2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

MgSFR: Multi-grained Semantic Fusion Retrieval for Multi-hop Reading Comprehension

Yingying Zhang, Bo Cheng(✉) and Yuli Chen

State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China
yingyingzhang@bupt.edu.cn

Abstract. Multi-hop Reading Comprehension (RC) has become a critical task in Natural Language Processing (NLP), requiring models to perform multiple reasoning steps and aggregate dispersed clues across multiple paragraphs to answer complex questions. Unlike single-hop RC, multi-hop RC aims to bridge information across diverse contexts and provide interpretable supporting facts, making it closer to human-like reasoning. To address these challenges, we propose a novel approach, Multi-grained Semantic Fusion Retrieval (MgSFR), which integrates semantic information from multiple granularities (word, phrase, sentence, and document). This fusion enhances the semantic relevance between questions and paragraphs, improving retrieval accuracy and efficiency. Additionally, MgSFR proposes a fine-grained semantic interaction mechanism that computes semantic similarity between different granularities, further boosting the model's performance. To complete the multi-hop RC pipeline, we introduce a multi-task reader that leverages this semantic fusion to enhance the model's reasoning capabilities. Experimental results on the HotpotQA benchmark dataset demonstrate that MgSFR significantly outperforms existing retrieval methods and provides high-quality context for multi-hop reasoning. Additionally, MgSFR achieves competitive performance compared to current state-of-the-art models in multi-hop reasoning tasks, validating its effectiveness in complex multi-hop tasks.

Keywords: Information Integration, Multi-grained Semantic Fusion, Multi-hop Reading Comprehension, Semantic Reasoning.

1 Introduction

1.1 A Subsection Sample

Machine Reading Comprehension (MRC) has gained significant attention in the field of Natural Language Processing (NLP), aiming to enable machines to read and answer questions based on textual information. With recent advancements in Deep Learning (DL), MRC has seen significant improvements, attracting attention from both academia and industry [1-3]. Datasets like SQuAD [4,5], TriviaQA [6], and CoQA [7] have become crucial benchmarks for MRC model development.

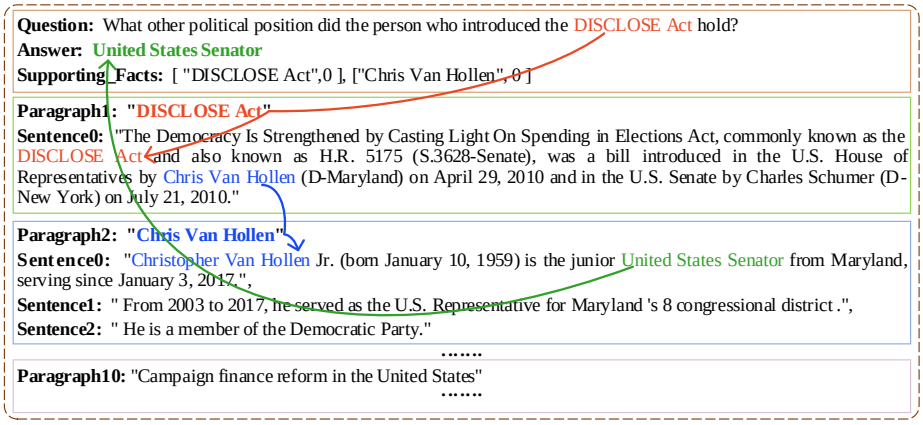

Fig. 1. An example from the HotpotQA dataset illustrating the multi-hop reasoning process. The model needs to predict the answer and supporting facts by reasoning across multiple sentences and paragraphs.

While early MRC research focused on single-hop reasoning [4-7], challenges arise when models need to reason across multiple paragraphs—an essential skill for answering complex, multi-hop questions where a single paragraph is insufficient [2,8,9]. To address this, multi-hop datasets such as NarrativeQA [10] and HotpotQA [11] have been developed, which require models to extract answers by reasoning across several supporting paragraphs. Among them, HotpotQA stands out as a widely used and challenging benchmark, featuring both relevant and irrelevant candidate paragraphs for each question. Fig. 1 presents an example from HotpotQA to demonstrate multi-hop reasoning.

Mainstream models for multi-hop RC typically adopt the "Retriever-Reader" architecture [12], where the retriever identifies candidate golden paragraphs relevant to the question while minimizing distracting information, and the reader jointly predicts supporting facts and extracts the answer span from the selected paragraphs. This architecture has achieved notable success, with models such as SAE [13], S2G [14], FE2H [15], Smoothing R^3 [8], and SuAN [9] leading the field. However, these approaches often rely on coarse-grained paragraph retrieval, offering word-level or document-level clues and failing to capture finer semantic dependencies, which may result in low-quality context for downstream tasks.

The primary challenge in multi-hop paragraph retrieval lies in selecting relevant paragraphs from a pool of candidates, where supporting facts are dispersed and often surrounded by noisy data. Traditional Sparse Retrieval (SR) methods, such as TF-IDF [16] and BM25 [17], focus on term matching and fail to account for semantic relationships between paragraphs. With the rise of deep learning and semantic-aware representations, several semantic retrieval models have been proposed with promising results. For example, DFGN [18] uses BERT-score as the relevance metric. HGN [19]introduces a paragraph ranker. FE2H [15]adopts a two-step approach to selecting relevant paragraphs. S2G [14] and Smoothing R^3 [8] focuses on a coarse-to-fine stepwise paragraph selection. While existing clue extraction methods exhibit strong performance in

retrieval, their limitations lie in providing only word-level or document-level clues and coarse-grained paragraph selection.

