SCJ SOFT COMPUTING JOURNAL
ISSN: 2322-3707
University of Kashan

A Comprehensive Survey on Multi-hop Machine Reading Comprehension Approaches

Azade Mohammadi[1],Reza Ramezani[2*] and Ahmad Baraani [3†]

[1,2,3]Department of Computer, University of Isfahan, Isfahan, Iran.

Abstract. Machine reading comprehension (MRC) is a long-standing topic in natural language processing (NLP). The MRC task aims to answer a question based on the given context. Recently studies focus on multi-hop MRC which is a more challenging extension of MRC, which to answer a question some disjoint pieces of information across the context are required. Due to the complexity and importance of multi-hop MRC, a large number of studies have been focused on this topic in recent years, therefore, it is necessary and worth reviewing the related literature. This study aims to investigate recent advances in the multi-hop MRC approaches based on 34 studies from 2018 to 2024. In this regard, first, the multi-hop MRC problem definition will be introduced, then 34 models will be reviewed in detail with a strong focus on their multi-hop aspects. They also will be categorized based on their main techniques. Finally, a fine-grain comprehensive comparison of the models and techniques will be presented.

Keywords: Machine Reading Comprehension, Multi-hop Machine Reading Comprehension, Natural Language Processing, Question Answering, Machine Reading Comprehension Approaches

1. Introduction

Natural language processing is a branch of artificial intelligence that uses linguistic and computational techniques to help computers understand and produce human language in the form of text and speech [**?**].Machine reading comprehension (MRC) is one of the most important and long-standing topics in Natural Language Processing (NLP). MRC provides a way to evaluate an NLP system's capability for natural language understanding. An MRC task, in brief, refers to the ability of a computer to read and understand natural language context and then find the answer to questions about that context. The emergence of large-scale single-document MRC datasets, such as SQuAD [1], CNN/Daily mail [2], has led to increased attention to this topic and different models have been proposed to address the MRC problem. However, for many of these datasets, it has been found that models don't need to comprehend and reason to answer a question. For example, it has been seen that that adversarial perturbation in candidate answers has a negative effect on the performance of the QA systems. Also, it has been showed that adding an adversarial sentence to the SQuAD [1] context will drop the result of many existing models. Also Chen, Bolton, and Manning [2] pointed out that the required reasoning in the CNN/Daily Maildataset is so simple that even a relatively simple algorithm can perform well on this

*Corresponding author. Email: r.ramezani@eng.ui.ac.ir

†Email(s): Azade.mohammadi@eng.ui.ac.ir, r.ramezani@eng.ui.ac.ir, ahmadb@eng.ui.ac.ir

dataset. The above problems seem to be due to the fact that answering the questions of these datasets doesn't require a deep understanding and reasoning. In other words, these models have focused only on answering questions based on a single or few nearby sentences of the context, mostly by matching information in the question and the context (known as single-hop MRC). However, in real-world cases, to answer a question it is required to read and comprehend multiple parts of disjoint evidence to find the valid information. Therefore, there are gaps between single-hop datasets and their models, and real-world applications. To eliminate these gaps, the single-hop MRC models are faced with three important challenges:

- Reasonability: Since in real-world cases, to answer a question it is necessary to integrate and synthesize information from multiple pieces of information, more complex reasoning named multi-hop reasoning is required. Multi-hop reasoning means reasoning over information taken from more than one document [3].
- Interpretability: Interpretability is a vital feature of any reliable system. Since in real-world cases, the evidence used to find the answer is not necessarily located closely and could be comprehended from disjointed pieces of information Then, finding all the supporting facts is more difficult [4]. .
- Scalability: In real-world cases, for each question, there may be multiple supporting documents sets but only a small part of them contains valid information. Then scalability became a serious challenge which means that the models could be able to handle the increasing amount of input with a limited cost. Most single-hop models do not scale well because they have been built for one passage.

