Hindawi
Complexity
Volume 2021, Article ID 7367181, 13 pages
https://doi.org/10.1155/2021/7367181

WILEY | Hindawi

Research Article

Dual-Channel Reasoning Model for Complex Question Answering

Xing Cao,[1,2] Yun Liu,[1,2] Bo Hu,[1,2] and Yu Zhang[1,2]

[1]*School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China*
[2]*Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing 100044, China*

Correspondence should be addressed to Yun Liu; liuyun@bjtu.edu.cn

Received 10 May 2021; Revised 23 June 2021; Accepted 8 July 2021; Published 26 July 2021

Academic Editor: Xuzhen Zhu

Multihop question answering has attracted extensive studies in recent years because of the emergence of human annotated datasets and associated leaderboards. Recent studies have revealed that question answering systems learn to exploit annotation artifacts and other biases in current datasets. Therefore, a model with strong interpretability should not only predict the final answer, but more importantly find the supporting facts' sentences necessary to answer complex questions, also known as evidence sentences. Most existing methods predict the final answer and evidence sentences in sequence or simultaneously, which inhibits the ability of models to predict the path of reasoning. In this paper, we propose a dual-channel reasoning architecture, where two reasoning channels predict the final answer and supporting facts' sentences, respectively, while sharing the contextual embedding layer. The two reasoning channels can simply use the same reasoning structure without additional network designs. Through experimental analysis based on public question answering datasets, we demonstrate the effectiveness of our proposed method

1. Introduction

One of the long-standing goals of natural language processing (NLP) is to enable machines to have the ability to understand natural language and make inferences in textual data. Many applications, such as dialogue systems [1, 2], recommendation systems [3–5], question answering [6, 7], and sentiment analysis [8], aim to explore the machine ability to understand textual data. Question answering, abbreviated as QA, has emerged as an important natural language processing task because it provides a quantifiable way to evaluate an NLP system's capability on language understanding and reasoning and its commercial value for real-world applications.

Most previous works have focused on questions answering only from a single paragraph, known as single-hop QA [9]. Although recent advances in QA and machine reading comprehension (MRC) had surpassed human performance on some single-hop datasets [10, 11], those datasets have gaps from real-world scenarios. In the real world, there are a lot of complex questions that need to be answered through multiple steps of reasoning by aggregating information distributed in multiple paragraphs, named multihop QA [12].

Jiang and Bansal [13] pointed out that because examples include reasoning shortcuts, some models may directly locate the answer by word-matching the question with sentences in the context. For the complex question "what was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?." The context contains the sentence "Peter Bolesław Schmeichel..... and was voted the IFFHS World's Best Goalkeeper in 1992 and 1993." At this time, the model may find the correct answer "World's Best Goalkeeper" through simple word matching, but does not infer that Peter Bolesław Schmeichel is the father of Kasper Schmeichel. Therefore, to enhance the interpretability of models and avoid answering complex questions through reasoning shortcuts, our study considers that, in addition to predicting the correct answer, it is also important to extract evidence sentences. However, most of the existing works only focused on improving the

accuracy of the model to answer complex questions, but pay less attention to the ability of the model on predicting the inference path.

An example from HotpotQA is illustrated in Figure 1. Ten paragraphs are given to answer complex questions ("*what government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?*"); the model first needs to identify passage 2 (P2) and passage 6 (P6) above as relevant paragraphs to correctly answer the question.

The first sentence of P6 and P2 are evidence sentences, which lead to the next-hop paragraph and the predicted answer, respectively. However, it is more difficult for the model to predict correct and complete evidence sentences than to answer a complex question because the question often does not contain information about the intermediate answer, such as "Shirley Temple" (green font) in Figure 1.

Most existing methods [14–16] predicted the final answer and the supporting facts in sequence or simultaneously, and the architecture of these methods is designed primarily to predict the right answer. In this paper, we propose a novel dual-channel reasoning architecture for complex question answering. Concretely, complex questions and documents pass through the word embedding layer and contextual embedding layer in succession. Thereafter, the output of the contextual embedding layer is the input of two reasoning channels: one for predicting the answer span or answer type, and the other for predicting evidence sentences.

Our contributions can be summarized as follows:

(1) We propose the dual-channel reasoning architecture, which is a novel architecture for the complex question answering task. The results of the experiment show that the dual-channel reasoning architecture is suitable for many kinds of existing neural network models, such as graph-based models.

(2) We perform comprehensive experiments on multihop QA datasets, and our proposed method outperforms previous approaches on complex questions, especially on the task of extracting evidence sentences. We conducted a detailed visual analysis of the baseline model and two channels in the dual-channel architecture and further explored the differences in the distribution of attention heat maps of several models.

2. Related Work

2.1. Multihop Question Answering over Knowledge Base. Knowledge-based question answering (KBQA) computes answers to natural language questions based on a knowledge base. Besides the traditional methods of defining templates and rules, KBQA methods can be mainly divided into two branches: semantic parsing (SP) based and information retrieval (IR) based. Semantic parsing methods focus on translating complex natural language questions into the executable query graph over the knowledge base. Lan and Jiang [17] proposed a modified staged query graph generation method by allowing longer relation paths. Sun et al. [18] proposed a novel skeleton grammar that uses the BERT-based parsing algorithm to improve the downstream fine-semantic parsing. To avoid generating noisy candidate queries, Chen et al. [19] proposed abstract query graphs (AQG) to describe query structures. The IR-based model first extracts the subject entities mentioned in the question and links them to the knowledge base [20]. Then, the subgraph centered on the subject entity is extracted, and all nodes in the subgraph are selected as candidate answers. Chen et al. [21] used a novel bidirectional attentional memory network to simulate the bidirectional interactive flow between a question and a knowledge base. Xu et al. [22] enhanced KV-MemNNs models by a new query updating strategy to perform interpretable reasoning for complex questions.

2.2. Multihop Question Answering over Text. Currently, there are two mainstream branches for complex question answering over textual data. The first direction is to apply the previous neural networks that are successful in single-hop QA tasks to multihop QA tasks. The Bidirectional Attention Flow (Bi-DAF) network proposed by Seo et al. [23] has achieved state-of-the-art results in single-hop QA datasets. Yang et al. [12] proposed the multihop dataset HotpotQA and used the model with the Bi-DAF module as the core which was used as the baseline model of this dataset. Zhong et al. [24] proposed a model combination of coarse-grained reading and fine-grained reading. The query-focused extractor model proposed by Nishida et al. [16] regards evidence extraction as a query-focused summarization task and reformulates the query in each hop. Because the semantics of the questions in the multihop QA task is more complex, it is difficult for the Bi-DAF module to fully understand the semantics. Min et al. [25] addressed HotpotQA by decomposing its multihop questions into single-hop subquestions to achieve better performance and interpretability. Jiang and Bansal [26] proposed a self-assembling modular model to make multihop reasoning and support fact selection more interpretable. However, their model needs to be trained by using a large amount of manually labeled data, which is undoubtedly expensive. Because answers to complex questions require aggregating information from multiple paragraphs and BERT cannot encode all documents at once, Bhargav et al. [27] proposed translucent answer prediction architecture to effectively capture the local context and the global interactions between the sentences.

The other direction is based on graph neural networks (GNNs) [28]. Graph is an effective way to represent complex relationships between entities and to obtain relationship information. Ding et al. [29] used the implicit extraction module and explicit reasoning module to build the reasoning process into a cognitive graph. Inspired by human's step-by-step reasoning behavior, Qiu et al. [15] proposed a dynamically fused graph network that can predict the subgraphs dynamically at each reasoning step. The multilevel graph network can represent the information in the context in more detail. The hierarchical graph network (HGN) proposed by Fang et al. [14] captures clues from different granularity levels and weaves heterogeneous nodes into a single unified graph.