To overcome these limitations, we propose Multi-grained Semantic Fusion Retrieval (MgSFR), a novel model that enhances multi-hop paragraph retrieval by fusing semantic information across multiple granularity levels—word, phrase, sentence, and document—from both the question and the paragraphs. MgSFR addresses the challenge of identifying high-quality, noise-free context for downstream tasks and can serve as a powerful plugin in multi-hop RC systems. We integrate MgSFR into a "Retriever-Reader" architecture [12], where a multi-task reader jointly predicts supporting facts and extracts the answer span. We validate MgSFR extensively on the HotpotQA dataset, using the distractor setting for rigorous evaluation. Our experimental results demonstrate that MgSFR outperforms existing retrieval methods, significantly enhancing retrieved-context quality and boosting the performance of downstream reading comprehension tasks. Additionally, MgSFR achieves competitive results with current state-of-the-art models in multi-hop reasoning tasks.

We highlight the main contributions of our work as follows:

- We propose Multi-grained Semantic Fusion Retrieval (MgSFR), a novel paragraph retrieval method that fuses semantic information across multiple granularity levels (word, phrase, sentence, and document), improving retrieval performance.
- We introduce a Fine-grained Semantic Interaction Mechanism (FgSIM), which efficiently captures semantic similarities across different granularity encodings of questions and paragraphs to improve paragraph retrieval efficiency.
- We conducted comprehensive evaluations on the HotpotQA dataset to validate the effectiveness of the MgSFR model. The experimental results show that MgSFR significantly outperforms existing retrieval methods, providing high-quality context for multi-hop reasoning. Additionally, MgSFR achieves competitive results with current state-of-the-art models in multi-hop reasoning tasks.

2 Related Work

2.1 Multi-hop Reading Comprehension

Multi-hop Reading Comprehension (RC) is an advanced form of Machine Reading Comprehension (MRC) that requires models to integrate information from multiple paragraphs to answer complex questions [2,20]. This mirrors human reasoning and presents significant challenges in Natural Language Processing (NLP). Datasets like HotpotQA [11], NarrativeQA [10], $\mathcal{R}^4\mathcal{C}$ [21], and 2WikiMultiHopQA [22] exemplify these challenges, with HotpotQA being the most representative and challenging. It not only tests the model's ability to extract answer spans but also its capability to provide supporting facts for reasoning.

Multi-hop RC models are typically categorized into three types based on their reasoning strategies: (i) *Graph-based Models*, which use graphs and GNNs to propagate information across entities and sentences (e.g., DFGN [18], HGN [19], SAE [13], AMGN [23], and SuAN [9]). While powerful, they suffer from high complexity and

error propagation due to techniques like Named Entity Recognition (NER) and graph construction, making them computationally expensive. (ii) *Question Decomposition Models*, which break down complex questions into simpler sub-questions for single-hop models (e.g., DecompRC [24], ONUS [25], RERC [26], and GenDec [27]). However, generating high-quality sub-questions remains challenging. (iii) *Clue Extraction Models*, which iteratively or non-iteratively extract relevant clues to bridge the semantic gap between questions and context (e.g., QFE [28], S2G [14]). These models address the semantic gap between the question and context, making them more aligned with human reasoning. Consequently, we adopt the clue-based reasoning approach, utilizing the "Retriever-Reader" architecture [12] for our novel multi-hop RC model, which aims to overcome these limitations while enhancing scalability and retrieval accuracy.

2.2 Paragraph Retrieval for Multi-hop RC

The primary goal of paragraph retrieval is to minimize redundant and noisy information while selecting question-relevant paragraphs for downstream tasks. Previous work has demonstrated that effective paragraph retrieval significantly improves the performance of RC systems. DFGN [18] and QUARK [29] respectively used BERT-score and sentence relevance scoring to rank paragraphs. However, these models overlook the semantic relationships between paragraphs, which limits their retrieval effectiveness. HGN [19] introduced a paragraph ranker to improve paragraph cascading. FE2H [15] proposed a two-stage retrieval method that iteratively selects paragraphs. S2G [14] and Smoothing R^3 [8] employed Coarse-to-Fine cascade retrieval approach to step-by-step select relevant paragraphs. Despite these advances, most existing methods are restricted to retrieval at the word or document level, which limits their ability to capture the semantic depth of the context.

To overcome these limitations, we introduce the Multi-grained Semantic Fusion Retrieval (MgSFR) framework. MgSFR enhances paragraph retrieval by fusing semantic information across multiple granularities—word, phrase, sentence, and document—captured from both the question and the paragraphs. This multi-grained approach ensures a more comprehensive and contextually relevant retrieval, ultimately improving performance in multi-hop RC tasks. By integrating semantic content from varying granularities, MgSFR significantly boosts retrieval accuracy, providing a more robust foundation for reasoning and answer prediction in multi-hop RC.

3 Proposed Model

In multi-hop RC tasks, each question Q is typically paired with a set of paragraphs $P_{ara} = \{P_{ara}^1, \cdots, P_{ara}^M\}$, where M denotes the total number of paragraphs. However, not all paragraphs are relevant to the question, so an efficient retrieval mechanism is essential for identifying relevant paragraphs. As illustrated in Fig. 2, our approach follows the "retrieve-and-reader" pipeline [12], where the paragraph retriever first filters the relevant paragraphs to form high-quality context C. This context is then processed by the multi-task reader to jointly predict the answer and supporting facts.

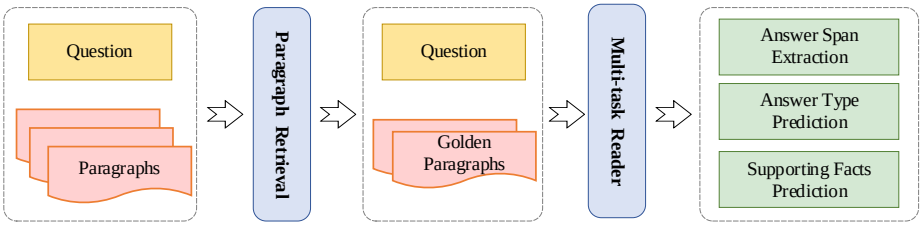

Fig. 2. Overview of the mainstream pipeline for Multi-hop RC tasks.

