*Electronic Research Archive*

http://www.aimspress.com/journal/era

ERA, 31(10): 6012–6026.
DOI: 10.3934/era.2023306
Received: 05 June 2023
Revised: 23 August 2023
Accepted: 27 August 2023
Published: 06 September 2023

*Research article*

# Word-level dual channel with multi-head semantic attention interaction for community question answering

**Jinmeng Wu [1], HanYu Hong[1], YaoZong Zhang[1,*], YanBin Hao[2], Lei Ma[1] and Lei Wang[1]**

[1] School of Electrical and Information Engineering, Wuhan Institute of Technology, Wuhan 430205, China

[2] School of Information Science and Technology, University of Science and Technology of China, Anhui 230026, China

* **Correspondence:** Email: zhangyaozong@wit.edu.cn.

**Abstract:** The semantic matching problem detects whether the candidate text is related to a specific input text. Basic text matching adopts the method of statistical vocabulary information without considering semantic relevance. Methods based on Convolutional neural networks (CNN) and Recurrent networks (RNN) provide a more optimized structure that can merge the information in the entire sentence into a single sentence-level representation. However, these representations are often not suitable for sentence interactive learning. We design a multi-dimensional semantic interactive learning model based on the mechanism of multiple written heads in the transformer architecture, which not only considers the correlation and position information between different word levels but also further maps the representation of the sentence to the interactive three-dimensional space, so as to solve the problem and the answer can select the best word-level matching pair, respectively. Experimentally, the algorithm in this paper was tested on Yahoo! and StackEx open-domain datasets. The results show that the performance of the method proposed in this paper is superior to the previous CNN/RNN and BERT-based methods.

**Keywords:** community question answering; interactive learning; multi-head attention; transformer; recurrent bidirectional neural network

## 1. Introduction

The task of Community Question Answering (cQA) is to enable computers to select highly relevant answers from a pool of candidate sentences given a question posted by a user in natural language. This task has wide applications such as information retrieval [1, 2], web search ranking, and dialogue system [3–5]. In order to compute an accurate relatedness measure, it is crucial to consider the syntactic,

lexical, and semantic information of text pairs. Among them, semantic information is a technical means to test the interactive relationship between texts. In actual dialogues, the meaning of sentences expressed by humans does not exist in a completely independent form. In the process of analyzing sentences, it is often necessary to conduct dialogues in combination with the context. The interactive semantic representation of enriched sentences can make the learned text-distributed feature representation more informative and thus enhance the learned similarity effect. Therefore, interactive relationship learning becomes a challenging problem for solving community question answering tasks.

To achieve the open-domain question-answer matching task, previous studies have proposed a feature-based model [6–9] utilizing semantic information provided by external resources such as WordNet. Although these methods use semantic features to improve similarity measures, they need to rely on extensive manual feature engineering, and the feature extraction stage will incur expensive cost loss and time loss. In recent years, deep learning methods have been widely used in question answering tasks [10–13] by using neural language models instead of manual feature engineering. The general strategy is based on the model of a Recurrent neural network (RNN) or Convolution neural network (CNN). The previous state-of-the-art [14] proposes the Compare-aggregation method, which uses a CNN model to capture the contextual relationship representation between the words in the question and the answer sentences one by one. Recently, the BERT model based Transformer has been widely used in question answering tasks, and has achieved very good results [15–19]. Transformer uses its multi-head mechanism to learn the self-attention representation of a single sentence to achieve the effect of paying more attention to learning similarity representation but ignores the order between words or the interaction information between sentences in the text matching process.

Aiming at the above-mentioned problems in the interactive learning process, we propose a dual channel with multi-head interaction (DMI) method based on multi-dimensional word-level interactive learning. First, for the issue of not considering the interaction of word sequences in the traditional Transformer, in order to introduce the correlation information between words, the method proposes the structure of the written head tuple by improving the linear transformation and using the recursive recurrent network combined with the word position, which obtains written multi-head tuple and sentence representations for the current moment. Second, a dual-channel interactive structure mapping is designed to obtain a multi-dimensional interactive representation, which can adaptively select the optimal word-level similarity measure for questions and answers. Therefore, the contributions of the DMI method proposed in this paper include the following points:

1) It proposes a dual-channel multi-head interactive attention method for learning sentence-based interactive expressions based on word-level multi-dimensionality and enhancing semantic interoperability.

2) It designs an interactive multi-dimensional correlation structure, from the two perspectives of questions and candidate answers to strengthen the impact of word-level pairs with greater weight on similarity, filter unimportant word-level pairs, and reduce unnecessary losses.

3) This proposed method is tested on different community question answering datasets. Compared with some previous representative works, including RNN and CNN-based methods, BERT, etc., this algorithm has achieved more robust results.

We first introduce the semantic interactive information of open domain question answering in Section 1, and outlines the mainstream deep learning model methods in recent years; then describes the overall

structure and algorithm of the DMI method in this paper in Section 2; then Section 3 introduces the setup of the experiment, including data sets, preprocessing methods, and evaluation indicators; Section 4 is the experimental results and analysis, including the experimental results and example analysis of method comparison; finally, the full text is summarized in Section 5.

## 2. Materials and methods

### *2.1. Bidirectional RNN preliminary*

A typical approach involves initially calculating the representations from a potential response collection of a sentence that aligns with the supplied question and answer sentences. This could be in the form of vectors or matrices, derived from the word content of the question and answer sentences. Next, similarity scores (or confidence of relevance) between the question and the potential answers are determined based on their corresponding representations. The candidate with the top score is then chosen.