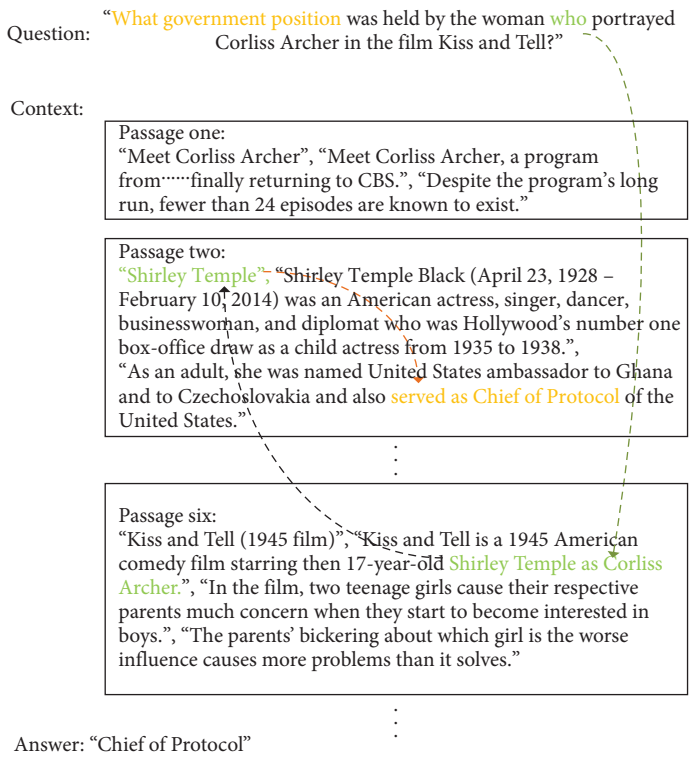

Figure 1: Example of multihop datasets HotpotQA.

3. Task Formulation

Suppose we are given a set of training data $\{C_i, Q_i, A_i, \text{Sup}_i\}$, where each context C_i is composed of many documents $\{P_1, P_2 \ldots P_n\}$ related to the question and is regarded as one connected text $C_i = \{x_1, x_2 \ldots x_T\}$, and $Q_i = \{q_1, q_2 \ldots q_J\}$is regarded as a complex query; context C_i and query Q_i have T words and J words, respectively.

The goal of the task is to design models to predict A_i and Sup_i. A_i includes answer type A_T and answer string A_S; the answer type A_T is selected from the answer candidates, such as "yes/no/span." The answer string A_S is a short span in context, which is determined by predicting the positions of the start indexes and the end indexes when there are not enough answer candidates to answer Q, expressed by <start_i, end_i>. Sup_i is regarded as evidence sentences, and supporting facts include more than one sentence in C_i, expressed by<paragraph title, sentence indexes>.

4. Solution Approach

4.1. Process Overview. We describe the dual-channel reasoning architecture in this section. Our proposed model consists of four components that are the input module, contextual module, reasoning module, and prediction module. To test the applicability of the proposed architecture, the input module, contextual module, and reasoning module, respectively, adopt different current mainstream neural networks. The overall dual-channel reasoning architecture is illustrated in Figure 2.

4.2. Input Module. An input question $Q_i = \{q_1, q_2 \ldots q_J\}$ and context $C_i = \{x_1, x_2 \ldots x_T\}$ are represented as sequences of word embeddings and character embeddings, respectively. The concatenation of the character and word embedding vectors is passed to the highway network, and the outputs of the highway network are two matrices $X_1 \in \mathbb{R}^{T \times d_1}$ for the context and $Q_1 \in \mathbb{R}^{J \times d_1}$ for the query, where d_1 is the dimension after fusion of the word embedding and character embedding. In addition, the input module can also use a pretrained model, BERT. The query Q_i and the context C_i are concatenated, and they pass the resulting sequence to a pretrained BERT model to obtain representations $X_2 \in \mathbb{R}^{T \times d_2}$ for the context and $Q_2 \in \mathbb{R}^{J \times d_2}$ for the query, where d_2 is the size of BERT hidden states.

4.3. Contextual Module. To model the temporal interactions between words in context and question, bidirectional long short-term memory (Bi-LSTM) networks are applied above the input module. The output representation of Bi-LSTM are $U \in \mathbb{R}^{J \times 2d_1}$ and $H \in \mathbb{R}^{T \times 2d_1}$, where $2d_1$ denotes the output dimension. For the graph neural network method, identifying supporting entities and the text span of potential answers from the output of BERT are used as nodes in the graph. Undirected edges are defined according to the positional properties of every node pair.

4.4. Reasoning Module. The reasoning module includes the context-query interaction layer and modeling layer. The typical implementation of the context-query interaction

layer is Bi-DAF. Bi-DAF is responsible for connecting and integrating the information of context and query words. Finally, the contextual module output and the vectors computed by the context-query interaction layer are combined to yield G:

$$\begin{aligned}
G_{t,:} &= \tilde{\beta}\left(H_{t,:}, \tilde{U}_{t,:}, \tilde{H}_{t,:}\right),\\
\tilde{\beta}(h,\tilde{u},\tilde{h}) &= [h;\tilde{u};h\circ\tilde{u};\tilde{h}\circ\tilde{u}],\\
\tilde{U}_{t,:} &= \sum_j a_{t,j}U_{j,:},\\
a_{t,:} &= \text{softmax}\left(S_{t,:}\right), \quad a_{t,:}\in\mathbb{R}^T,\\
S_{tj} &= [h+u+\alpha(H,U)], \quad S_{tj}\in\mathbb{R}^{T\times J},\\
\alpha(H,U) &= U^T H, \alpha(H,U)\in\mathbb{R}^{T\times J},\\
h &= \text{linear}(H), \quad h\in\mathbb{R}^{T\times 1},\\
u &= \text{permute}(\text{linear}(U)), \quad u\in\mathbb{R}^{1\times J},\\
\tilde{h} &= \sum_t b_t H_{t,:}, \quad \tilde{h}\in\mathbb{R}^{2d},\\
b &= \text{softmax}\left(\max_{\text{col}}(S)\right), \quad b\in\mathbb{R}^J.
\end{aligned} \tag{1}$$

where $\tilde{h}$ is tiled T times across the column, thus giving $\tilde{H}\in\mathbb{R}^{T\times 2d}$, [;] is vector concatenation across row, S is the similarity matrix, and $\tilde{U}$ and $\tilde{H}$ represent the output of context-to-query attention and query-to-context attention, respectively. The output G of the context-query interaction layer is taken as the input to the modeling layer, which encodes the query-aware representations of context words. We use one layer of the bidirectional GRU to capture the interaction among the context words conditioned on the query. Since multiple documents contain thousands of words, the long-distance dependency problem is obvious, so a self-attention module is added to alleviate this problem.

For the graph neural network method, graph attention networks, graph recurrent networks, and graph convolutional networks, their variants can propagate messages across different entity nodes in graphs and update the vector representation of the original entity.

4.5. Prediction Module. The prediction module consists of four homogeneous Bi-GRU and linear layers. Corresponding to the channels used to predict the answer are three sets of Bi-GRU and linear layers, and they have three output dimensions, including (1) the start indexes of the answer, (2) the end indexes of the answer, and (3) the answer type. The prediction module corresponding to the evidence sentences extraction channel only outputs the supporting sentences predicted by the model.