Paragraph Retrieval: Given a question Q and a set of paragraphs P_{ara}, the goal is to filter out irrelevant information and identify the relevant golden paragraphs, denoted as $P'_{ara} = \{P_{ara}^{gold1}, P_{ara}^{gold2}, \cdots\}$, which will provide a high-quality context C for downstream processing.

Multi-Task Reader: Given a question Q and its relevant golden paragraphs P'_{ara}, the task is to predict both the answer span and the supporting fact sentences that are related to the question Q.

3.1 MgSFR Paragraph Retrieval

As shown in Fig. 3, we propose a novel and effective Multi-grained Semantic Fusion Retrieval (MgSFR) approach that enhances paragraph retrieval by fusing semantic information across multiple granularity levels (word, phrase, sentence, and document) in both questions and paragraphs, ensuring high-quality context for improved multi-hop reasoning. The MgSFR model consists of several key components: **Multi-grained Semantic Encoder:** Encodes the question and paragraph at four granularities (word, phrase, sentence, and document) using pre-trained models. Multi-head attention mechanisms are applied across granularities to enhance semantic interactions. **Multi-grained Semantic Interaction:** Computes the relevance between question and paragraph at each granularity using the Fine-grained Semantic Interaction Mechanism (FgSIM), which calculates similarity scores at the word, phrase, sentence, and document levels. **Multi-grained Semantic Fusion:** Aggregates the relevance scores from all granularities into a final score that reflects the overall relevance between the question and the paragraph.

Multi-grained Semantic Encoder. In the multi-grained semantic encoding process, we extract semantic units at four levels(word, phrase, sentence, and document) in both questions and paragraphs. Special tokens are added to distinguish between the question and the paragraph: token $[Q]$ is prepended to the question, and token $[P_{ara}]$ is prepended to the paragraph.

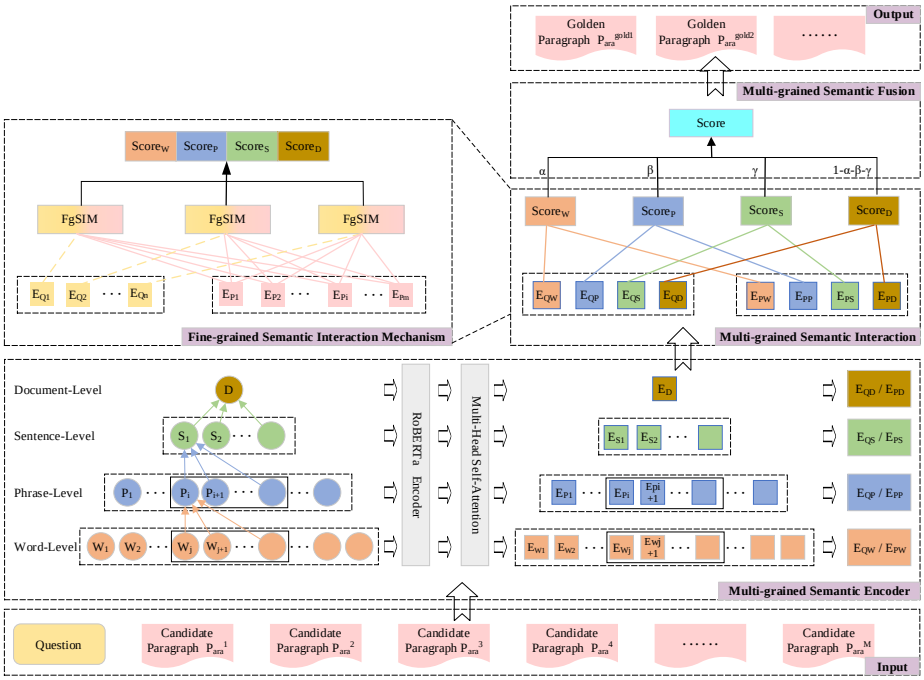

Fig. 3. Overview of the MgSFR model architecture. The architecture integrates semantic information across multiple granularities: word (W), phrase (P), sentence (S), and document (D).

Question Semantic Encoder. For the question Q, we extract semantic units at the four granularity levels. The multi-grained semantic unit for the question is constructed as follows:

$$W_q = [CLS][Q]w_1 w_2 \cdots w_{l_w}\#\#\# \tag{1}$$

$$P_q = [CLS][Q]p_1 p_2 \cdots p_{l_p}\#\#\# \tag{2}$$

$$S_q = [CLS][Q]s_1 s_2 \cdots s_{l_s}\#\#\# \tag{3}$$

$$D_q = [CLS][Q]d\#\#\# \tag{4}$$

where W_q, P_q, S_q, and D_q represent the word, phrase, sentence, and document-level semantic unit sequences of the question. $[CLS]$ denotes the start token, and Q denotes the input sequence is a question. $w_1 w_2 \cdots w_{l_w}$ denotes that the WordPiece model [30] tokenized word sequences, $p_1 p_2 \cdots p_{l_p}$ denotes the sequence of phrases extracted by the phrase extractor, $s_1 s_2 \cdots s_{l_s}$ stands for the sequence of sentences, and d represents the document sequence. # denotes the [mask] token used for question enhancement. Specifically, we define a fixed query length l_q. If the actual question is shorter than l_q, we pad it with special [mask] tokens to reach the standard length. These [mask] tokens serve a dual purpose. First, they act as placeholders to ensure uniform input dimensions

for batched processing. More importantly, they function as semantic expansion markers. During training, the model learns to interpret the [mask] tokens as cues for query augmentation, allowing it to implicitly infer common suffix patterns or contextual extensions that often follow such queries. This mechanism enhances the expressiveness of shorter questions, mitigates semantic ambiguity, and ultimately improves their semantic alignment with relevant document content.