Multi-Hop Reading Comprehension (MHRC) is a more challenging extension of MRC in which the models need to properly integrate multiple pieces of evidence and reason over them to correctly answer a question. In contrast to the question in single-hop MRC that can be answered by matching information in some nearby sentences, multi-hop MRC requires answering more complex questions based on a deep understanding of the full information. The multi-hop reasoning (also known as multi-step reasoning) is considered the basic key in multi-hop MRC. Multi-hop reasoning is the ability of reach from some intermediate steps to a final step through a reasoning chain [2]. Figure 1 shows an example of Multi-hop MRC problem from HotpotQA [3]. In this case, the answer span Greenwich Village, New York City could be found in a shortcut as it is closely related to the span New York City. Also, finding evidence paragraph requires the information from paragraph Big Stone Gap (film).

Paragraph 1: Adriana Trigiani is an Italian American best-selling author of sixteen books, television writer, film director, and entrepreneur based in Greenwich Village, New York City. Trigiani has published a novel a year since 2000.

Paragraph 2: Big Stone Gap is a 2014 American drama romantic comedy film written and directed by Adriana Trigiani and produced by Donna Gigliotti for Altar Identity Studios, a subsidiary of Media Society. Based on Trigiani's 2000 best-selling novel of the same name, the story is set in the actual Virginia town of Big Stone Gap circa 1970s. The film had its world premiere at the Virginia Film Festival on November 6, 2014.

Question: The director of the romantic comedy "Big Stone Gap" is based in what New York city?

Answer: Greenwich Village, New York City

Evidence Sentences: ["Big Stone Gap (film)", 1], ["Adriana Trigiani", 1]

FIGURE 1. Multi-hop MRC from HotPotQA [3]

Then, single-hop MRC typically involves retrieving or inferring an answer directly from a single piece of evidence, often within one sentence or passage, meaning that datasets for this task (e.g., SQuAD) are designed with questions answerable from localized contexts. In contrast, multi-hop MRC requires reasoning across multiple pieces of information, which may be distributed across different sentences, paragraphs, or even documents. Datasets supporting this setting (e.g., HotpotQA, QAngaroo) are constructed to demand integration of evidence from various sources. Consequently, while single-hop QA mainly tests retrieval and shallow comprehension, multi-hop QA emphasizes more complex reasoning skills such as logical inference, comparison, and causal or temporal connections, making it a more challenging benchmark for evaluating advanced reasoning capabilities in QA models. The first attempt to improve the simple single-hop MRC task happened with emerging of some datasets like TriviaQA [5]. These datasets addressed more challenges by introducing multiple passages per each question and also presenting a more complex kind of questions that couldn't be answered with one single sentence. Although this kind of question was more complex than single-hop questions, they still could be answered by a few nearby sentences within one passage, which means they mostly do not need multi-hop reasoning. They are generally known as the multi-passage or multi-document dataset that is closer to open-domain Question Answering or retrieving-reading problems, which means models have to focus on retrieving the most related passage and then answer the question based on that passage instead of reasoning over information from multiple passages. HotpotQA [3] and WikiHop [6] can be mentioned as the first and most popular multi-hop datasets which in addition to providing multiple passages per each question, ensure that the question can only be answered by reasoning over disjoint pieces of information across different passages. It has been shown that the models with successful results in single-hop MRC datasets have limited success on these datasets [7]. Recently, a loattention in recent years. Due to the importance of this task, and also the high speed of presenting new models, it is necessary to present a comprehensive investigation of current models. It would clarify the advantages at of studies have been done in the field of multi-hop MRC, they focus on different aspects of the task. One of the most basic aspects is to propose a model for solving the multi-hop MRC problem, which has received great nd disadvantages of existing solutions and help improve future models. To have an accurate view of the growing trend of multi-hop MRC, Figure 2 has been prepared. In 2017, several datasets were introduced (like TriviaQA [5]) that are not considered multi-hop datasets, but they attracted attention to this task by proposing more complicated questions than single-hop questions. Although proposing multi-hop models has been beginning from 2018, there was a serious shortage of multi-hop datasets, so in 2018 some MRC datasets were been proposed with a main focus on the multi-hop challenges. These datasets made a proper situation to present the multi-hop MRC models, and as you can see a significant number of multi-hop models were being proposed in 2019. The trend of proposing new multi-hop datasets and models has been continued till 2024. It can be concluded from Figure 2 that after proposing each new dataset, some models will be proposed as well to address the new challenges of that dataset, as you can see the number of models in 2019 because of the large number of datasets in 2018.