5. Experiments

5.1. Datasets. HotpotQA is a recently introduced multihop QA dataset with 113k Wikipedia-based question-answer pairs. HotpotQA has two benchmark settings, namely, distractor setting and full wiki setting. In the distractor setting, for each example, there are two golden paragraphs related to complex questions and eight unrelated ones. The two gold paragraphs and the eight distractors are shuffled before they are fed to the model. Full wiki setting requires the model to answer the question given the first paragraph of all Wikipedia articles, in which no specified golden paragraphs are given. Here, we focus on the HotpotQA dataset under the distractor setting to challenge the model to find the true supporting facts in the presence of noise. For the full wiki setting where all Wikipedia articles are given as input, we consider the bottleneck to be about information retrieval, thus we do not include the full wiki setting in our experiments. In HotpotQA, only the training and validation data are publicly available, while the test data are hidden. For further analysis, we report only the performance on the validation set, as we do not want to probe the unseen test set by frequent submissions. According to the observations from our experiments and previous works, the validation score is well correlated with the test score.

5.2. Model Comparison. We compared the results with those of the three categories' model. The first category is the model which follows the feature interaction framework, such as models with Bi-DAF as the core component, specifically, NMN, etc. The second category is the reasoning model based on a graph neural network, such as KGNN and DFGN. The third category is the pretrained model, such as BERT.

5.2.1. Baseline. The baseline model was proposed in the original HotpotQA paper. The network architecture is composed of context and question embedding layer, contextual embedding layer, modeling layer, and prediction layer from the bottom to the top.

5.2.2. NMN. NMN is a self-assembling modular model for multihop QA. Four atomic neural modules were designed, namely, Find, Relocate, Compare, and NoOp, where four neural modules were dynamically assembled to make multihop reasoning and support fact selection more interpretable.

5.2.3. KGNN. KGNN is a knowledge-enhanced graph neural network (KGNN), which performs reasoning over multiple paragraphs.

5.2.4. DFGN. DFGN is a dynamically fused graph network that can predict the subgraphs dynamically and update query at each reasoning step.

5.2.5. BERT. BERT has been shown to be successful on many NLP tasks, and recent papers have also examined complex QA using the BERT model.

5.2.6. Coarse-Grained Decomposition Strategy. To solve the problem that the original Bi-DAF module cannot obtain a query-aware context representation correctly for complex

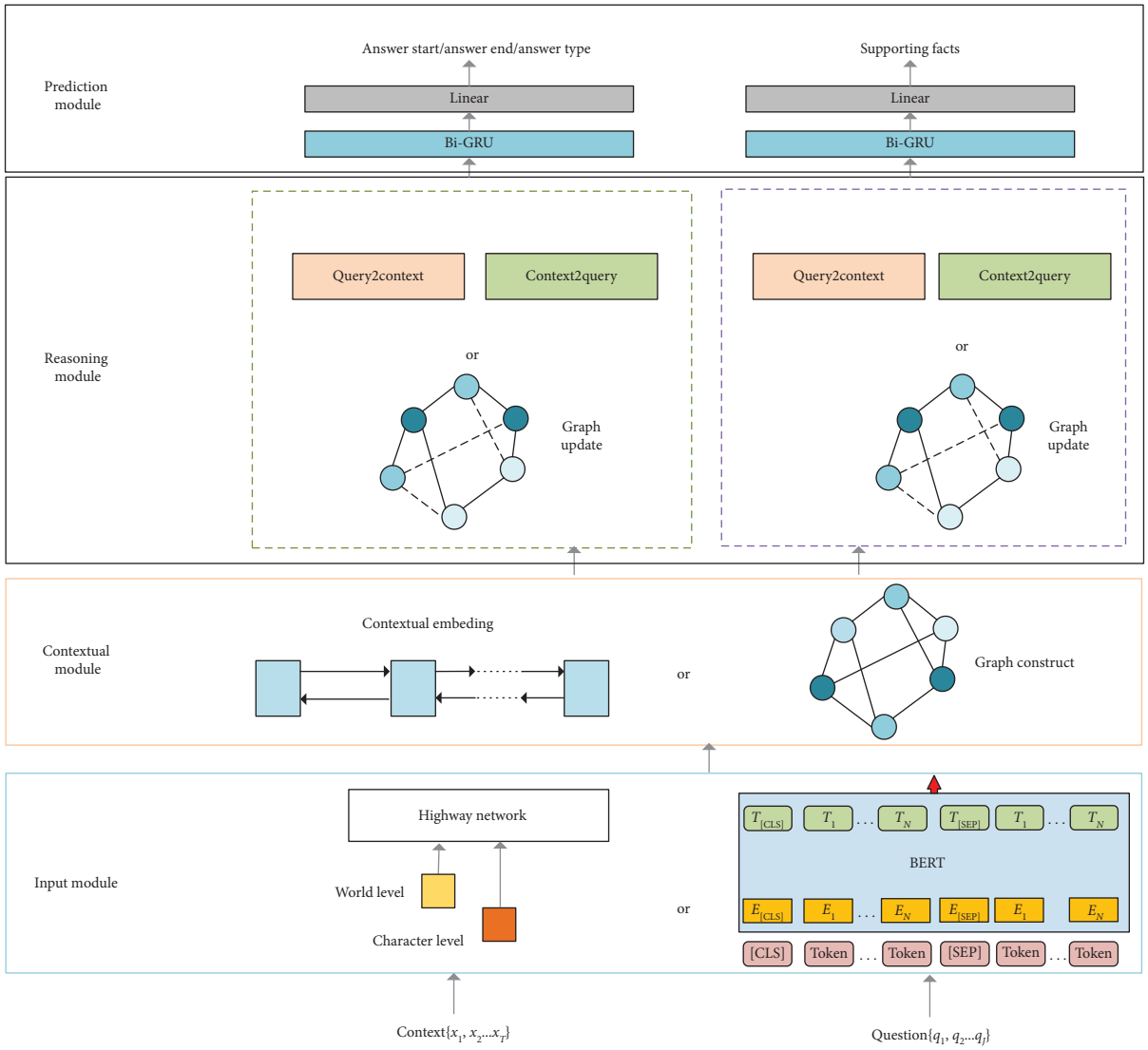

Figure 2: Overview of the dual-channel reasoning architecture.

questions, Cao and Liu [6] proposed the coarse-grained decomposition strategy, named CGDe strategy. The CGDe is responsible for decomposing complex questions and generating a new question which contains the semantics of intermediate answers that appear in the text to a certain extent.

5.2.7. Fine-Grained Interaction Strategy. Cao and Liu [6] proposed the fine-grained interaction strategy to solve deficiencies of vanilla Query2Context, named FGIn strategy. Instead of max pooling operation, softmax is used for each column of the attention matrix, and then, the document vector is dotted with each column weight. The method obtains J vector matrices of size $(T, 2d)$, where J is the number of words in the question. Finally, the J matrices are added to obtain the output matrix of the same size as the original Query2Context module. Comprehensive experiments showed that FGIn strategy predicts the number of evidence sentences more accurately than the baseline.

5.3. Implementation Details. To prove that our model components and model architecture have absolute performance advantages over the baseline model, we reimplemented the architecture described in the works by Yang et al. [12] and Qiu et al. [15].