To enhance the semantic representation of input text, we incorporate a CNN layer that captures local n-gram features from the embeddings generated by RoBERTa and refined by multi-head self-attention. This process allows the model to focus on meaningful patterns within the text while suppressing irrelevant noise. The multi-grained semantic units are encoded as follows:

$$E_{Wq} = Normalize(CNN(MHSA(RoBERTa(W_q)))) \tag{5}$$

$$E_{Pq} = Normalize(CNN(MHSA(RoBERTa(P_q)))) \tag{6}$$

$$E_{Sq} = Normalize(CNN(MHSA(RoBERTa(S_q)))) \tag{7}$$

$$E_{Dq} = Normalize(CNN(MHSA(RoBERTa(D_q)))) \tag{8}$$

where E_{Wq}, E_{Pq}, E_{Sq}, and E_{Dq} denote the embedding representations of the question at the word, phrase, sentence, and document levels, respectively.

Paragraph Semantic Encoder. The paragraph encoding follows a similar process. Unlike the question, no [mask] tokens are added because the paragraph length does not need expansion. The multi-grained semantic unit for the paragraph is constructed as follows:

$$W_p = [CLS][P_{ara}]w_1 w_2 \cdots w_{l_w} \tag{9}$$

$$P_p = [CLS][P_{ara}]p_1 p_2 \cdots p_{l_p} \tag{10}$$

$$S_p = [CLS][P_{ara}]s_1 s_2 \cdots s_{l_s} \tag{11}$$

$$D_p = [CLS][P_{ara}]d \tag{12}$$

where W_p, P_p, S_p, and D_p respectively denote the word, phrase, sentence, and document-level semantic unit sequences for the paragraph. $[CLS]$ denotes the start token, and $[P_{ara}]$ denotes that the input sequence belongs to a paragraph. Unlike questions, we do not append [mask] tokens to the paragraph because the paragraph does not face the issue of being too short in length; hence, semantic expansion is not required.

For paragraph inputs, we introduce a punctuation filtering step before applying the CNN layer. This step eliminates unnecessary punctuation marks, ensuring that the model's attention is directed towards the core content, improving both retrieval efficiency and the overall accuracy of the model in capturing relevant information. The embeddings for the paragraph are encoded similarly to the question:

$$E_{Wp} = Filter(Normalize(CNN(MHSA(RoBERTa(W_P))))) \tag{13}$$

$$E_{Pp} = Filter(Normalize(CNN(MHSA(RoBERTa(P_P)))))) \quad (14)$$

$$E_{Sp} = Filter(Normalize(CNN(MHSA(RoBERTa(S_P)))))) \quad (15)$$

$$E_{Dp} = Filter(Normalize(CNN(MHSA(RoBERTa(D_P)))))) \quad (16)$$

where E_{Wp}, E_{Pp}, E_{Sp}, and E_{Dp} denote the embedding representations of the paragraph at the word, phrase, sentence, and document levels, respectively.

This encoding process ensures both question and paragraph are represented with rich semantic information across multiple granularities, which is essential for effective multi-hop reasoning in downstream tasks.

Multi-grained Semantic Interaction. In the multi-grained semantic interaction stage, we perform semantic relevance calculations on embeddings from different granularities obtained in the previous encoding phase. The primary objective of this stage is to compare and align the semantic representations extracted at various granularities, thereby facilitating more effective retrieval. We employ the Fine-grained Semantic Interaction Mechanism (FgSIM), which computes relevance scores between the question and paragraph at the word, phrase, sentence, and document levels. The interaction is based on cosine similarity between the embeddings, which measures the strength of the match between the question and paragraph.

$$Score_* = \sum_i^m \underset{j \le n}{FgSIM}\left(E_{qi} \cdot E_{pj}\right) \quad (17)$$

where $Score_*$ denotes the relevance score between the question and paragraph at different levels. When $*= W$, it indicates the score of the question and paragraph at the word level, m and n respectively indicate the number of semantic units in the question and paragraph at the word level. When $*= P$, it indicates the score at the phrase level, with m and n representing the number of semantic units in the question and paragraph at the phrase level. When $*= S$ or $*= D$, they represent the scores at the sentence and document levels, respectively.

This multi-grained interaction mechanism significantly improves the model's ability to retrieve paragraphs that are most relevant to the question, thereby providing a better foundation for subsequent reasoning and understanding.

Multi-grained Semantic Fusion. Relying on a single-granularity semantic representation for paragraph retrieval may lead to the omission of critical relevance signals, as different questions and contexts may align more strongly with different semantic levels. To address this, we design a multi-grained semantic fusion mechanism that aggregates relevance scores computed at four granularity levels—word, phrase, sentence, and document—into a final retrieval score.

Rather than assigning fixed weights to each granularity, we introduce an adaptive fusion strategy based on learnable parameters. Specifically, we assign trainable raw scores α, β, and γ to the word-, phrase-, and sentence-level representations, respectively. These scores are normalized using a softmax function to obtain the final fusion

weights $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{\gamma}$, while the residual weight is automatically allocated to the document level. The fused relevance score is computed as follows:

$$Score = \tilde{\alpha} \cdot Score_W + \tilde{\beta} \cdot Score_P + \tilde{\gamma} \cdot Score_S + (1 - \tilde{\alpha} - \tilde{\beta} - \tilde{\gamma}) \cdot Score_D, \quad (18)$$

$$[\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}] = softmax([\alpha, \beta, \gamma])$$

where $Score$ denotes the final fused relevance score. The parameters α, β, and γ are learnable scalar values that capture the model's preference for each semantic level. The normalized weights $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{\gamma}$ reflect the relative contribution of each granularity and are dynamically adjusted during training to optimize retrieval effectiveness. This adaptive fusion mechanism enables the model to learn task- and query-specific weighting patterns, thereby improving both flexibility and generalization across diverse multi-hop reasoning scenarios.