Consider sentence embedding is indicated as $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$, where $\boldsymbol{x}_t$ is the $t$-th word representation in the sentence. A bidirectional RNN-based language model acquires a vector representation to encapsulate the semantic and sequential information of the words in the sentence. The forward and backward propagation of hidden states are usually denoted as

$$\boldsymbol{h}_t^f = f(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}^f), \tag{2.1}$$

$$\boldsymbol{h}_t^b = f(\boldsymbol{x}_t, \boldsymbol{h}_{t+1}^b), \tag{2.2}$$

where the $t$-th word representation vector $\boldsymbol{x}_t$ relates to a forward hidden state at time step $t$. The hidden representation vector $\boldsymbol{h}_t = RNN(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}^f, \boldsymbol{h}_{t+1}^b)$ includes word context information accumulated up to the $t$-th word in the sentence. It is derived from the vector representation $\boldsymbol{x}_t$ of the current word and the previous accumulation $\boldsymbol{h}_{t-1}^f$ and $\boldsymbol{h}_{t+1}^b$ for forward and backward propagation, respectively. Different types of RNNs arise from various implementations of the activation function $f(\cdot)$. For example, a traditional RNN utilizes a standard linear operation with a sigmoid activation $\mathrm{sig}(\cdot)$ to handle the input $\boldsymbol{x}_t$ and $\boldsymbol{h}_{t-1}$. In contrast, a bidirectional LSTM employs a set of recurrent functions [20] which is defined as

$$\begin{aligned}
\boldsymbol{i}_t &= \mathrm{sig}\left(\mathbf{W}_{xi}\boldsymbol{x}_t + \mathbf{W}_{hi}[\boldsymbol{h}_{t-1}^f; \boldsymbol{h}_{t+1}^b] + \boldsymbol{b}_i\right), && (2.3)\\
\boldsymbol{f}_t &= \mathrm{sig}\left(\mathbf{W}_{xf}\boldsymbol{x}_t + \mathbf{W}_{hf}[\boldsymbol{h}_{t-1}^f; \boldsymbol{h}_{t+1}^b] + \boldsymbol{b}_f\right), && (2.4)\\
\boldsymbol{o}_t &= \mathrm{sig}\left(\mathbf{W}_{xo}\boldsymbol{x}_t + \mathbf{W}_{ho}[\boldsymbol{h}_{t-1}^f; \boldsymbol{h}_{t+1}^b] + \boldsymbol{b}_o\right), && (2.5)\\
\boldsymbol{g}_t &= \tanh\left(\mathbf{W}_{xc}\boldsymbol{x}_t + \mathbf{W}_{hc}[\boldsymbol{h}_{t-1}^f; \boldsymbol{h}_{t+1}^b] + \boldsymbol{b}_g\right), && (2.6)\\
\boldsymbol{c}_t &= \boldsymbol{f}_t \odot [\boldsymbol{c}_{t-1}^f; \boldsymbol{c}_{t+1}^b] + \boldsymbol{i}_t \odot \boldsymbol{g}_t, && (2.7)\\
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t), && (2.8)
\end{aligned}$$

where the symbol $\odot$ represents the Hadamard product. The word representation vector $\boldsymbol{x}_t$, along with distinct subscript symbols of the weight matrices $\mathbf{W}$ and the bias vectors $\boldsymbol{b}$ are the model parameters that need to be fine-tuned.

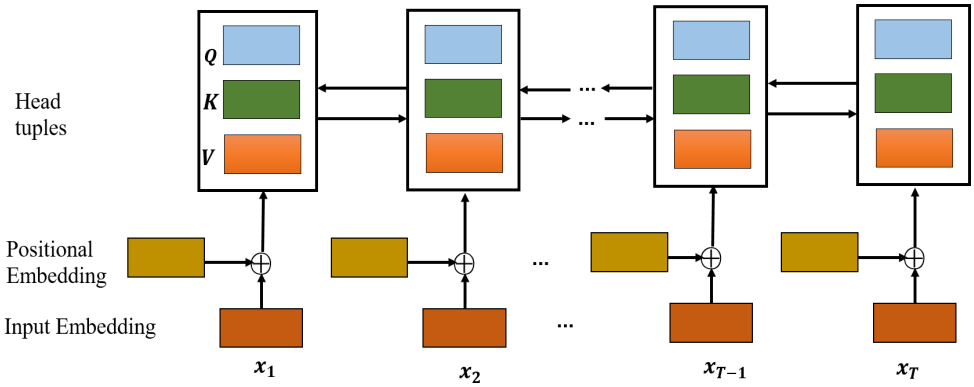


**Figure 1.** Bidirectional multi-head recursive recurrent network structure.

### *2.2. Bidirectional multi-head recurrent network*

Given a community question **q** and a candidate answer **a**, where the maximum length of the question and answer sentences is $m$ and $n$. We mainly explore the relevance of question-answer pairs depends on the semantic similarity between the words they contain. The word embedding of each sentence at time $t$ is expressed as a $d$-dimensional vector $\mathbf{x}_t$, $0 < t \leq T$ using distributed vector representation. In the traditional Transformer [15], the head tuple is obtained through simple linear transformation: query vector, key vector and value vector. However, such a mapping method is only for spatial projection of a single word, without taking into account the semantic correlation between words, and is not suitable for learning interactive sentence expressions. Therefore, we design the structure of the recursive network to sequentially write the word input to obtain the written head group. The architecture of the bidirectional multi-head recursive recurrent network is shown in Figure 1.