5.3.1. Baseline Model for HotpotQA Dataset. We use the standard 300-dimensional pretrained GloVe as word embeddings. The dimensions of hidden states in Bi-GRU are set as $d = 80$. Using the Adam optimizer, with a minibatch size of 32 and an initial learning rate of 0.01, an early stopping strategy is adopted, with patience = 1.

5.3.2. Dynamically Fused Graph Network. We also used a pretrained BERT model as the encoder, d is 768. All the hidden state dimensions are set to 300 using the Adam optimizer and an initial learning rate of 0.0001.

5.4. Main Results. The performance of multihop QA on HotpotQA is evaluated by using the exact match (EM) and $F1$ as two evaluation metrics for answer prediction and evidence sentences extraction. Exact match (EM) means that the answer or evidence sentences predicted by the model are exactly the same as the golden label. Joint EM is 1 only if the answer string and supporting facts are both strictly correct. The calculation formula of Joint $F1$ is

$$\begin{aligned} P^{(\text{joint})} &= P^{(\text{ans})}P^{(\text{sup})}, \\ R^{(\text{joint})} &= R^{(\text{ans})}R^{(\text{sup})}, \\ \text{Joint}F_1 &= \frac{2P^{(\text{Joint})}R^{(\text{Joint})}}{P^{(\text{Joint})} + R^{(\text{Joint})}}. \end{aligned} \tag{2}$$

To verify the general applicability of dual-channel reasoning in various neural network models, we apply dual-channel reasoning architecture to the feature interaction framework model and graph-based model, respectively. Correspondingly, we selected the baseline model proposed by Yang et al. [12] and the DFGN model proposed by Qiu et al. [15], which are the Baseline-Dual model and DFGN_Dual model in Table 1, respectively.

We integrate the CGDe strategy and the FGIn strategy proposed by Cao and Liu [6] into the dual-channel architecture. The CGDe strategy is conducive to finding the answer, so the channel for predicting the answer in the dual-channel architecture uses the CGDe strategy, and the other reasoning channel uses the baseline reasoning module. Similarly, FGIn is conducive to extracting evidence sentences, and the FGIn strategy is used for supporting facts prediction channels, which means that the dual-channel architecture is FGIn-Baseline.

We compare our approach with several previously published models and present our results in Table 1, where * represents the result of our reimplementation of the model. As shown in Table 1, all the results of our proposed model are superior to those of the baseline model, especially in supporting fact prediction tasks, both EMsup and F1sup have greatly improved. It is worth noting that although our model does not use any pretrained language model such as BERT for encoding, it outperformed the methods that used BERT such as DFGN, DFGN/BERT, and BERT Plus in the supporting facts prediction task.

5.5. Ablation Studies. In this paper, a dual-channel reasoning architecture is designed for complex question answering. To study the contributions of the dual-channel structure and these two strategies to the performance of our model, we perform an ablation experiment on the HotpotQA datasets.

As shown in Table 2, the three models in the dual-channel reasoning architecture are superior to all single-channel models on all metrics of supporting the fact prediction task (see the bottom of Table 2). Table 2 shows that when the baseline only performs answer prediction or supporting fact prediction tasks, both EM and $F1$ metrics are higher than models that simultaneously perform answer prediction and supporting fact prediction. It shows that when the single-channel reasoning structure is adopted, the two tasks not only do not promote each other but also reduce the model's ability to extract evidence. Using the dual-channel reasoning structure, the two tasks promote each other, and the score of supporting facts' extraction tasks is higher than those of complex methods that use graph neural networks and pretrained language models. In the CGDe-Baseline architecture, there is a significant improvement in the indicators on the answer prediction task, while the performance on the supporting facts prediction task drops slightly. As Cao and Liu [6] concluded, the CGDe model's ability to predict supporting facts is limited because the new question generated contains the intermediate answer required for the first subquestion, so the support sentence that answers the first question may not be predicted as a supporting fact. In the CGDe-Baseline architecture, the performance of the supporting facts prediction task is also affected, which further proves that there is a soft interaction between two reasoning channels in the dual-channel reasoning architecture.

5.6. Analysis and Visualization. In this section, we conduct a series of visual analyses with different settings using our approach.

For a more intuitive analysis, on the HotpotQA validation set, we evaluate the baseline model proposed by Yang et al. [12], the dual-channel reasoning model, and the model that only performs answer prediction task or supporting facts prediction task. At the same time, heat maps of the attention matrices of these models are generated. As the heat map of the model that only performs the answer prediction task (Figure 3) shows, the phrase "*who portrayed Corliss Archer in the film Kiss and Tell?*" used to describe constraints in the complex question has low correlation with all words in the document (the part within the red frame in the figure). This means that the model only answers part of the question, and complex questions are mistakenly regarded as simple questions. Similar to Figure 3, the phrase "*who portrayed Corliss Archer in the film Kiss and Tell?*" in Figures 4 and 5 is also low in correlation with the words in the document, but the correlation in Figure 4 is better than that in Figure 3, and the correlation in Figure 5 is better than that in Figure 4.

The reason for the high correlation of the corresponding positions in Figure 4 is that the baseline model also extracts evidence sentences while predicting the answer, using a single-channel reasoning structure. The correlation of corresponding words shown in Figure 5 is further superior to that shown in Figure 4, indicating that the supporting facts prediction task has a greater impact on the answer prediction task in dual-channel reasoning architecture. It is worth noting that although the EM_{ans} and F1_{ans} values of the only-ans model are slightly higher than those of the baseline model, it may be that the only-ans model mistakenly regards complex questions as simple questions and happens to find the correct answer by using the reasoning shortcut pointed out by Jiang and Bansal [13].

Table 1: The performance of our model and competing approaches on the HotpotQA dataset.

Model	Answer		Sup fact		Joint	
	EM	*F*1	EM	*F*1	EM	*F*1
Baseline	45.60	59.02	20.32	64.49	10.83	40.16
NMN	49.58	62.71	—	—	—	—
KGNN	50.81	65.75	38.74	76.69	22.40	52.82
BERT Plus	55.84	69.76	42.88	80.74	27.13	58.23
DFGN	56.31	69.69	51.50	81.62	33.62	59.82
DFGN*	55.19	68.68	49.72	80.67	31.53	58.26
DFGN/BERT	55.17	68.49	49.85	81.06	31.87	58.23
Our model						
Baseline-Dual	49.56	64.15	47.61	83.45	26.44	55.41
CGDe-Baseline	51.23	65.42	46.71	83.03	27.69	55.77
FGIn-Baseline	50.42	64.82	48.93	**84.10**	27.95	56.47
DFGN_Dual	**55.42**	**68.90**	**50.70**	81.70	**31.76**	**58.83**

Table 2: Ablation results on the HotpotQA dataset.