Inspired by works like S2G [14] and FE2H [15], we adopt a two-step retrieval strategy. First, the question and each candidate paragraph are processed by MgSFR to identify the most relevant paragraph. Then, this selected paragraph is combined with the question to form a new query, which is re-entered into MgSFR to retrieve the second-hop relevant paragraphs. The final high-quality context is formed by concatenating the retrieved paragraphs and is passed to the downstream reader module.

3.2 Multi-task Reader

The Multi-task Reader module utilizes a multi-task learning framework to simultaneously extract answer spans and predict supporting fact sentences.

Question and Context Encoding. For encoding, we first construct the input sequence for RoBERTa. Since the HotpotQA dataset includes both "Yes/No" questions and span-based questions, we prepend special tokens for "yes" and "no" to the sequence. The input format is as follows: "[CLS]+yes+no+question+ [SEP]+context+[SEP]". This sequence is passed through RoBERTa, generating initial representations. The resulting output is represented as $H = [h_1, \cdots, h_l] \in R^{l \times d_1}$, where l is the sequence length and d_1 is the hidden vector dimension. The context and question embeddings are denoted by $H_c = [h_1, \cdots, h_{l_c}] \in R^{l_c \times d_1}$ and $H_q = [h_{l_c+1}, \cdots, h_l] \in R^{l_q \times d_1}$, respectively, where l_c is the context length and $l_q = l - l_c$ is the question length.

$$H_q, H_c = RoBERTa(Q, C) \quad (19)$$

To further enhance performance, we apply a bi-attention mechanism to model interactions between the question and the context, refining the representations. The final output representations are denoted as $\widehat{H} \in R^{l \times d_2}$, with the context and question representations as $\widehat{H}_c \in R^{l_c \times d_2}$ and $\widehat{H}_q \in R^{l_q \times d_2}$, where d_2 is the output embedding dimension.

$$\widehat{H}_q, \widehat{H}_c = Bi - Attention(H_q, H_c) \quad (20)$$

This bi-attention mechanism is crucial for enabling multi-hop reasoning across paragraphs, allowing the model to learn richer interactions between the question and context.

Multi-task Prediction. Inspired by methods like DFGN [18], HGN [19], and FE2H [15], we use multi-task learning to jointly predict the answer span and supporting facts. The prediction task is divided into three sub-tasks: (i) Answer Start Position Extraction: identifying the start position of the answer span; (ii) Answer End Position Extraction: identifying the end position of the answer span; (iii) Supporting Facts Prediction: determining whether a sentence in the selected paragraph is a supporting fact.

For answer span extraction, we apply linear prediction layers to predict the start and end positions of the answer. We use the cross-entropy loss for these tasks, denoted as L_{start} and L_{end}. For supporting facts prediction, we use a binary classifier, with cross-entropy loss denoted as L_{sup}.

$$L_{start} = CrossEntropy(o_{start}, \hat{o}_{start}) \tag{21}$$

$$L_{end} = CrossEntropy(o_{end}, \hat{o}_{end}) \tag{22}$$

$$L_{sup} = CrossEntropy(o_{sup}, \hat{o}_{sup}) \tag{23}$$

where $\hat{o}_{start}$ and $\hat{o}_{end}$ denote the predicted probability distributions for the start and end positions, and $\hat{o}_{sup}$ denotes the predicted probability distribution for supporting fact sentences.

The final loss function is a weighted sum of the individual losses:

$$L_{joint} = \lambda_1 L_{start} + \lambda_2 L_{end} + \lambda_3 L_{sup} \tag{24}$$

where λ_1, λ_2, and λ_3 are hyperparameters that control the relative importance of each task. Typically, we set $\lambda_1 = \lambda_2 = \lambda_3 = 1$ to treat all tasks equally.

4 Experiments

4.1 Experimental Settings

Dataset. We evaluated our method on the *distractor* setting of the HotpotQA [11] dataset. HotpotQA is a popular multi-hop RC dataset and the first interpretable multi-hop RC dataset with sentence-level evidence. HotpotQA comprises 113K question-answer pairs sourced from Wikipedia and crowdsourcing. In the *distractor* setting, each question has two golden paragraphs with ground-truth answers and supporting facts and eight "distractor" paragraphs retrieved from Wikipedia via bigram TF-IDF [31] based on the question.} HotpotQA also provides supporting facts for each question, encouraging models to explain the reasoning paths in multi-hop RC. Table 1 shows the details of the dataset, which contains 90,564 examples in the train set, 20% of which are easy

single-hop questions, 80% of which are hard multi-hop questions, and 7,405 questions and multi-hop examples in both the dev and test sets.

Table 1. Statistics of dataset.

Name	Level	Description	Usage	Examples
Train	easy	single-hop	training	18089
	medium	multi-hop		56814
	hard	multi-hop		15661
Dev	hard	multi-hop	dev	7405
Test	hard	multi-hop	test	7405

Evaluation Metrics. The HotpotQA dataset consists of two sub-tasks: Answer Span Prediction (Ans) and Supporting Facts Prediction (Sup). For each sub-task, model performance is evaluated using Exact Match (EM) and Partial Match (F1) scores. We also use Joint EM and F1 scores to measure the overall performance of the model, encouraging accuracy in both sub-tasks for each example. Since the test set is unavailable, the reported performance is obtained by submitting our best model on the development set.