It can be seen from the figure that the input of each word into the recurrent network is composed of word embedding combined with position embedding expression, and the input into the network is expressed as:

$$\mathbf{\Phi}_t = \mathbf{x}_t \oplus PE_{pos}(\mathbf{x}_t), \tag{2.9}$$

where $PE_{pos}(\cdot)$ is the relative position information, which is calculated by the sine and cosine functions of different frequencies. Next, input $\mathbf{\Phi}_t$ and the tuple at time $t-1$ into the recurrent network for forward and reverse bidirectional transmission, and solve the tuple at time $t$. The expression formula of fresh bidirectional recurrent can be written as:

$$\begin{bmatrix} \mathbf{Q}_t^f \\ \mathbf{K}_t^f \\ \mathbf{V}_t^f \end{bmatrix} = \begin{bmatrix} \mathbf{W}_Q^f \\ \mathbf{W}_K^f \\ \mathbf{W}_V^f \end{bmatrix} \cdot \begin{bmatrix} \mathbf{\Phi}_t, \mathbf{W'}_Q^f \mathbf{Q}_{t-1}^f \\ \mathbf{\Phi}_t, \mathbf{W'}_K^f \mathbf{K}_{t-1}^f \\ \mathbf{\Phi}_t, \mathbf{W'}_V^f \mathbf{V}_{t-1}^f \end{bmatrix}, \tag{2.10}$$

$$\begin{bmatrix} \mathbf{Q}_t^b \\ \mathbf{K}_t^b \\ \mathbf{V}_t^b \end{bmatrix} = \begin{bmatrix} \mathbf{W}_Q^b \\ \mathbf{W}_K^b \\ \mathbf{W}_V^b \end{bmatrix} \cdot \begin{bmatrix} \mathbf{\Phi}_t, \mathbf{W'}_Q^b \mathbf{Q}_{t+1}^b \\ \mathbf{\Phi}_t, \mathbf{W'}_K^b \mathbf{K}_{t+1}^b \\ \mathbf{\Phi}_t, \mathbf{W'}_V^b \mathbf{V}_{t+1}^b \end{bmatrix}. \tag{2.11}$$

For forward and backward propagation, the weights $\mathbf{W}_Q$, $\mathbf{W}_K$ and $\mathbf{W}_V$ are matrices of $d_k \times d$ dimensions, corresponding to query matrix (**Q**), key matrix (**K**) and value matrix (**V**). The matrices $\mathbf{W'}_Q$, $\mathbf{W'}_K$ and $\mathbf{W'}_V$ are $d \times d_k$ dimensions used to map $t-1$ time written head tuples. The written head tuple is

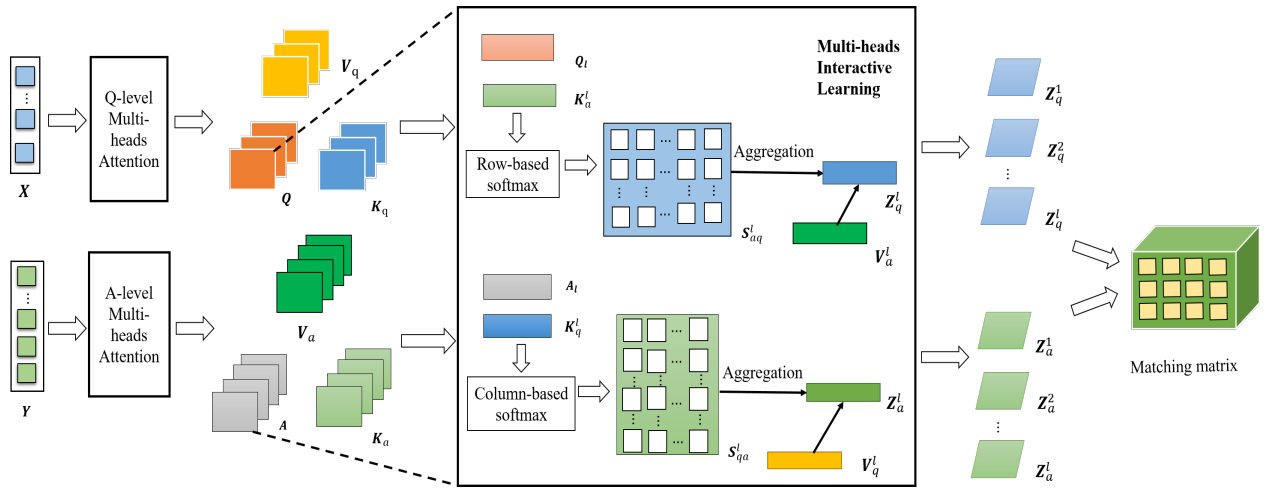


**Figure 2.** Dual-channel multi-head interaction (DMI) model structure.

composed of the results obtained by forward and reverse transmission through aggregation, and can be obtained by:

$$\begin{bmatrix} \mathbf{Q}_t \\ \mathbf{K}_t \\ \mathbf{V}_t \end{bmatrix} = LSTM\left(\mathbf{x}_t, \begin{bmatrix} \mathbf{Q}^f_{t-1}, \mathbf{Q}^b_{t+1} \\ \mathbf{K}^f_{t-1}, \mathbf{K}^b_{t+1} \\ \mathbf{V}^f_{t-1}, \mathbf{V}^b_{t+1} \end{bmatrix}\right). \tag{2.12}$$

In this work, each element of the tuple at time $t$ is a horizontal quantity of $2d_k$ dimension. For the representation of each sentence, the bidirectional muti-head RNN function $LSTM(\cdot)$ aggregates $\mathbf{Q}_t$, $\mathbf{K}_t$ and $\mathbf{V}_t$ of the $t$-th time into $2d_k \times T$ dimensions. The sentence length $T$ is equal to $m$ and $n$ for questions and candidate answers, respectively.