There are some review papers on the MRC task that we will mention here to explain the difference and the innovation aspect of our paper. Liu et al. [8] reviewed 85 studies from

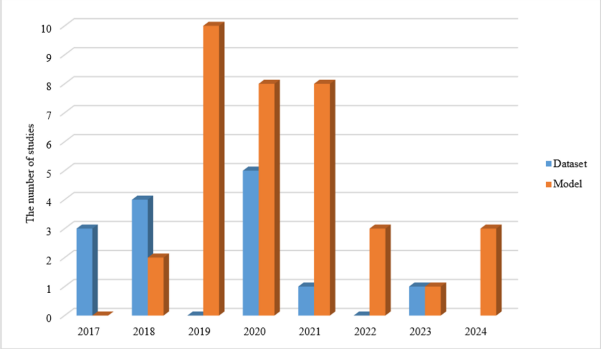


FIGURE 2. Frequency of the MRC models and datasets

2015 to 2018 with a focus on neural network solutions for the MRC problem to investigate the neural network methods in the MRC task. Baradaran, Ghiasi, and Amirkhani [9] presented a survey on the MRC field based on 124 reviewed papers from 2016 to 2018 with a focus on presenting a comprehensive survey on different aspects of machine reading comprehension systems, including their approaches, structures, input/outputs, and research novelties. Thayaparan, Valentino, and Freitas [10] proposed a systematic review of explainable MRC, from 2014 to 2020 with a focus on the explainable feature of the recent MRC methods. Zhang, Zhao, and Wang [11] presented a survey on the role of contextualized language models (CLMs) on MRC from 2015 to 2019. Bai and Wang [12] presented a survey on textual question answering with a focus on datasets and metrics, they investigate 47 datasets and 8 metrics. Although the above studies investigate different aspects of MRC/QA, none of them have focused on the multi-hop challenges. Due to the importance and increasing attention to multi-hop MRC it is necessary to investigate the multi-hop MRC studies separately. Our contribution in this paper is to focus on the multi-hop MRC models and techniques. In this regard, we first proposed the problem definition of the multi-hop MRC task, then categorize 34 models from 2018 to 2024 based on their main techniques and also investigate each model in detail. Also, a comprehensive comparison of the models and techniques will be presented. Finally, open issues in this field have been discussed.In this regard, 34 papers from 2015 to 2024 have been reviewed. Our main resources were Google Scholar, IEEE Explore, and Elsevier. We searched for papers with these keywords: "multi-hop machine reading comprehension", "multi-hop question answering", "multi-passage reading comprehension", "multi-passage question answering", "reading comprehension over multiple documents", "reading comprehension over multiple passages", "question answering over multiple documents", and "question answering over multiple passages". our selection criteria for the reviewed papers were based on their compliance with the definition of a multi-hop MRC model as presented in this study. Specifically, we ensured that the models were categorized as both MRC and multi-hop in terms of dataset, evaluation metrics, and methodology. It is important to note that since there is a close relationship between MRC and Question Answering, most of the existing machine reading comprehension tasks are in the form of textual question answering [22], also MRC is known as a basic task of textual question answering [14]. Thus, we consider cloze domain textual question answering as a typical MRC task in this paper. Figure 3 shows the relationship between QA and MRC and multi-hop MRC [17].

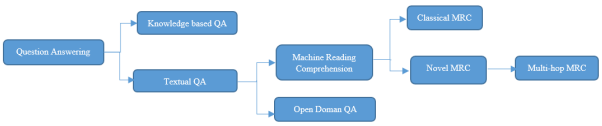


FIGURE 3. QA, MRC and multi-hop MRC relationships [17]

The rest of this paper is organized as follows: In Section 2, the definition of the multi-hop MRC problem is explained. In section 3, besides categorizing 34 models based on their main techniques, each model will be reviewed in detail with a focus on its main idea and superiority. Next, a comprehensive comparison of the models and techniques will be done by presenting some figures and tables in section 4. Then, open issues are expressed in Section 5, and finally, Section 6 concludes the paper.