Baseline Models. To evaluate the effectiveness of our proposed MgSFR model, we compare it with two categories of baseline models:

Graph-based reasoning models: **DFGN** [18] proposes a dynamic fusion graph for multi-hop reasoning along entity connections. **SAE-large** [13] constructs an inference graph with sentence nodes, where contextual sentence embeddings are used as graph nodes. **KIFGraph** [32] integrates structured knowledge and contextual information across multiple granularity levels using asynchronous node updates. **BFR-Graph** [33] proposes a breadth-first reasoning graph that propagates reasoning step-by-step through sentence nodes. **HGN-large** [19] constructs a hierarchical graph that integrates multiple granularity levels (question, paragraph, sentence, and entity) for reasoning. **SuAN** [9] proposes a subgraph-based attention network that sequentially updates nodes based on the subgraph order. **AMGN+** [23] utilizes an asynchronous multi-grained graph with entity, sentence, and paragraph nodes to simulate human-like multi-hop reasoning.

Clue extraction-based reasoning models: **C2F Reader** [34] points out that graph attention is a special form of self-attention, validating that graph structures are not necessary for multi-hop reasoning. **FE2H** [15] adopts a two-stage paragraph retriever and reader from easy to hard. **S2G+EGA** [14] applies a Coarse-to-Fine paragraph retrieval process with Sentence-aware Self-Attention and evidence-guided attention for reasoning. **Smoothing R^3** [8] employs the Coarse-to-Fine retrieval and label-smoothing technique to extract both answer spans and supporting facts.

4.2 Implementation Details

The model is implemented based on Pre-trained Language Models (PLMs) from the Hugging Face Transformers library [35]. The framework is built on Pytorch and trained on Tesla A40 GPUs. The following provides implementation details for the paragraph retrieval and multi-task reader modules.

MgSFR Paragraph Retrieval. For the paragraph retrieval module, we utilize both the base **(-base)** and the large **(-large)** versions of RoBERTa [36] to obtain contextual embeddings for questions and paragraphs at multiple granularity levels. Due to resource constraints, ablation studies on paragraph retrieval are conducted solely using RoBERTa-base. All experiments are run on Tesla A40 GPUs, with the maximum number of epochs set to 16 and the batch size set to 8. For the optimizer, we use the BERT-Adam optimizer with a learning rate of 2e-5. We employ the WordPiece [30] tokenizer to tokenize words and T5 [37] (trained for phrase extraction tasks) to recognize phrases in questions and paragraphs.

Multi-task Reader. For the multi-task reader module, we employ the large **(-large)** version of RoBERTa [36] and the XXLarge **(-xxlarge)** version of DeBERTa [38] as PLMs. Experiments are conducted on Tesla A40 GPUs, using the AdamW optimizer with a learning rate of 1e-5, a warm-up rate of 0.1, and an L2 weight decay of 1e-2. For the RoBERTa-large model, we set the number of epochs to 16 and the batch size to 8. For the DeBERTa-v2-xxlarge model, due to its larger number of parameters, we set the number of epochs to 12 and the batch size to 4.

4.3 Experimental Results

MgSFR Retrieval Performance. We compared our MgSFR retrieval method with several previous retrieval models, including HGN [19], SAE [13], S2G [14], FE2H [15] and Smoothing R^3 [8], on the distractor setting of the development set in the multi-hop dataset HotpotQA. These models also employed carefully designed retrieval techniques to identify relevant paragraphs required for downstream tasks from the set of candidate paragraphs. The retrieval module was trained using the base and large versions of RoBERTa as PLMs and evaluated using EM and F1 scores. As shown in Table 2, the experimental results indicate that our MgSFR method significantly outperformed all other methods across all metrics, validating its effectiveness in enhancing retrieval performance through a multi-grained semantic fusion strategy. Notably, even the base version of MgSFR achieved competitive performance comparable to state-of-the-art retrieval methods, further confirming its practicality and effectiveness in multi-grained semantic retrieval tasks.

Table 2. Comparison of retrieval performance on the HotpotQA dev set with previous work.

Models	Retrieval	
	EM	F1
HGN [19]	-	94.53
SAE-large [13]	91.98	95.76
S2G-large [14]	95.77	97.82
FE2H-large [15]	96.32	98.02
Smoothing R^3 [8]	96.85	98.32
MgSFR on RoBERTa-base(Ours)	96.65	98.21
MgSFR on RoBERTa-large(Ours)	**97.32**	**98.56**

Table 3. Performance comparison on HotpotQA dataset in the distractor setting. MgSFR achieves overall competitive performance and outperforms all graph-based reasoning models.

Models	Ans		Sup		Joint	
	EM	F1	EM	F1	EM	F1
Official Baseline						
Baseline Model [11]	45.60	59.02	20.32	64.49	10.83	40.16
Graph-based Reasoning Models						
DFGN [18]	56.31	69.69	51.50	81.62	33.62	59.82
SAE-large [13]	66.92	79.62	61.53	86.86	45.36	71.45
KIFGraph [32]	69.53	82.42	61.79	87.98	46.49	74.12
BFR-Graph [33]	70.06	82.20	61.33	88.41	45.92	74.13
HGN-large [19]	69.22	82.19	62.76	88.47	47.11	74.21
SuAN [9]	69.55	82.71	63.43	88.82	47.44	74.92
AMGN+ [23]	70.53	83.37	63.57	88.83	47.77	75.24
Clue Extraction-based Reasoning Models						
C2F Reader [34]	67.98	81.24	60.81	87.63	44.67	72.73
FE2H on ELECTRA [15]	69.54	82.69	64.78	88.71	48.46	74.90
S2G+EGA [14]	70.92	83.44	63.86	88.68	48.76	75.47
R^3 [8]	71.27	83.57	65.25	88.98	49.81	76.02
FE2H on ALBERT [15]	71.89	**84.44**	64.98	89.14	**50.04**	76.54
Smoothing R^3 [8]	**72.07**	84.34	65.44	89.55	49.73	**76.69**
MgSFR on RoBERTa(ours)	71.35	84.13	65.94	90.12	49.36	76.16
MgSFR on DeBERTa(ours)	71.95	84.36	**66.85**	**90.51**	49.92	76.61