### *2.3. Word-level interactive expression based on dual-channel multiple written heads*

We get the written head tuples of the question $[\mathbf{Q}, \mathbf{K}_q, \mathbf{V}_q]$ and the written head tuple of the answer $[\mathbf{A}, \mathbf{K}_a, \mathbf{V}_a]$ based on the bidirectional multi-head recurrent neural network. We design an interactive attention representation learning method based on multiple written heads, which deeply integrates the word-level multi-dimensional representation of questions and answers. Two types of interactive matrices are obtained by substituting the query matrix of the question or answer and the key matrix of the answer or question using the dual-channel dot-scaled product. The interactive operations of the question and answer are expressed as:

$$\mathbf{S}^l_{a\to q} = \frac{\mathbf{Q}^T_l \cdot \mathbf{K}^l_a}{\sqrt{d_k}}, \tag{2.13}$$

$$\mathbf{S}^l_{q\to a} = \frac{\mathbf{A}^T_l \cdot \mathbf{K}^l_q}{\sqrt{d_k}}, \tag{2.14}$$

where $l$ is defined as the index value of the written head, and the total number of written heads is $L$. The interaction matrix $\mathbf{S}^l_{a\to q}$ of the answer to the question is the $m \times n$ matrix of the first $l$ written head, and the interaction matrix of the question to the answer $\mathbf{S}^l_{q\to a}$ is the $n \times m$ matrix of the $l$-th written head. After that, using softmax to regularize the word row vector corresponding to the row-based of $\mathbf{S}^l_{a\to q}$. Similarly, at the same time, columed-based word-level regularization is performed on $\mathbf{S}^l_{q\to a}$ to

determine the contribution of each word in the question sentence to the words in the current answer sentence, as follows:

$$\mathbf{S}_{aq}^{l} = softmax\left(\mathbf{S}_{a\to q}^{l}[i,:]\right), \tag{2.15}$$

$$\mathbf{S}_{qa}^{l} = softmax(\mathbf{S}_{q\to a}^{l}[:,j]). \tag{2.16}$$

For each word representation of the interaction matrix of the question or answer, the weighted sum of the column or row vectors is performed, and then each word vector in the weight and value matrix is used as a dot product. The formula can be expressed as:

$$\mathbf{Z}_{q}^{l} = \sum_{j}^{n} \mathbf{S}_{aq}^{l}[j] \odot \mathbf{V}_{q}^{l}, \tag{2.17}$$

$$\mathbf{Z}_{a}^{l} = \sum_{i}^{m} \mathbf{S}_{qa}^{l}[i] \odot \mathbf{V}_{a}^{l}, \tag{2.18}$$

where $L$ is the written heads in the model, and the encoded $Z_q \in R^{L\times m\times d_k}$ and $Z_a \in R^{L\times n\times d_k}$ assemble and write more the head tuples to form a three-dimensional matrix.

### *2.4. Prediction layer*

After that, the multi-head output of the questions and answers obtained by the feed-forward layer encoding is accumulated, and layer normalization is added to the final prediction layer to calculate the similarity matrix, which is defined as:

$$\mathbf{S} = LayerNorm(\sum_{l}^{L} \mathbf{Z}_{q}^{l}\mathbf{W}_{q} + \mathbf{Z}_{a}^{l}\mathbf{W}_{a}). \tag{2.19}$$

Given a collection of question and answer candidate sentences containing available ground truth knowledge about whether they are relevant, the model variables are optimized by minimizing a regularized cross-entropy cost function, which can be expressed as:

$$\begin{aligned} L(\boldsymbol{\theta}) = -\sum_{(i,j)\in I} \Big[ & t_{ij} \log p(t_{ij} = k|\mathbf{S}_{ij}) \\ & +(1 - t_{ij}) \log\left(1 - p(t_{ij} = k|\mathbf{S}_{ij})\right)\Big] + \frac{\lambda}{2}\left\|\boldsymbol{\theta}\right\|_2^2, \end{aligned} \tag{2.20}$$

where the index set $I$ represents the training question-answer sentence pair used, and the regularization parameter $\lambda > 0$. In this paper, softmax is used to calculate the correlation probability $p(t|\mathbf{s})$ between the candidate answer and the question.

## 3. Experimental settings

In this section, we conduct experimental design and evaluation on the model DMI. The dataset and experimental setup of the model are introduced first, followed by the training configuration and parameters, and several model evaluation metrics.

### 3.1. Dataset

The Q&A Yahoo! collection is a large-scale data set formed by collecting community-based Webscope Program. It includes about 4 million questions and answers, each of which is associated with the best answer and a category. The BM25 retrieval algorithm is used to retrieve the top 100 answers to each question. These retrieved answers are also marked as the correct answer for each corresponding question, ranked after the best answer provided by the set [21].

Stack Exchange [22] is a popular forum-based question and answer service. These forums allow users to post questions and answers, and one of the answers can be marked as the accepted answer. Additionally, users can vote on questions and answers to associate a score with each post. These forums cover a wide range of topics, and the evaluation dataset in this paper is based on specific forums that focus on legal aspects. The statistics of the data set are shown in Table 1.

**Table 1.** Dataset statistics.

| Parameter | Yahoo! | StackEx(L) |
|---|---|---|
| No. of Questions | 90,000 | 6939 |
| No. of Answers | 4.5M | 8595 |
| Mean Question Length(words) | 9.73 | 136.03 |
| Mean Answer Length(words) | 99.38 | 217.61 |

### 3.2. Text pre-processing and training configuration

To process the dataset, a special end-of-sentence symbol ⟨EOS⟩ is added to the end of each sentence, while out-of-vocabulary words are mapped to special token symbols ⟨UNK⟩. We follow the same text pre-processing procedure as in [23, 24]. The bidirectional recurrent network used consists of 2 layers, and each layer contains $d_k = 64$ hidden units. The dimensionality of each word embedding vector is set to $d = 215$. The number of multiple written heads for the model is set to 8. This work uses BERT [15] and RoBERTa [25] to initialize the word embedding vectors of the data. For words not present in the corpus, random values uniformly sampled from the interval [-0.3, 0.3] are assigned to each embedding dimension. Model variables are initialized using the normal distribution $\mathcal{N}(0, 0.1)$.