Model	Answer		Sup fact		Joint	
	EM	*F*1	EM	*F*1	EM	*F*1
Baseline						
—Yang et al.	45.60	59.02	20.32	64.49	10.83	40.16
—Yang et al.*	49.06	63.62	32.01	75.40	18.12	50.36
—Only ans	49.88	64.42	—	—	—	—
—Only sup	—	—	43.15	79.61	—	—
Single channel						
CGDe/FGIn	**51.04**	**65.57**	39.45	79.80	**23.49**	**54.79**
Only CGDe	50.81	65.20	38.78	79.35	22.62	53.86
Only FGIn	49.72	64.39	**41.03**	**80.72**	23.03	53.97
Dual channel						
—Baseline-Dual	49.56	64.15	47.61	83.45	26.44	55.41
—CGDe-Baseline	**51.23**	**65.42**	46.71	83.03	26.79	55.77
—FGIn-Baseline	50.42	64.82	**48.93**	**84.10**	**27.95**	**56.47**

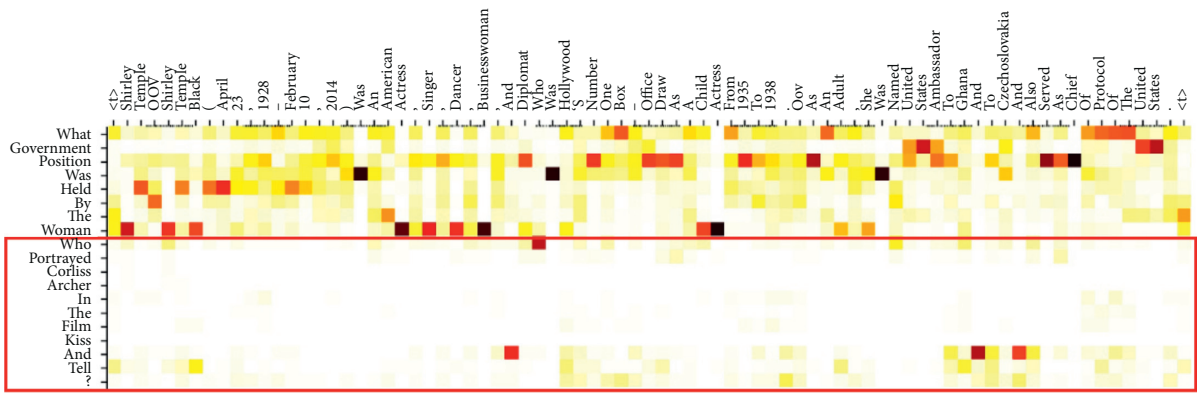

Figure 3: Attention heat map of the only-ans model.

The attention heat map shown in Figures 6 and 7 is no longer very sparse in the corresponding part of the phrase "who portrayed Corliss Archer in the film Kiss and Tell?," indicating that the model has further captured the semantics of the phrase. This is very important for the model to extract evidence sentences because "who portrayed Corliss Archer in the film Kiss and Tell?" is a constraint on the complex question.

The main difference between our dual-channel reasoning model and the single-channel reasoning model is the supporting facts prediction task. Figure 8 reveals that the reason for the high EM and F1 is that the dual-channel reasoning model (baseline-baseline) rarely extracts too many supporting facts. That is, it predicts the number of evidence sentences more accurately than the baseline model. In addition, Figure 8 shows that the dual-channel reasoning model has a similar

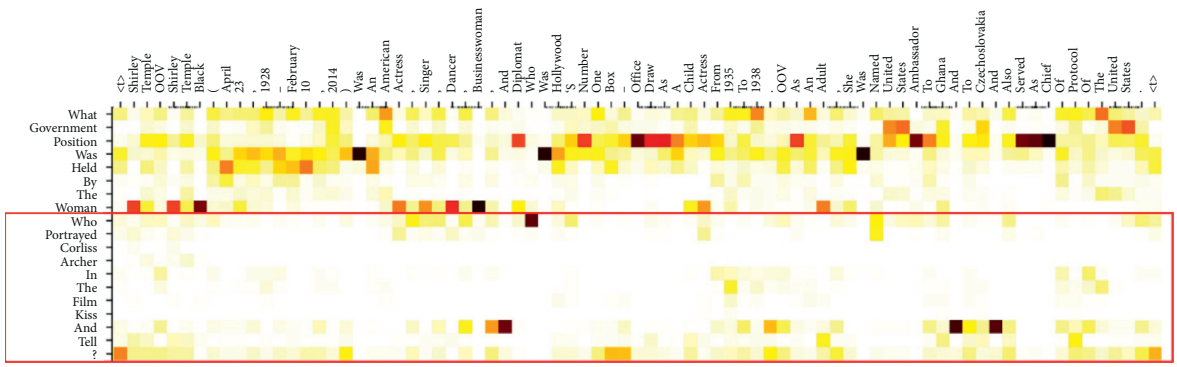

Figure 4: Attention heat map of the baseline model.

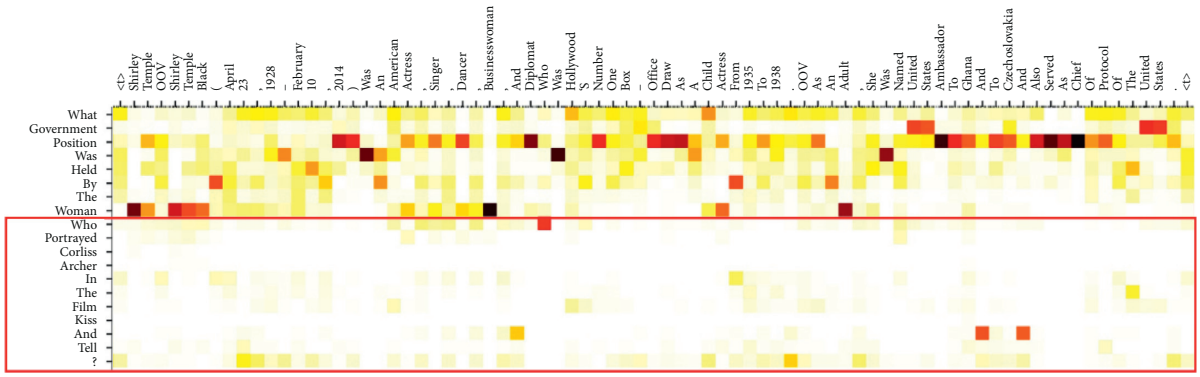

Figure 5: Attention heat map of the dual/ans model.

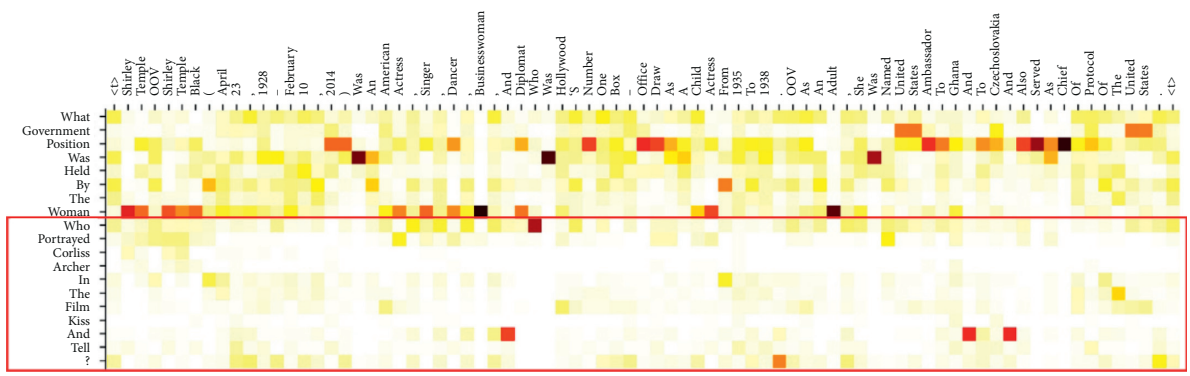

Figure 6: Attention heat map of the dual/sup model.