Multi-task Reading Comprehension Performance. We conducted a comprehensive evaluation of the MgSFR model on the development set of the HotpotQA dataset in the distractor setting (since the test set is not publicly available) and compared its perfor-

mance against other strong multi-hop RC models listed in Table 3. As shown in the experimental results in Table 3, the MgSFR model achieved improvements of 1.41 and 0.96 in EM and F1 scores for supporting facts prediction, respectively, surpassing the previously best-performing Smoothing R^3 [8] model and achieving state-of-the-art performance. These results demonstrate that MgSFR effectively enhances multi-hop RC tasks by improving retrieval performance and providing higher-quality context for downstream reasoning. Furthermore, MgSFR consistently outperformed existing multi-hop graph reasoning models across all evaluation metrics, further validating the effectiveness of the multi-grained semantic fusion strategy in capturing complex semantic relationships and improving reasoning capabilities. Utilizing DeBERTa-v2-xxlarge [38] as the PLM further enhanced performance, highlighting the pivotal role of powerful pre-trained models in addressing complex retrieval and reasoning tasks and underscoring the applicability of our approach in multi-hop RC scenarios.

4.4 Ablation Studies

To evaluate the contribution of each semantic granularity in our MgSFR framework, we perform an ablation study using RoBERTa-base [36]. We remove each granularity (word, phrase, sentence, document) individually and measure its impact on retrieval performance.

As shown in Table 4, the removal of any granularity leads to a noticeable drop in performance, confirming the effectiveness of multi-granular semantic fusion. Notably, removing phrase-level semantics causes the most substantial performance decline (↓3.41 F1), significantly larger than the impact of removing any other level. This result underscores the central role of phrase-level representations in multi-hop paragraph retrieval. Unlike word-level embeddings, which may lack context, or sentence/document-level semantics, which can be overly coarse, phrases provide semantically compact and contextually meaningful units—such as named entities, noun phrases, and short relational expressions—that are often directly aligned with the question's focus and critical for bridging reasoning steps. In this sense, phrase-level information forms a semantic pivot between fine-grained and coarse-grained signals, playing a unique and indispensable role in identifying supporting evidence.

Table 4. Ablation study results on the dev set of HotpotQA.

Models	Retrieval	
	EM	F1
MgSFR on RoBERTa-base(Ours)	**96.65**	**98.21**
w/o word-level	96.24 ↓0.41	97.96 ↓0.25
w/o phrase-level	90.61 ↓6.04	94.80 ↓3.41
w/o sentence-level	94.44 ↓2.21	96.96 ↓1.25
w/o document-level	96.12 ↓0.53	97.83 ↓ 0.38

4.5 Analysis of Adaptive Fusion Weights

To further evaluate the effectiveness of our proposed adaptive fusion strategy, we conducted an empirical analysis of the learned fusion weights across different types of questions in the HotpotQA development set. Specifically, we grouped questions based on their leading interrogatives (e.g., "Who", "What", "When", "Yes/No"), and recorded the average values of the normalized weights $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{\gamma}$ for the word-, phrase-, and sentence-level semantics, respectively. The remaining weight was assigned to the document level. As shown in Table 5, the model consistently assigned the highest weights to phrase-level semantics across most question types, particularly for "Who" and "What" questions that rely heavily on fine-grained entity and attribute matching. Interestingly, Yes/No questions tended to place greater emphasis on sentence-level representations, reflecting the need for more contextual or logical consistency in such cases. These results demonstrate that the adaptive fusion mechanism allows the model to dynamically prioritize semantic granularities that are most relevant to the question type, thereby improving both retrieval accuracy and interpretability. This observation aligns with our ablation study, which showed that removing phrase-level semantics led to the largest performance drop, confirming the importance of phrase-level signals in multi-hop reasoning.

Table 5. Average learned fusion weights by question type.

Question Type	Word ($\tilde{\alpha}$)	Phrase ($\tilde{\beta}$)	Sentence ($\tilde{\gamma}$)	Document
Who	0.23	**0.45**	0.21	0.13
When	0.29	**0.37**	0.18	0.16
What	0.25	**0.41**	0.22	0.12
Yes/No	0.19	**0.34**	0.33	0.14

5 Conclusion

In this paper, we propose a novel and effective Multi-grained Semantic Fusion Retrieval (MgSFR) model to tackle the challenges of multi-hop RC. The MgSFR model significantly enhances the semantic relationships between questions and paragraphs by fusing semantic information across different granularities (word, phrase, sentence, and document). This approach mitigates the risk of semantic information loss by ensuring a comprehensive understanding of both the question and the context, improving retrieval accuracy and multi-hop reasoning performance. Experimental results on the HotpotQA benchmark dataset demonstrate that the MgSFR model not only enhances paragraph retrieval performance in multi-hop reasoning tasks but also excels in multi-task reading comprehension, particularly in supporting fact prediction. These results highlight the model's superior semantic reasoning capabilities. In the future, we plan to explore joint optimization and answer-aware refinement strategies to enable dynamic interaction between the retriever and reader, thereby further improving the synergy between retrieval and reasoning.

Acknowledgments. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFF0902701; in part by the National Natural Science Foundation of China under Grants U21A20468, 62372058, U22A2026.