For model optimization, we use the RMSProp algorithm. The training process contains a mini-batch of 50 training examples, where the learning rate is 0.1 and the dropout rate is 0.5. After training for 15 rounds, set the learning rate to be halved. Gradient clipping is used to scale the gradient descent value when the gradient norm exceeds a threshold of 5. The data set is divided into the training set, verification set and test set, and the data distribution is shown in Table 2.

**Table 2.** Distribution of data set.

| Data Set | Q/A Pairs | Development | Training | Testing |
|---|---|---|---|---|
| Yahoo! [21] | 4M | 2500 | 50,000 | 25,000 |
| StackEx(L) [22] | 7760 | 1500 | 4760 | 1500 |

### 3.3. Model evaluation metrics

For the model effect test, three different evaluation algorithms are used in this paper: Mean Reciprocal Rank (MRR), Mean Average Precision (MAP) and $\text{MRT}_N$. The MRR algorithm focuses on the sorting of correct answers, and the formula is expressed as:

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{r_i^1}, \tag{3.1}$$

among them, for the $i$th question, $r_i^j$ is the ranking of the $j$th answer in the ground truth. $|Q|$ is the total number of test questions and MAP is the cumulative average rank of all correct answers in each question. The formula is:

$$\text{MAP} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{n_i^j}{r_i^j}, \tag{3.2}$$

where $n_i^j$ is the number of correct answers in the sorted list of the $j$th answer. $n_i$ corresponds to the number of correct answers in the question. $\text{MRT}_N$ or $p_N$ is to calculate the average ranking in the top-N answers, which is more flexible than the MAP index, mainly by selecting the value of $N$ to solve.

$$\text{MRT}_N = p_N = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{n_i^N}{N}, \tag{3.3}$$

where $N$ is a positive integer, and $n_i^N$ is the number of true correct answers among the first $N$ answers to the $i$th question. In this paper, $N = \{3, 5, 10\}$ is used as the evaluation index.

## 4. Experimental results and analysis

### 4.1. Algorithm comparison experiment results

In order to verify the performance of the model proposed in this paper, we use three performance indicators to test the Yahoo! and Stack Exchange datasets, namely MRR, MAP and $p_N$. Moreover, we show the advantages and disadvantages of model performance by comparing the effects of RNN, CNN and Transformer-based methods. The results of the effect comparison between different models are shown in Table 3. It can be seen from the table that the DMI proposed in this paper has achieved the best performance of the two data sets in terms of various performance indicators: 1) The best MRR performance is obtained on the Yahoo! dataset, and the improvement is 1.68% when compared with the BERT model. In particular, the proposed model DMI+RoBERTa beats the second best SLP + RoBERTa model by 0.59% in MRR result over Yahoo! dataset. 2) StackEx(L) data has a large database, and the preprocessing model shows performance advantages on this data. Compared with traditional RNN and CNN-based models, the proposed DMI+RoBERTa model obtains the best performance in evaluation algorithms.

From Table 3, it can be seen that for all datasets, except that the MAP performance of DMI model is 1.45% lower than that of BERT-base over Yahoo! data due to Yahoo's community dataset has some slang. The proposed model DMI provides a significant improvement in the evaluation results of all datasets.

**Table 3.** Comparison of performance effects of different models. The best results are indicated in bold and the next best results are underlined.

| | Yahoo! | | | | StackEx | | | |
|---|---|---|---|---|---|---|---|---|
| Models | MRR | MAP | $p_3$ | $p_5$ | MRR | MAP | $p_3$ | $p_5$ |
| ARC-II [26] | 0.4962 | 0.4213 | 0.3494 | 0.2830 | 0.5738 | 0.5068 | 0.4438 | 0.3879 |
| Bigram-CNN [27] | 0.6096 | 0.5321 | 0.4752 | 0.4176 | 0.6842 | 0.6223 | 0.5672 | 0.5252 |
| Add-CNN [24] | 0.6920 | 0.4254 | 0.5634 | 0.5041 | 0.7043 | 0.6523 | 0.6044 | 0.5513 |
| Bi-LSTM [28] | 0.6476 | 0.5779 | 0.5243 | 0.4728 | 0.7199 | 0.6528 | 0.5900 | 0.5310 |
| ATTN-Bi-LSTM [14] | 0.6753 | 0.6125 | 0.5455 | 0.4835 | 0.7766 | 0.7096 | 0.6466 | 0.5916 |
| ATTN-LSTM-CNN [29] | 0.6865 | 0.6208 | 0.5439 | 0.4882 | 0.7878 | 0.7130 | 0.6537 | 0.6010 |
| ELMo [30] | 0.7163 | 0.6758 | 0.6046 | 0.5612 | 0.7942 | 0.7537 | 0.6833 | 0.6320 |
| BERT-based [15] | 0.7218 | 0.6792 | 0.6068 | 0.5921 | 0.8045 | 0.7641 | 0.7196 | 0.6728 |
| TANDA [31] | 0.7259 | 0.6631 | 0.6048 | 0.5907 | 0.8136 | 0.7608 | 0.7245 | 0.6790 |
| SLP+RoBERTa [32] | 0.7566 | **0.6924** | 0.6379 | 0.6255 | 0.8396 | 0.7852 | 0.7481 | 0.6943 |
| DMI | 0.7386 | 0.6647 | 0.6102 | 0.6079 | 0.8220 | 0.7831 | 0.7334 | 0.6812 |
| DMI+RoBERTa | **0.7625** | 0.6912 | **0.6416** | **0.6389** | **0.8461** | **0.7964** | **0.7572** | **0.7059** |