2. PROBLEM DEFINITION

In general, the multi-hop MRC problem can be defined as: Given a collection of training examples (C;Q;A), the goal is to find a function F which takes a context C and a corresponding question Q as inputs, and gives answer A as output.

$$F : (C, Q) \rightarrow A \tag{1}$$

For the multi-hop MRC problem, $C = (P_1, P_2, \ldots, Plp)$ can be a set of paragraphs (or documents) where lp denotes the number of paragraphs (or documents) and also question Q is such a way that needs the multiple disjoint pieces of information from C to be answered.In other words, it needs multi-hop reasoning, each intermediate step of the reasoning chain can be considered a hop.

$$h = [s_1, s_2, ..., s_n] \quad n >= 2 \tag{2}$$

where h is for hop and s_i is the ith intermediate step. The number of hops have to be more or equal to 2. Like general MRC task, Answer A in multi-hop MRC can be in different forms, where have been divided into four categories [19]: Span-extraction, Multiple-choice, Free-form and Cloze-style. To show the frequency of each task among multi-hop studies Figure 4 has been prepared.

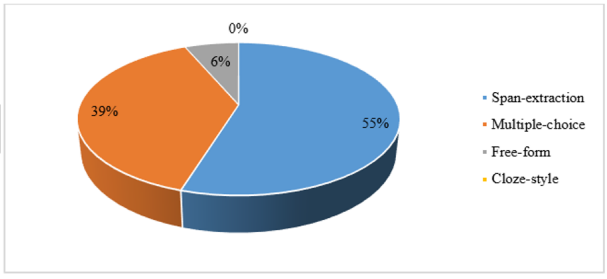


FIGURE 4. The frequency of Multi-hop MRC tasks in reviewed studies

3. TECHNIQUES

In this paper, 34 studies have been investigated, which propose a model for multi-hop MRC based on the presented problem in section 2. It is important to note that since there is a close relationship between MRC and Question Answering, most of the existing machine reading comprehension tasks are in the form of textual question answering [2], and also MRC is known as a basic task of textual question answering [8]. Thus, we consider cloze domain textual question answering as a typical MRC task in this paper. Existing studies for multi-hop MRC can be mainly divided into three categories: decomposition, recurrent reasoning based on memory retrieval, and multi-step reasoning based on graph neural networks. For each category, after explaining the main technique, the multi-hop MRC models will be reviewed in detail; beside reviewing the detail architectures of each model, we also focus on the superiority and the motivation of them. Also, the disadvantages of each technique will be discussed at the end. In the next section (4) a comprehensive comparison of the techniques and models will be presented. It should also be emphasized that this categorization is based on the primary reasoning strategy adopted in each model. However, the boundaries between the three categories are not absolute. Several recent approaches integrate multiple strategies—for instance, combining decomposition with memory retrieval, or leveraging graph-based reasoning together with recurrent reasoning. Such models can be considered as hybrid approaches, reflecting the fact that multi-hop MRC often requires complementary techniques to achieve stronger performance. By explicitly acknowledging these hybrid models, the proposed taxonomy gains both flexibility and explanatory power, while still providing a clear structure for analyzing the literature. The techniques do not have a specific order, because all three techniques have been used by models from 2018 to 2022, but as much as possible, the studies within the techniques have been according to the order of published time.

3.1. Decomposition. Complicated question is a basic challenge of multi-hop MRC, unlike the single-hop questions, they cannot be answered easily and require complicated reasoning. Since the human reasoning about complex questions is done by decomposition, answering sub-questions, summarizing, and comparing [14], then this technique has focused on simplifying the problem by decomposition of a complex question into multiple simple sub-questions. It means it reduces multi-hop MRC to multiple single-hop MRC. This technique mostly uses the single-hop MRC models to find the answers to sub-questions and then combine the answers to obtain the final answer. In the following, the models which use this technique for multi-hop MRC will be reviewed in detail.
Self-assembling MNM: Jiang and Bansal [7] focused on identifying the sub-questions in the correct reasoning order and presented an interpretable and controller-based self-assembling neural modular network for the multi-hop reasoning process which includes two main parts, Modular Network with a Controller (top) and the Dynamically-assembled Modular Network (bottom) that can be seen in Figure 5. The main idea of the model to handle multi-hop questions is done with Controller that computes an attention distribution over all question words at every reasoning step, which finds the sub-question that should be answered at the current step. The mentioned modules are described as follows:

- The Find module first builds a similarity matrix between the question and context.
- The Relocate module pops an attention map from the stack and computes the bridge entity's representation, which is a weighted average over context representation.