References

1. Han, W., Ouyang, Z., Wang, Y., Zheng, W.: Question answering over knowledge graphs via machine reading comprehension. In: DASFAA. pp. 577–594. Springer (2023)
2. Mavi, V., Jangra, A., Jatowt, A., et al.: Multi-hop question answering. Foundations and Trends® in Information Retrieval **17**(5), 457–586 (2024)
3. Mihaylov, T.B.: Knowledge-Enhanced Neural Networks for Machine Reading Comprehension. Ph.D. thesis (2024)
4. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for squad. In: ACL. pp. 784–789 (2018)
5. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. In: EMNLP. pp. 2383–2392 (2016)
6. Joshi, M., Choi, E., Weld, D.S., Zettlemoyer, L.: Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In: ACL. pp. 1601–1611 (2017)
7. Reddy, S., Chen, D., Manning, C.D.: Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics **7**, 249–266 (2019)
8. Yin, Z., Wang, Y., Hu, X., Wu, Y., Yan, H., Zhang, X., Cao, Z., Huang, X., Qiu, X.: Rethinking label smoothing on multi-hop question answering. In: China National Conference on Chinese Computational Linguistics. pp. 72–87 (2023)
9. Gong, C., Wei, Z., Wang, X., Wang, R., Li, Y., Zhu, P.: Subgraph-based attention network for multi-hop question answering. In: International Joint Conference on Neural Networks. pp. 1–9 (2024)
10. Kočiskỳ, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K.M., Melis, G., Grefenstette, E.: The narrativeqa reading comprehension challenge. Transactions of the Association for Computational Linguistics, 317–328 (2018)
11. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., Manning, C.D.: Hotpotqa: A dataset for diverse,explainable multi-hop question answering. In: EMNLP. pp. 2369–2380 (2018)
12. Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., Chua, T.S.: Retrieving and reading: A comprehensive survey on open-domain question answering. arXiv preprint arXiv:2101.00774 (2021)
13. Tu, M., Huang, K., Wang, G., Huang, J., He, X., Zhou, B.: Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In: AAAI. vol. 34, pp. 9073–9080 (2020)
14. Wu, B., Zhang, Z., Zhao, H.: Graph-free multi-hop reading comprehension: A select-to-guide strategy. arXiv preprint arXiv:2107.11823 (2021)
15. Li, X.Y., Lei, W.J., Yang, Y.B.: From easy to hard: Two-stage selector and reader for multi-hop question answering. In: ICASSP. pp. 1–5 (2023)
16. Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. vol. 242, pp. 29–48 (2003)
17. Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends®in Information Retrieval **3**(4), 333–389 (2009)
18. Qiu, L., Xiao, Y., Qu, Y., Zhou, H., Li, L., Zhang, W., Yu, Y.: Dynamically fused graph network for multi-hop reasoning.In: ACL. pp. 6140–6150 (2019)

19. Fang, Y., Sun, S., Gan, Z., Pillai, R., Wang, S., Liu, J.: Hierarchical graph network for multi-hop question answering. In: EMNLP. pp. 8823–8838 (2020)
20. Xu, M., Zheng, W., Yang, D.: Macre: Multi-hop question answering via contrastive relation embedding. In: DASFAA. pp. 595–605. Springer (2023)
21. Inoue, N., Stenetorp, P., Inui, K.: R4c: A benchmark for evaluating rc systems to get the right answer for the right reason. In: ACL. pp. 6740–6750 (2020)
22. Ho, X., Nguyen, A.K.D., Sugawara, S., Aizawa, A.: Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In: COLING. pp. 6609–6625 (2020)
23. Li, R., Wang, L., Wang, S., Jiang, Z.: Asynchronous multi-grained graph network for interpretable multi-hop reading comprehension. In: IJCAI. pp. 3857–3863 (2021)
24. Min, S., Zhong, V., Zettlemoyer, L., Hajishirzi, H.: Multi-hop reading comprehension through question decomposition and rescoring. In: ACL. pp. 6097–6109 (2019)
25. Perez, E., Lewis, P., Yih, W.t., Cho, K., Kiela, D.: Unsupervised question decomposition for question answering. In: EMNLP. pp. 8864–8880 (2020)
26. Fu, R., Wang, H., Zhang, X., Zhou, J., Yan, Y.: Decomposing complex questions makes multi-hop qa easier and more interpretable. In: EMNLP-findings. pp. 169–180 (2021)
27. Wu, J., Yang, L., Ji, Y., Huang, W., Karlsson, B.F., Okumura, M.: Gendec: A robust generative question-decomposition method for multi-hop reasoning. arXiv preprint arXiv:2402.11166 (2024)
28. Nishida, K., Nishida, K., Nagata, M., Otsuka, A., Saito, I., Asano, H., Tomita, J.: Answering while summarizing: Multi-task learning for multi-hop qa with evidence extraction. In: ACL. pp. 2335–2345 (2019)
29. Groeneveld, D., Khot, T., Sabharwal, A., et al.: A simple yet strong pipeline for hotpotqa. In: EMNLP. pp. 8839–8845 (2020)
30. Song, X., Salcianu, A., Song, Y., Dopson, D., Zhou, D.: Fast wordpiece tokenization. In: EMNLP. pp. 2089–2103 (2021)
31. Chen, D., Fisch, A., Weston, J., Bordes, A.: Reading wikipedia to answer open-domain questions. In: ACL. pp. 1870–1879 (2017)
32. Deng, Z., Zhu, Y., Qi, Q., Witbrock, M.J., Riddle, P.: Explicit graph reasoning fusing knowledge and contextual infor-mation for multi-hop question answering. In: DLG4NLP. pp. 71–80 (2022)
33. Huang, Y., Yang, M.: Breadth first reasoning graph for multi-hop question answering. In: NAACL. pp. 5810–5821 (2021)
34. Shao, N., Cui, Y., Liu, T., Wang, S., Hu, G.: Is graph structure necessary for multi-hop question answering? In: EMNLP. pp. 7187–7192 (2020)
35. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Transformers: State-of-the-art natural language processing. In: EMNLP: system demonstrations. pp. 38–45 (2020)
36. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
37. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
38. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654 (2020)