**Table 4.** Comparison of the proposed DMI method with alternative model settings, under the Yahoo! dataset. The best performance is highlighted in bold and the second best is underlined.

| Experimental settings | | MRR | MAP | $p_3$ |
|---|---|---|---|---|
| Exp. 1 | DMI+RNN | 0.7165 | 0.6316 | 0.5920 |
| | DMI+Bi-RNN | 0.7290 | 0.6483 | 0.6011 |
| Exp. 2 | DMI wo atten | 0.7158 | 0.6420 | 0.5957 |
| Exp. 3 | Bi-MiStack | 0.7205 | 0.6491 | 0.4738 |
| | Bi-MiGRU | 0.7269 | 0.6613 | 0.6077 |
| DMI | | **0.7386** | **0.6647** | **0.6102** |

Compared with the past RNN, CNN and Transformer-based models, the BERT-based preprocessing DMI model has performed well on large-scale open-domain data.

It is found through the table that when comparing the performance of the DMI model in the Yahoo! and StackEx datasets, the performance of the model on the Yahoo! dataset is generally lower than that of other datasets. From Table 1, we can observe that the questions and answers in the dataset are different. The average sentence length is much shorter than the StackEx(L) dataset because there are a lot of internet terms in the sentences in the Yahoo! data, such as slang, emoticons and idioms, etc., for example, Internet sentences similar to the following: *" hahhaha! you might want to go with espeon because none of those can really learn any psycic moves and gyarados learns bite which is dark so... yeah... you might just get rid of it though because of lugia? "*. Therefore, the existence of the above factors adds some difficulties to the question answering task.

### *4.2. Ablation study*

**Table 5.** Sorting comparison of Top-3 models based on Yahoo! dataset.

| **Example (Yahoo! Dataset)** | |
|---|---|
| Question | pokemon soul silver/heart gold umbreon or espeon? |
| Top 3 answers by DMI | **No.1**: hahhaha! you might want to go with espeon because none of those can really learn any psycic moves and gyarados learns bite which is dark so... yeah... you might just get rid of it though because of lugia? and you get the 3 legendary dogs and birds and ho-oh eventually.(∗) |
| | **No.2**: well i can do 150 straight, but i went to military school for 4 years, and i had a drill sergeant up my ass every time i slowed down, i was pissed at the time but i realized it helped. get someone to do them with you and make a game to see who can do the most, or get someone to push you while your doing them. |
| | **No.3**: what an excellent question for your vet, after they run some tests! could be urinary tract infection, kidney failure, kidney or bladder stones, or a variety of other issues. what they all have in common is that they require vet attention. none of the possibilities are treatable at home. |
| Top 3 answers by ATTN-LSTM-CNN [29] | **No.1**: what an excellent question for your vet, after they run some tests! could be urinary tract infection, kidney failure, kidney or bladder stones, or a variety of other issues. what they all have in common is that they require vet attention. none of the possibilities are treatable at home. |
| | **No.2**: well i can do 150 straight, but i went to military school for 4 years, and i had a drill sergeant up my ass every time i slowed down, i was pissed at the time but i realized it helped. get someone to do them with you and make a game to see who can do the most, or get someone to push you while your doing them. |
| | **No.3**: hahhaha! you might want to go with espeon because none of those can really learn any psycic moves and gyarados learns bite which is dark so... yeah... you might just get rid of it though because of lugia? and you get the 3 legendary dogs and birds and ho-oh eventually.(∗) |
| Top 3 answers by BERT-based [15] | **No.1**: well i can do 150 straight, but i went to military school for 4 years, and i had a drill sergeant up my ass every time i slowed down, i was pissed at the time but i realized it helped. get someone to do them with you and make a game to see who can do the most, or get someone to push you while your doing them. |
| | **No.2**: hahhaha! you might want to go with espeon because none of those can really learn any psycic moves and gyarados learns bite which is dark so... yeah... you might just get rid of it though because of lugia? and you get the 3 legendary dogs and birds and ho-oh eventually.(∗) |
| | **No.3**: what an excellent question for your vet, after they run some tests! could be urinary tract infection, kidney failure, kidney or bladder stones, or a variety of other issues. what they all have in common is that they require vet attention. none of the possibilities are treatable at home. |

**Table 6.** The ranking results of Top-5 models for failure example based on Yahoo! dataset.

| **Example (Yahoo! Dataset)** | |
|---|---|
| Question | weight issue - is it possible to lose 21pounds in 2 weeks? |
| Top 5 answers by DMI | **No.1**: it's possible to lose 50 pounds a week if you lived on celery & worked out 12 hours a day. but honestly, who has the time for that? 21 pounds in two weeks seems like a lot. but, it is possible to lose at least 10 in a week (not sure about 20 in two weeks because the 1st week you push yourself to the limit; your body loses weight—after that you build up an immunity to your workout).(∗) |
| | **No.2**: it alll depends on where you live. you'll need to invest in a few pairs of jeans (they last you a while!) as for clothing, get some cute tanks, and blouses. like you said forever 21 but you can also try: www.wetseal.com www.sidecca.com www.cutesygirl.com www.pinkice.com www.modcloth.com these sites all have cute clothes for trendy girls, and the clothes are all decent priced/affordable.hope this helped. |
| | **No.3**: the methodist branch of protestant religion traces its roots back to 1739 where it developed in england as a result of the teachings of john wesley. wesley's three basic precepts that began the methodist tradition consisted of: 1.shun evil and avoid partaking in wicked deeds at all costs, 2.perform kind acts as much as possible, and 3.abide by the edicts of god the almighty father. |
| | **No.4**: it should be 4-door, 4-cyls, not sporty. i would prefer honda or toyota... first time drivers usually have much more expensive car insurance, and it depends from car. |
| | **No.5**: actually, you can follow this tutorial to practice, here's the official link: http://link.brightcove.com/services/player/bcpid1873832296?bctid=207499896001.(∗) |