- The Compare module pops two attention maps from the stack and computes two weighted averages over the context representation using the attention maps.

However, this system approaches question decomposition by having a decomposer model trained via human labels [4].

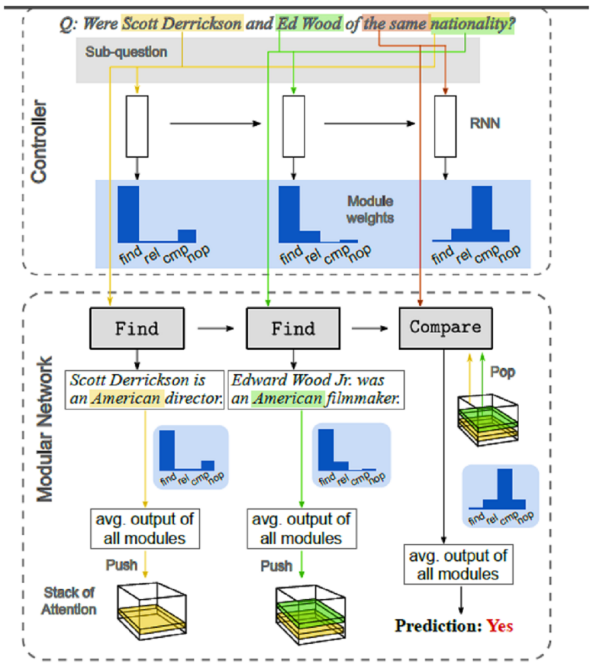


FIGURE 5. Self-assembling MNM Architecture

ONUS: Perez et al. [15] proposed an algorithm for One-to-N Unsupervised Sequence transduction (ONUS) that map a complicate multi-hop question to some simpler single-hop sub-questions. Unlike other decomposition studies use a combination of hand-crafted heuristics, rule-based algorithms, and learning from supervised decompositions to decompose multi-hop question which require significant human effort, this model automatically learns to decompose different kinds of questions. The main idea has been shown in Figure 6. To decompose multi-hop question Q to simpler corpus D, First some candidate sub-question from a simple corpus S will be created by mining 10M possible sub-questions from Common Crawl with a classifier. It then trains a decomposition model on the mined data using Q and D with unsupervised sequence-to-sequence learning to map multi-hop questions to sub-question. With this idea, the model is able to train a large transformer model to generate decompositions and avoid heuristic/extractive decompositions.

CGDe-FGIn: Coa and Liu [4] proposed a Coarse-Grained Decomposition Fine-Grained Interaction (CGDe-FGIn) model to handle the complicated multi-hop questions. Existing studies use Bi-directional Attention Flow (Bi-DAF) to capture semantic feature interaction between documents and questions, but Bi-DAF generally captures the surface semantics of words, instead of the implied semantic feature of intermediate answers, so it cannot extract

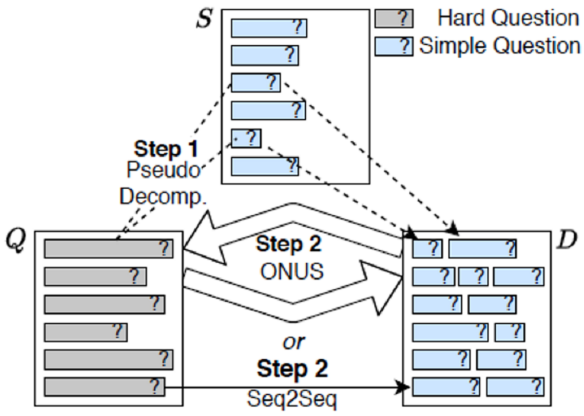


FIGURE 6. ONUS Architecture

the most important parts of multiple documents in the multi-hop MRC task. CGDe-FGIn consists of:

- Coarse-Grain complex question Decomposition (CGDe) to decompose complex questions into simple sub-questions without any additional human annotations (Figure 7).
- Fine-Grained Interaction (FGIn) to better represent each word in the document and extract more comprehensive and accurate sentences related to the inference path instead of using Bi-DAF (Figure 8).