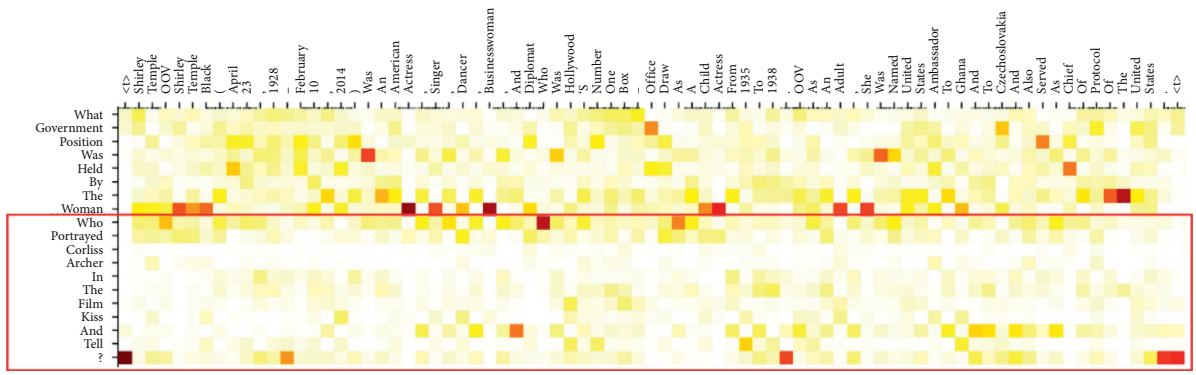

Figure 7: Attention heat map of the only-sup model.

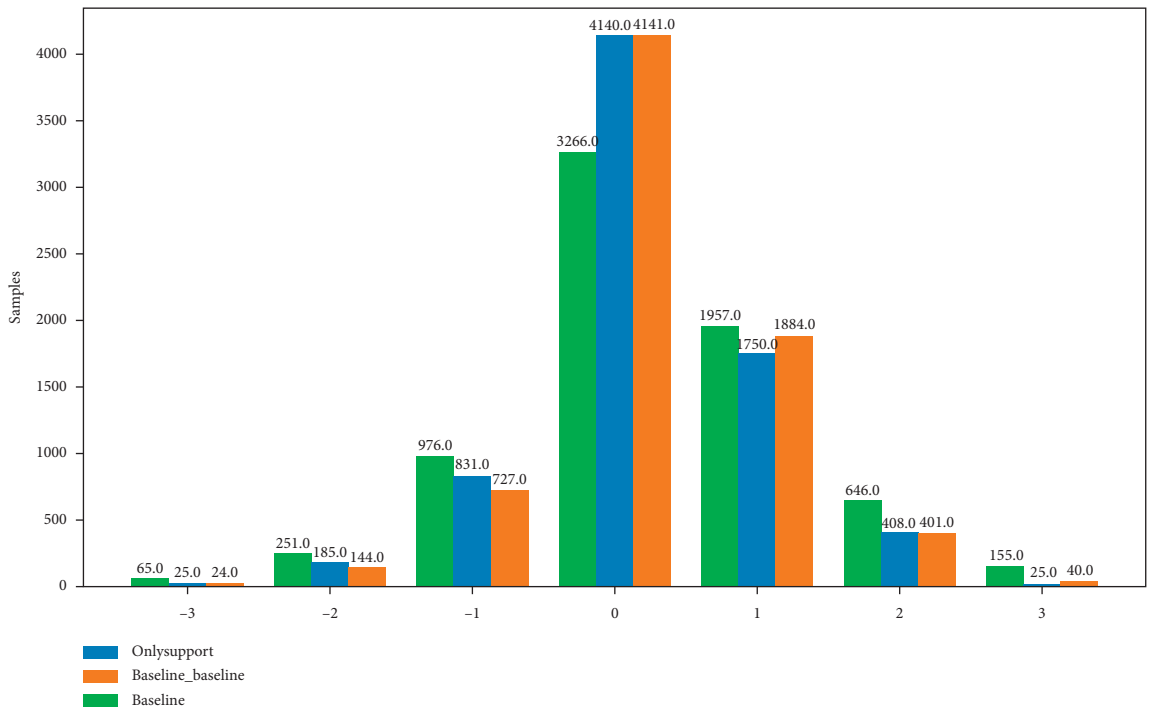

Figure 8: Number of predictions minus number of gold sentences.

distribution to the model that only performs supporting fact prediction tasks, and in the left half of the graph, the number of the latter is generally higher than the former, while in the right half of the graph, the opposite is true. This situation shows that the dual-channel reasoning model tends to predict more evidence sentences than the only-sup model.

To further explore the advantages of the dual-channel reasoning model, we calculated Kendall's tau correlation between the number of predicted evidence sentences of the three models and the gold evidence sentences. As shown in Figure 9, the dual-channel reasoning model improves EM_{sup}, $F1_{sup}$, precision, recall, and Kendall's tau.

We also introduced the two strategies of CGDe and FGIn into the dual-channel reasoning architecture. Similarly, Figures 10 and 11, respectively, show the number of predictions minus the number of gold sentences of several models and the scores of all evaluation metrics.

In supporting facts prediction task, the model with the FGIn strategy performs best; the model with the CGDe strategy performs slightly lower than the other two models. The reason for this result is that CGDe decomposes complex questions so that it is easy to ignore evidence sentences, while FGIn can better represent each word in multiple documents. In contrast, the answer prediction ability of the model containing the CGDe strategy is significantly stronger than the other two models. Finally, the dual-channel inference model is not affected by the difference in the proportion of the two training losses when the two tasks are jointly optimized.

Tang et al. [30] used a neural decomposition model [25] to generate subquestions for multihop questions to explain the reasoning process of the question answering system to answer complex questions. In order to be able to further evaluate the ability of our proposed dual-channel reasoning architecture to perform true multihop reasoning, we evaluated the dual-channel reasoning model, the single-channel reasoning model, and the only-answer model on the subquestion datasets proposed by Tang et al.

As shown in Table 3, the complex question is decomposed into two subquestions. Tang et al. [30] divided complex questions in the HotpotQA verification set into two subquestions and extracted the answers of subquestion 1 from the original text. Then, they saved the answer to subquestion 1, subquestion 1, and all contexts to the JSON file Dev_sub1, and they saved the answer to subquestion 2 (also the final answer to the original complex question), subquestion 2, and all contexts to the JSON file Dev_sub2. The model is trained using the HotpotQA dataset and tested on three validation sets (Dev_ori, Dev_sub1, and Dev_sub2) to evaluate the ability of different models to answer subquestions.

As shown in Table 4, under the first three columns, *correct* represents the model answered correctly and *wrong* represents the model answered incorrectly. For example, the sixth line indicates that the model correctly answers the first subquestion but incorrectly answers the complex question and the second subquestion. For all experiments, we measure EM scores for question, $question_{sub1}$, and $question_{sub2}$ on 1,000 human-verified examples. When the answer predicted by the model is the same as the correct answer (both the start index and the end index are predicted correctly), the score is 1. The last three columns in Table 4 indicate the number of examples in the corresponding situation. For example, the number of examples in which the dual-channel model answers all complex questions and subquestions correctly is 282.