#### 4.2.1. The effect of Bi-MiRNN of the proposed DMI model

To verify the effect of a bidirectional multi-head recurrent network (Bi-MiRNN) in the proposed DMI model, we compare the performance of the proposed DMI model with the other two variants in the Exp.1: DMI with the recurrent neural network (DMI+RNN), and DMI with the bidirectional recurrent neural network (DMI+Bi-RNN). It can be seen in Table 4 that DMI achieves the best performance, which is higher than 0.96% than the second best DMI+Bi-RNN. The comparison results show that the design of Bi-MiRNN has a competitive strategy for the proposed DMI model.

#### 4.2.2. The variant interactive learning expressions of DMI model

In the second experiment, we compare the proposed DMI model with/without (wo) word-level interactive attention mechanism based on dual-channel multiple written heads. To calculate the question and answer representations without attention, we use the key vectors $\mathbf{K}_q^l$, $\mathbf{K}_a^l$ instead of $\mathbf{K}_a^l$, $\mathbf{K}_q^l$ in the Eqs (2.13) and (2.14), respectively. From Table 4, it can be seen that the DMI with a word-level interactive attention mechanism has a better performance than the proposed model without an attention mechanism. Thus, the design of a word-level interactive attention mechanism is beneficial to compute the semantic similarity according to fusing the dual-channel multiple-written heads.

4.2.3. The choices of aggregation function in Bi-MiRNN

In the bidirectional multi-head RNN, we use the alternative design choices of aggregation function to combine the forward and backward tuples, including the simple stacked bidirectional multi-head (Stacked Bi-MiSatck), bidirectional multi-head GRU (Bi-MiGRU), and the LSTM of Eq (2.12) in the proposed model. The results of Exp.3 in Table 4 show that the proposed model with the bidirectional muti-head LSTM function has a robust performance against the other two choices due to the long-term property of the LSTM model.

### *4.3. Example analysis*

The comparison of the DMI, ATTN-LSTM-CNN and BERT examples in work, are shown in Table 5, where $*$ is the ground truth annotation of the correct answer. It can be seen from the table that compared to the traditional LSTM and CNN, which use attention to perform feature fusion algorithm and BERT-based to independently learn the feature representation of a single sentence, although the above methods have global relevance in the words in the corpus, However, the interaction order between local word levels is ignored. In the task of sorting candidate answers, DMI in this paper uses a bidirectional multi-head recursive recurrent network to more accurately select the best candidate answer. In addition, we analyze an example where the DMI model fails on the Yahoo dataset. The ranking results are shown in Table 6. There are multiple groundtruth answers for a question, although the DMI model can rank the groundtruth answer with more semantic information first. DMI has comprehension limitations on similar phrases ('forever 21' in the candidate's answer is actually a brand name) or user-provided URL content, especially when the candidate's answer sentences are short, which interferes with word-level interactive learning.

## 5. Conclusions

In summary, this paper proposes an interactive method based on dual-channel multiple written heads for solving open-domain question answering tasks. The DMI method we proposed improves the traditional linear transformation to solve the written head trigram and uses the architecture of a bidirectional recursive recurrent network to fuse the word-level representation effectively. Moreover, a dual-channel multi-head mechanism is designed to map sentence representations in multiple dimensions and adaptively extract interactive three-dimensionality semantic features of questions and answers. By testing on open-domain question answering data, the DMI model is able to achieve competitive performance. The main work in the future will focus on enhancing the learning structure of semantic representation so that the model can handle short text and Internet term data in actual scenarios.

## Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

## Acknowledgments

This work is supported by the Natural Science Foundation of HuBei Province (No.2021CFB255), and the National Natural Science Foundation of China (No.62171329, No.62201406).

## Conflict of interest

The authors declare there is no conflict of interest.