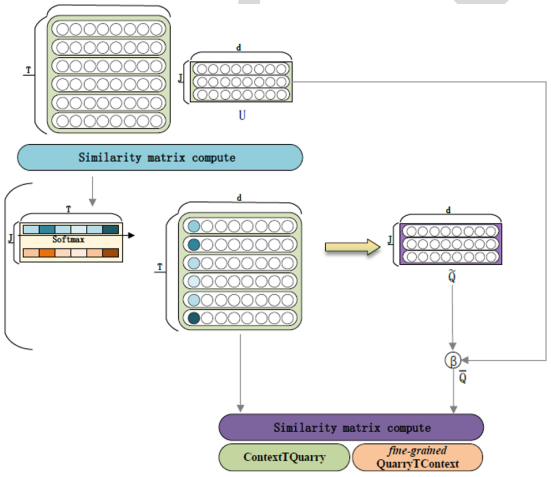


FIGURE 7. Coarse-grained decomposition component

RERC: Fu et al. [14] proposed a three-stage Relation Extractor-Reader and Comparator (RERC) model. RERC is consist of 1) Relation Extractor to decompose the complex questions into simple sub-questions by automatically extracting the subject and key relations of the complex question, 2) Reader to find the answers to the sub-questions in turn

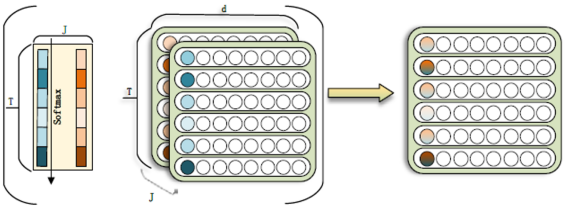


FIGURE 8. Fine-grained interaction component

by an advanced ALBERT model, and finally, 3) Comparator to perform the numerical comparison and summarizes all the answers to get the final answer (Figure 9).

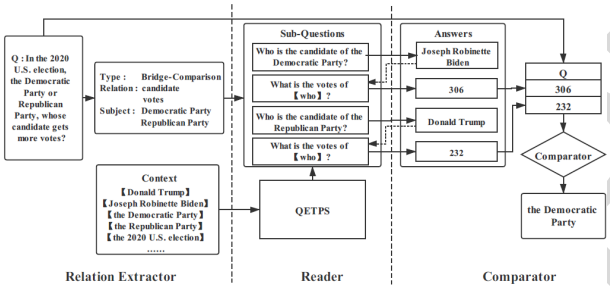


FIGURE 9. RERC Architecture

The Decomposition technique was one of the first ideas for multi-hop MRC and as you can see in recent years (2021) still has attracted attention. The main disadvantage of this technique is that, instead of focusing on multi-hop reasoning as an important key of multi-hop MRC, it focuses on reducing the problem to a single-hop MRC. Thus, they actually do not go far beyond single-hop models.

3.2. Recurrent reasoning-based. Recurrent neural networks are designed to capture temporal dependencies in sequential data [34].The sequence models have been first used for single-hop MRC tasks, and most of them are based on Recurrent Neural Networks (RNNs), some studies focus on using them in multi-hop MRC. It can be called state-based reasoning models and they are closer to a standard attention-based RC model with an additional "state" representation that is iteratively updated [35]. Most models, presented in this section, have used advanced neural network concepts, such as attention mechanism and network memory for multi-hop reasoning. In the following the models which use this technique will be reviewed in detail including the architecture alongside the superiority and the motivation of them.

TAP: Bhargav et al. [18] proposed a deep neural architecture, called TAP (Translucent Answer Prediction) cover of two main ideas: (1) Local Interaction: Each sentence should be understood in the context of its neighboring sentences and the question, (2) Global Interaction: A global (inter-passage) interaction among sentences must be identified and used for supporting facts. TAP is a hierarchical architecture that tries to capture the local and global interactions between the sentences and consists of two main parts: (Figure 10)

- Local and Global Interaction eXtractor (LoGIX) with three layers: local layer to obtain intra-passage dependencies, Global Layer to obtain inter-passage dependencies, and Supporting Facts Prediction Layer to calculate the probability that a sentence is a supporting fact.
- Answer Predictor (AP) to predict the final answer by reasoning over the supporting facts. It consists of four parts: Input Data Shaping to prepossess and concatenate the supporting facts, Context Encoding to encode the context using a pre-trained BERT model, Answer Type Predictor to classify the question into one of the three classes (yes, no and, span), and Answer Span Predictor to predict the final answer.