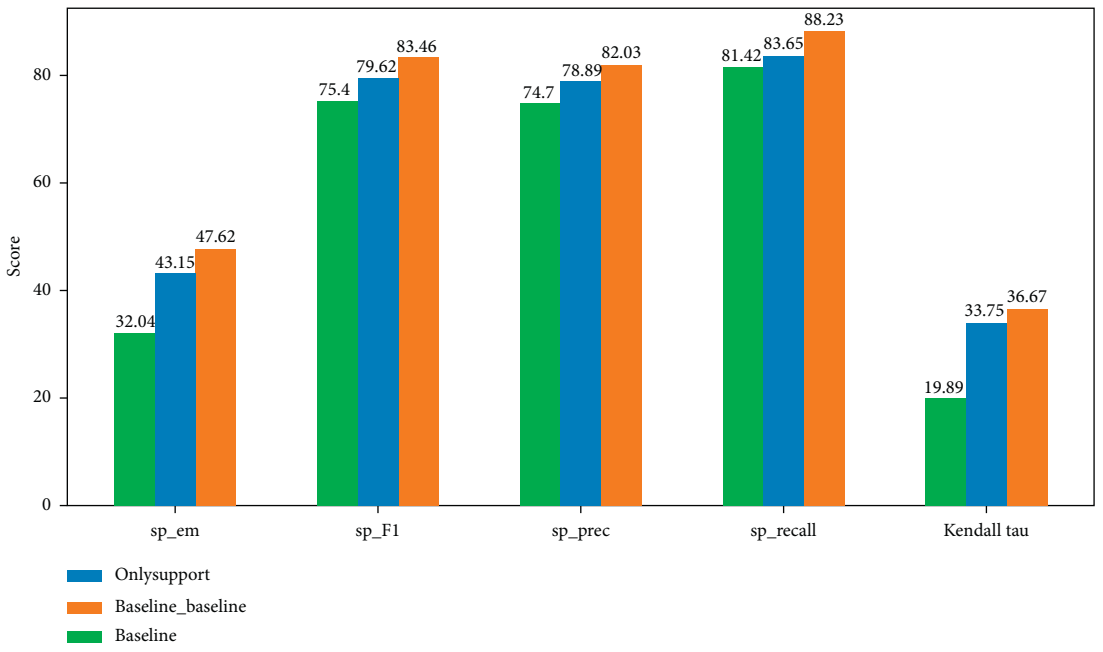

Figure 9: Evaluation metrics.

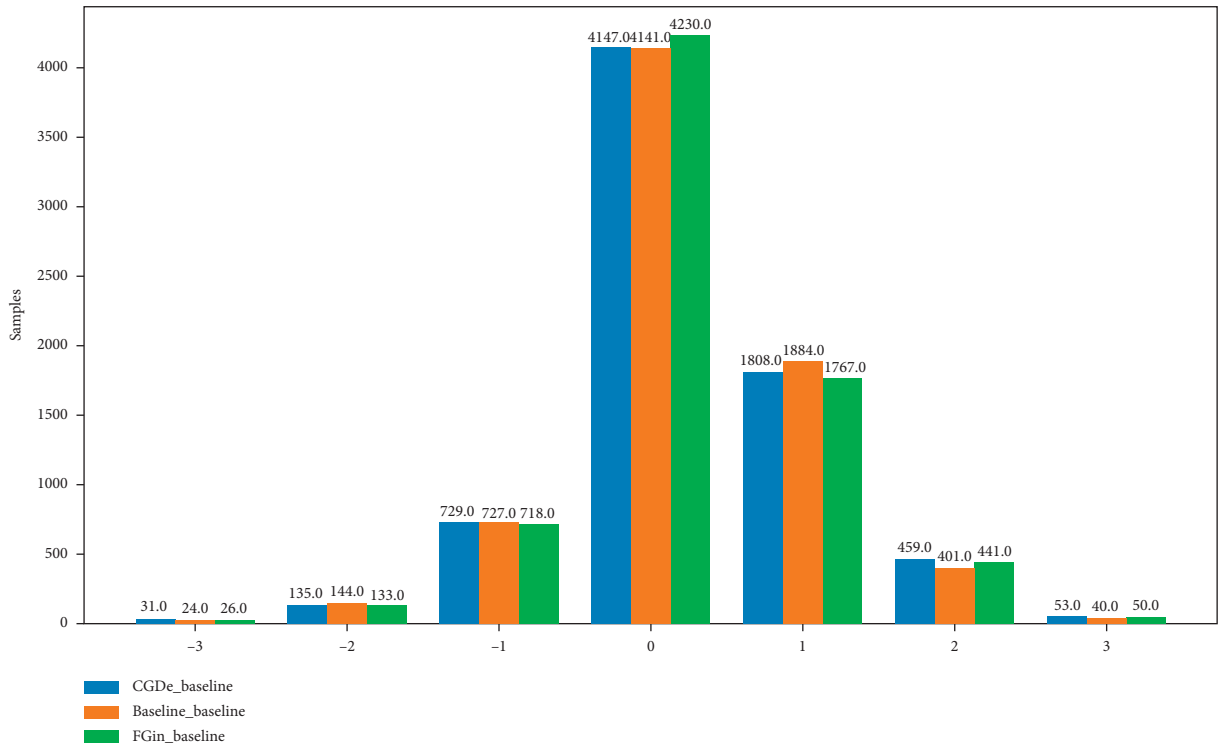

Figure 10: Number of predictions minus number of gold sentences (with two strategies).

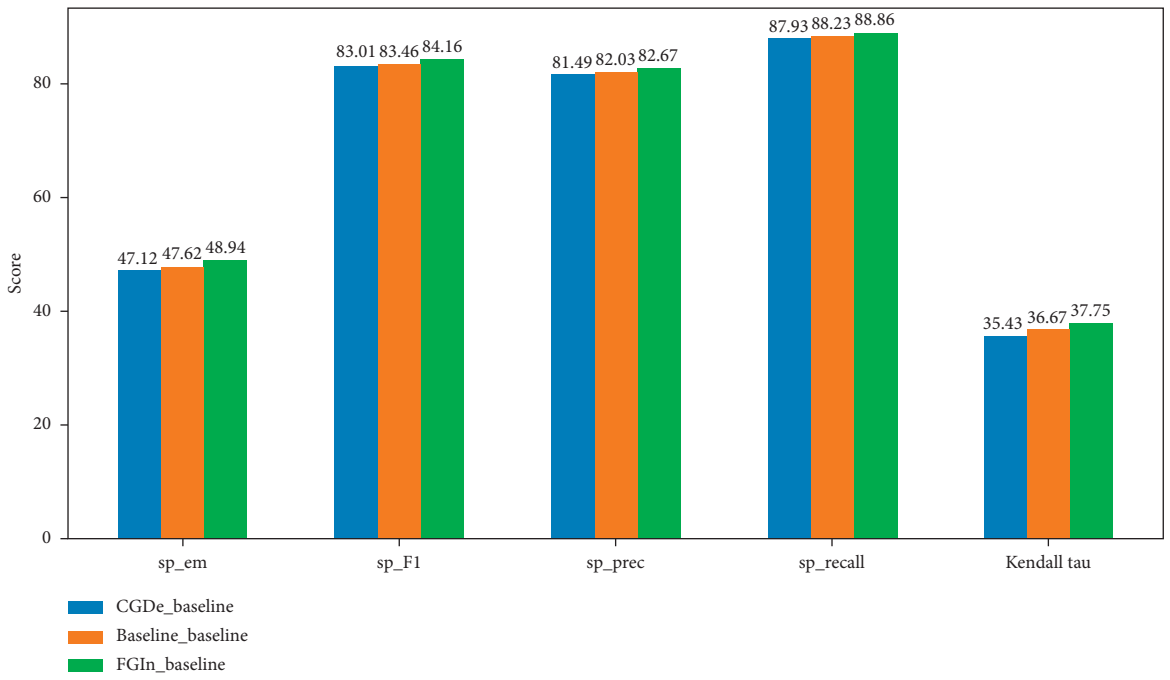

Figure 11: Evaluation metrics (with two strategies).