## References


1. M. Pan, Q. Pei, Y. Liu, T. Li, E. A. Huang, J. Wang, et al., Sprf: a semantic pseudo-relevance feedback enhancement for information retrieval via conceptnet, *Knowl.-Based Syst.*, **274** (2023), 110602. https://doi.org/10.1016/j.knosys.2023.110602
2. L. Ma, H. Hong, F. Meng, Q. Wu, J. Wu, Deep progressive asymmetric quantization based on causal intervention for fine-grained image retrieval, *IEEE Trans. Multimed.*, **2023** (2023). https://doi.org/10.1109/TMM.2023.3279990
3. H. Hasan, H. Huang, Mals-net: A multi-head attention-based lstm sequence-to-sequence network for socio-temporal interaction modelling and trajectory prediction, *Sensors*, **23** (2023), 530. https://doi.org/10.3390/s23010530
4. J. Wu, T. Mu, J. Thiyagalingam, J. Y. Goulermas, Memory-aware attentive control for community question answering with knowledge-based dual refinement, *IEEE Trans. Syst. Man Cybern. Syst.*, **53** (2023), 3930–3943. https://doi.org/10.1109/TSMC.2023.3234297
5. X. Li, B. Wu, J. Song, L. Gao, P. Zeng, C. Gan, Text-instance graph: Exploring the relational semantics for text-based visual question answering, *Pattern Recognit.*, **124** (2022), 108455. https://doi.org/10.1016/j.patcog.2021.108455
6. X. Bi, H. Nie, X. Zhang, X. Zhao, Y. Yuan, G. Wang, Unrestricted multi-hop reasoning network for interpretable question answering over knowledge graph, *Knowl.-Based Syst.*, **243** (2022), 108515. https://doi.org/10.1016/j.knosys.2022.108515
7. W. Zheng, L. Yin, X. Chen, Z. Ma, S. Liu, B. Yang, Knowledge base graph embedding module design for visual question answering model, *Pattern Recognit.*, **120** (2022), 108153. https://doi.org/10.1016/j.patcog.2021.108153
8. S. Lv, D. Guo, J. Xu, D. Tang, N. Duan, M. Gong, et al., Graph-based reasoning over heterogeneous external knowledge for commonsense question answering, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2022), 8449–8456. https://doi.org/10.48550/arXiv.1909.05311
9. Z. Wang, X. Xu, G. Wang, Y. Yang, H. T. Shen, Quaternion relation embedding for scene graph generation, *IEEE Trans. Multimedia*, **2023** (2023). https://doi.org/10.1109/TMM.2023.3239229
10. J. Wu, F. Ge, H. Hong, Y. Shi, Y. Hao, L. Ma, Question-aware dynamic scene graph of local semantic representation learning for visual question answering, *Pattern Recognit. Lett.*, **170** (2023), 93–99. https://doi.org/10.1016/j.patrec.2023.04.014
11. H. Zhang, L. Cheng, Y. Hao, C. W. Ngo, Long-term leap attention, short-term periodic shift for video classification, in *Proceedings of the 30th ACM International Conference on Multimedia*, (2022), 5773–5782. https://doi.org/10.1145/3503161.3547908
12. L. Peng, Y. Yang, Z. Wang, Z. Huang, H. Shen, Mra-net: Improving vqa via multi-modal relation attention network, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2020), 318–329. https://doi.org/10.1109/TPAMI.2020.3004830

13. Z. Wang, Z. Gao, G. Wang, Y. Yang, H. T. Shen, Visual embedding augmentation in fourier domain for deep metric learning, *IEEE Trans. Circuits Syst. Video Technol.*, **2023** (2023). http://doi.org/10.1109/TCSVT.2023.3260082

14. M. Tan, C. Santos, B. Xiang, B. Zhou, Lstm-based deep learning models for non-factoid answer selection, preprint, arXiv:1511.04108. https://doi.org/10.48550/arXiv.1511.04108

15. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, preprint, arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

16. Y. Li, W. Li, L. Nie, Mmcoqa: Conversational question answering over text, tables, and images, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (2022), 4220–4231.

17. X. Zhao, J. X. Huang, Bert-qanet: Bert-encoded hierarchical question-answer cross-attention network for duplicate question detection, *Neurocomputing*, **509** (2022), 68–74. https://doi.org/10.1016/j.neucom.2022.08.044

18. Y. Guan, Z. Li, Z. Lin, Y. Zhu, J. Leng, M. Guo, Block-skim: efficient question answering for transformer, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **36** (2022), 10710–10719.

19. Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, et al., An empirical study of gpt-3 for few-shot knowledge-based vqa, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **36** (2022), 3081–3089.

20. H. Sak, A. Senior, F. Beaufays, Long short-term memory recurrent neural network architectures for large scale acoustic modeling, in *Interspeech 2014*, (2014), 338–342.

21. G. Zhou, Y. Zhou, T. He, W. Wu, Learning semantic representation with neural networks for community question answering retrieval, *Knowl.-Based Syst.*, **93** (2016), 75–83. https://doi.org/10.1016/j.knosys.2015.11.002

22. A. Anderson, D. Huttenlocher, J. Kleinberg, J. Leskovec, Discovering value from community activity on focused question answering sites: a case study of stack overflow, in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, (2012), 850–858.

23. J. Wu, T. Mu, J. Thiyagalingam, J. Y. Goulermas, Building interactive sentence-aware representation based on generative language model for community question answering, *Neurocomputing*, **2020** (2020), 93–107. https://doi.org/10.1016/j.neucom.2019.12.107

24. A. Severyn, A. Moschitti, Learning to rank short text pairs with convolutional deep neural networks, in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (2015), 373–382.

25. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., Roberta: A robustly optimized bert pretraining approach, preprint, arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

26. B. Hu, Z. Lu, H. Li, Q. Chen, Convolutional neural network architectures for matching natural language sentences, *Adv. Neural Inf. Process. Syst.*, (2014), 2042–2050.

27. L. Yu, K. M. Hermann, P. Blunsom, S. Pulman, Deep learning for answer sentence selection, preprint, arXiv:1412.1632. https://doi.org/10.48550/arXiv.1412.1632

28. M. Seo, A. Kembhavi, A. Farhadi, H. Hajishirzi, Bidirectional attention flow for machine comprehension, preprint, arXiv:1611.01603. https://doi.org/10.48550/arXiv.1611.01603

29. S. Wan, Y. Lan, J. Guo, J. Xu, L. Pang, X. Cheng, A deep architecture for semantic matching with multiple positional sentence representations, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **30** (2016), 2835–2841.

30. M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, et al., Deep contextualized word representations, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, (2018), 2227–2237.

31. S. Garg, T. Vu, A. Moschitti, Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2020), 7780–7788.

32. L. D. Liello, S. Garg, L. Soldaini, A. Moschitti, Pre-training transformer models with sentence-level objectives for answer sentence selection, preprint, arXiv:2205.10455. https://doi.org/10.48550/arXiv.2205.10455


AIMS Press