CQT) model performs better when the text in the dataset involves more specialized terms and entity information.

In future work, we will continue to expand the data volume of ATSMRC, increase the difficulty of ATSMRC question-answer pairs and add some unanswerable types of questions. At the same time, in terms of the model, we will optimize the network structure of the model and the algorithm of candidate answers predicted by the model further to improve the model performance and generalization ability.

ACKNOWLEDGEMENTS

This paper is funded by: National key research and development program (2020AAA0108500), National Natural Science Foundation of China Project (No. U1836118) and Key Laboratory of Rich Media Digital Publishing, Content Organization and Knowledge Service (No.: ZD2022-10/05).

AUTHOR CONTRIBUTIONS

F. Gao (feng.gao86@wust.edu.cn) and J. Hou (1371707917@qq.com) are responsible for the design of the research, the implementation of the approach, the analysis of the results, and the writing of the manuscript. J.G. Gu (simon@wust.edu.cn) has contributed to the design of the research, the analysis of the results, and the writing of the manuscript. L.H. Zhang (lhzhang@ecjtu.edu) revised the whole paper.

REFERENCES

[1] Therasa, M., Mathivanan, G.: Survey of Machine Reading Comprehension Models and its Evaluation Metrics. In: 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1006–1013. IEEE (2022, March)

[2] Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Advances in Neural Information Processing Systems 30(20) (2017)

[3] Du, H., Le, Z., Wang, H., et al.: COKG-QA: Multi-hop Question Answering over COVID-19 Knowledge Graphs. Data Intelligence 4(3), 471–492 (2022). doi: https://doi.org/10.1162/dint a 00154

[4] Van Nguyen, K., Van Huynh, T., Nguyen, D.V., et al.: New vietnamese corpus for machine reading comprehension of health news articles. Transactions on Asian and Low-Resource Language Information Processing 21(5), 1–28 (2022)

[5] Putri, R.A., Oh, A.: IDK-MRC: Unanswerable Questions for Indonesian Machine Reading Comprehension. arXiv preprint arXiv:2210.13778 (2022)

[6] Devlin, J., Chang M.W., Lee, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. https://arxiv.org/abs/1810.04805 (2018)

[7] Liu, Y., Ott, M., Goyal, N.: Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692. https://arxiv.org/abs/1907.11692 (2019)

[8] Joshi, M., Chen, D., Liu, Y.: Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, pp. 64–77 (2020)

[9] Lan, Z., Chen, M., Goodman, S.: Albert: A lite bert for self-supervised learning of language representations. arXiv:1909.11942. https://arxiv.org/abs/1909.11942 (2019)

[10] Yang, A., Wang, Q., Liu, J., et al.: Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2346–2357 (2019)

[11] Mavi, V., Jangra, A., Jatowt, A.: A Survey on Multi-hop Question Answering and Generation. arXiv preprint arXiv:2204.09140 (2022)

[12] Zhang, Z., Han, X., Liu, Z., et al.: ERNIE: Enhanced language representation with informative entities. arXiv:1905.07129. https://arxiv.org/abs/1905.07129 (2019)

[13] Li, D., Hu, B., Chen, Q., et al.: Towards medical machine reading comprehension with structural knowledge and plain text. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp. 1427–1438 (2020, November)

[14] Lapchaicharoenkit, T., Vateekeul, P.: Machine Reading Comprehension Using Multi-Passage BERT with Dice Loss on Thai Corpus. ECTI Transactions on Computer and Information Technology (ECTI-CIT) 16(2), 125–134 (2022)

[15] Rajpurkar, P., Zhang, J., Lopyrev K., et al.: Squad: 100,000+ questions for machine comprehension of text. arXiv:1606.05250. https://arxiv.org/abs/1606.05250 (2016)

[16] Wei, H., Kai, L., Haifeng, W.: DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications. arXiv:1711.05073. https://arxiv.org/abs/1711.05073 (2017)
[17] Karl Moritz, H., Tomas, K., Edward, G., et al.: Teaching Machines to Read and Comprehend. Advances in Neural Information Processing Systems (2015)
[18] Cui, Y., Liu, T., Che, W., et al.: A span-extraction dataset for Chinese machine reading comprehension. arXiv:1810.07366. https://arxiv.org/abs/1810.07366 (2018)
[19] Sun, Y., Wang, S., Li, Y., et al.: Ernie 2.0: A continual pre-training framework for language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence 34(5), 8968–8975 (2020)
[20] Yang, J., Xiao, G., Shen, Y., et al.: A survey of knowledge enhanced pre-trained models. arXiv preprint arXiv:2110.00269 (2021)
[21] Bai, J., Wang, Y., Chen, Y., et al.: Syntaxbert: Improving pre-trained transformers with syntax trees. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 3011–3020 (2021)
[22] He, B., Zhou, D., Xiao, J., et al.: Integrating graph contextualized knowledge into pretrained language models. arXiv preprint arXiv:1912.00147. https://arxiv.org/abs/1912.00147 (2019)
[23] Su, Y., Han, X., Zhang, Z., et al.: Cokebert: Contextual knowledge selection and embedding towards enhanced pre-trained language models. AI Open 2, 127–134 (2021)
[24] Sun, Y., Shi, Q., Qi, L., et al.: JointLK: Joint Reasoning with Language Models and Knowledge Graphs for Commonsense Question Answering. In: NAACL-HLT (2022)
[25] Liu, W., Zhou, P., Zhao, Z., et al.: K-bert: Enabling language representation with knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence 34(3), 2901–2908 (2020)
[26] Sun, T., Shao, Y., Qiu, X., et al.: Colake: Contextualized language and knowledge embedding. arXiv:2010.00309. https://arxiv.org/abs/2010.00309 (2020)
[27] Wang, X., Gao, T., Zhu, Z., et al.: Kepler: A unified model for knowledge embedding and pretrained language representation. Transactions of the Association for Computational Linguistics 9, 176–194 (2021)
[28] Liu, Y., Wan, Y., He, L., et al.: Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence 35(7), 6418–6425 (2021)
[29] Seo, M., Kembhavi, A., Farhadi, A., et al.: Bidirectional attention flow for machine comprehension. arXiv:1611.01603. https://arxiv.org/abs/1611.01603 (2016)
[30] Zhao, Z., Chen, H., Zhang, J., et al.: UER: An open-source toolkit for pre-training models. arXiv preprint arXiv:1909.05658. https://arxiv.org/abs/1909.05658 (2019)
[31] Cui, Y., Che, W., Liu, T., et al.: Revisiting pre-trained models for Chinese natural language processing. arXiv:2004.13922. https://arxiv.org/abs/2004.13922 (2020)
[32] Cui, Y., Che, W., Liu, T., et al.: Cross-lingual machine reading comprehension. arXiv:1909.00361. https://arxiv.org/abs/1909.00361 (2019)

AUTHOR BIOGRAPHY

Feng Gao (1986–), male, from Wuhan, Hubei, lecturer, master supervisor, doctoral degree, the main research directions are knowledge graph, semantic web, E-mail: feng.gao86@wust.edu.cn

Jin Hou (1998–), male, from Wuhan, Hubei, master superior degree, the main research directions are natural language processing, knowledge graph, E-mail: 1371707917@qq.com

Jinguang Gu (1974–), male, from Xiantao, Hubei, professor, doctoral supervisor, doctoral degree, the main research direction is distributed computing, knowledge graph, E-mail: simon@wust.edu.cn

Lihua Zhang (1972–), male, from Jingshan, Hubei, associate professor, doctoral degree, the main research direction are electrical information technology, industrial control network information security, E-mail: lhzhang@ecjtu.edu.cn