Table 3: An example in the subquestion dataset.

Dev_ori: Complex question: what government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?
Dev_sub1: Subquestion 1: which woman portrayed Corliss Archer in the film Kiss and Tell?
Dev_sub2: Subquestion 2: what government position was held by Shirley Temple?

Table 4: Categorical EM statistics (%) of subquestion evaluation for the three models.

Question	$question_{sub1}$	$question_{sub2}$	Baseline model	Dual-channel model	Only-ans model
Correct	Correct	Correct	26.7	**28.2**	26.3
Correct	Correct	Wrong	8.6	6.0	8.2
Correct	Wrong	Correct	14.6	14.7	17.4
Correct	Wrong	Wrong	4.8	5.4	4.3
Wrong	Correct	Correct	2.9	3.2	4.0
Wrong	Correct	Wrong	24.7	21.4	21.4
Wrong	Wrong	Correct	1.9	3.0	2.2
Wrong	Wrong	Wrong	15.8	18.1	16.2

As shown in Table 4, the dual-channel model has the largest number of examples that can correctly answer complex questions and subquestions. The model has the least number of examples when only one subquestion can be answered correctly, and the complex question is still answered correctly because this situation is not consistent with common sense.

6. Conclusion and Future Work

In this paper, we propose a dual-channel reasoning architecture for complex question answering. The dual-channel reasoning architecture is applied to the feature interaction framework and graph-based models to verify its general applicability. In the experiments, we show that our models

significantly and consistently outperform the baseline model, especially in supporting fact prediction tasks. After more detailed experimental analysis, it is proved that the dual-channel reasoning structure has stronger step-by-step reasoning ability than the single-channel reasoning structure. In the future, we believe that the following issue will be worth studying. For the dual-channel reasoning architecture, the interaction strategy between the two channels, such as the soft parameter sharing of the homogeneous neural network components of the two channels, is worthy of further study.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the National Key Research and Development Program of China (Grant no. 2018YFC0832304) and Fundamental Research Funds for the Central Universities (Grant no. 2020YJS012).

References

[1] L. Qin, X. Xu, W. Che, Y. Zhang, and T. Liu, "Dynamic fusion network for multi-domain end-to-end task-oriented dialog," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6344–6354, Seattle, WA, USA, April 2020.

[2] Y. Dai, H. Li, C. Tang, Y. Li, J. Sun, and X. Zhu, "Learning low-resource end-to-end goal-oriented dialog for fast and reliable system deployment," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 609–618, Seattle, WA, USA, July 2020.

[3] F. Xiong, X. Wang, S. Pan, H. Yang, H. Wang, and C. Zhang, "Social recommendation with evolutionary opinion dynamics," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 10, pp. 3804–3816, 2020.

[4] Y. Hu, F. Xiong, S. Pan, X. Xiong, L. Wang, and H. Chen, "Bayesian personalized ranking based on multiple-layer neighborhoods," *Information Sciences*, vol. 542, pp. 156–176, 2021.

[5] F. Xiong, W. Shen, H. Chen, S. Pan, X. Wang, and Z. Yan, "Exploiting implicit influence from information propagation for social recommendation," *IEEE Transactions on Cybernetics*, vol. 50, no. 10, pp. 4186–4199, 2020.

[6] X. Cao and Y. Liu, "Coarse-grained decomposition and fine-grained interaction for multi-hop question answering," *Journal of Intelligent Information Systems*, 2021, https://arxiv.org/abs/2101.05988.

[7] Y. Feldman and R. El-Yaniv, "Multi-hop paragraph retrieval for open-domain question answering," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2296–2309, Florence, Italy, July 2019.

[8] Y. Fu and Y. Liu, "CGSPN: cascading gated self-attention and phrase-attention network for sentence modeling," *Journal of Intelligent Information Systems*, vol. 56, no. 1, pp. 147–168, 2021.

[9] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, TX, USA, November 2016.

[10] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: unanswerable questions for SQuAD," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, Melbourne, Australia, July 2018.

[11] S. Reddy, D. Chen, and C. D. Manning, "CoQA: a conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019.

[12] Z. Yang, P. Qi, S. Zhang et al., "HotpotQA: a dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October 2018.

[13] Y. Jiang and M. Bansal, "Avoiding reasoning shortcuts: adversarial evaluation, training, and model development for multi-hop QA," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2726–2736, Florence, Italy, July 2019.

[14] Y. Fang, S. Sun, Z. Gan, R. Pillai, S. Wang, and J. Liu, "Hierarchical graph network for multi-hop question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8823–8838, November 2020.

[15] L. Qiu, Y. Xiao, Y. Qu et al., "Dynamically fused graph network for multi-hop reasoning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6140–6150, Florence, Italy, July 2019.

[16] K. Nishida, K. Nishida, M. Nagata et al., "Answering while summarizing: multi-task learning for multi-hop QA with evidence extraction," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2335–2345, Florence, Italy, July 2019.

[17] Y. Lan and J. Jiang, "Query graph generation for answering multi-hop complex questions from knowledge bases," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Seattle, WA, USA, July 2020.

[18] Y. Sun, L. Zhang, G. Cheng, and Y. Qu, "SPARQA: skeleton-based semantic parsing for complex questions over knowledge bases," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 8952–8959, New York, NY, USA, February 2020.

[19] Y. Chen, H. Li, Y. Hua, and G. Qi, "Formal query building with query structure prediction for complex question answering over knowledge base," in *International Joint Conference on Artificial Intelligence (IJCAI)*, Yokohama, Japan, July 2020.

[20] B. Fu, Y. Qiu, C. Tang, Y. Li, H. Yu, and J. Sun, "A survey on complex question answering over knowledge base: recent advances and challenges," 2020, https://arxiv.org/abs/2007.13069.

[21] Y. Chen, L. Wu, and M. J. Zaki, "Bidirectional attentive memory networks for question answering over knowledge bases," in *Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 2913–2923, Minneapolis, MN, USA, June 2019.

[22] K. Xu, Y. Lai, Y. Feng, and Z. Wang, "Enhancing key-value memory neural networks for knowledge based question answering," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2937–2947, Minneapolis, MN, USA, June 2019.

[23] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," 2016, https://arxiv.org/abs/1611.01603.

[24] V. Zhong, C. Xiong, N. S. Keskar, and R. Socher, "Coarse-grain fine-grain coattention network for multi-evidence question answering," 2019, https://arxiv.org/abs/1901.00603.

[25] S. Min, V. Zhong, L. Zettlemoyer, and H. Hajishirzi, "Multi-hop reading comprehension through question decomposition and rescoring," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6097–6109, Florence, Italy, June 2019.

[26] Y. Jiang and M. Bansal, "Self-assembling modular networks for interpretable multi-hop reasoning," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4464–4474, Hong Kong, China, November 2019.

[27] G. P. S. Bhargav, M. Glass, D. Garg et al., "Translucent answer predictions in multi-hop reading comprehension," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 7700–7707, New York, NY, USA, February 2020.

[28] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," 2018, https://arxiv.org/abs/1810.00826.

[29] M. Ding, C. Zhou, Q. Chen, H. Yang, and J. Tang, "Cognitive graph for multi-hop reading comprehension at scale," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2694–2703, Florence, Italy, July 2019.

[30] Y. Tang, H. T. Ng, and A. K. H. Tung, "Do multi-hop question answering systems know how to answer the single-hop sub-questions?," 2020, https://arxiv.org/abs/2002.09919.