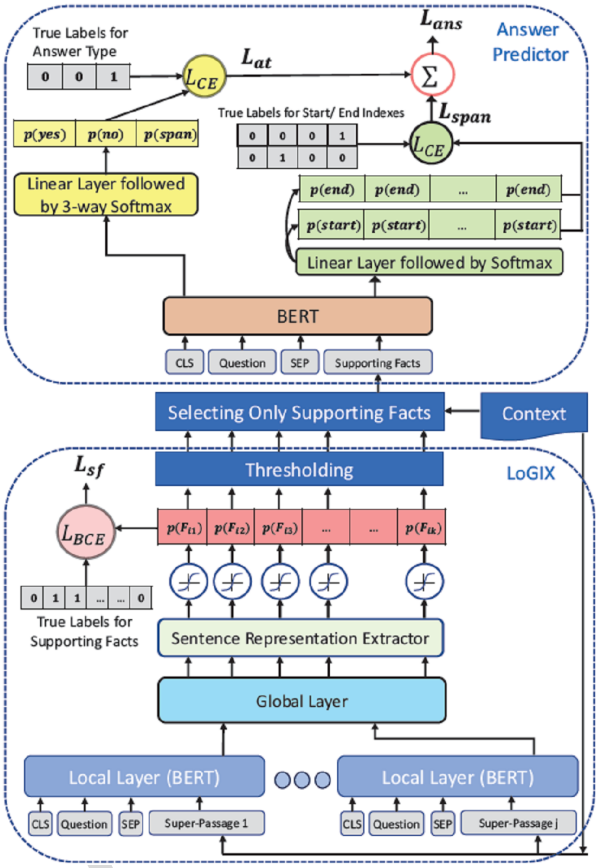


FIGURE 10. TAP Architecture

PH-Model: Cong et al. [19] focused on using the benefit of the hierarchical structure of the natural language text (document-passage-sentence-word-character). As you can see in Figure 11, PH-Model consists of multiple main parts: Bi_ONLSTM, that is an ordered neuron LSTM is used to obtain hierarchical information from a passage (Instead of traditional LSTM), Bidirectional Attention Flow is used to extract the hierarchical information form paragraphs to get the query-aware context and the context-aware query representation, Fused layer is used to merge all information and finally Pointer network to obtain the probability of the start and end positions of the answer.

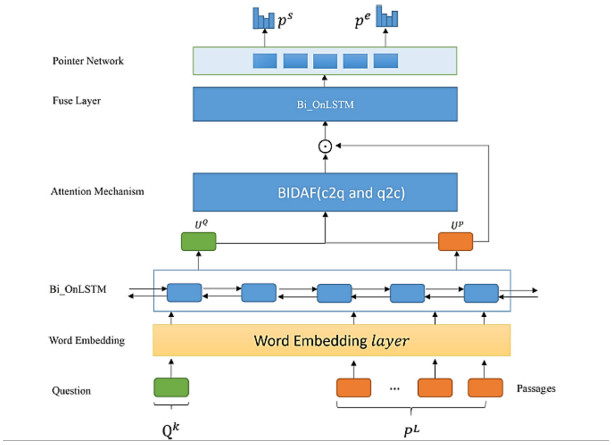


FIGURE 11. PH-Model Architecture

3.2.1. *Path-based.* As multi-hop MRC faces more information and complex questions, finding the right path has become more important and difficult. One of the important advantages of path-based models is that they are interpretable because they can provide the evidence chain to the final answer.

PathNet: Kundu et al. [16] proposed a path-based reasoning approach for multi-hop MRC which first extracts all paths in the passages based on implicit relations between entities, and then composes the passage representations along each path to compute a passage-based representation. This model first finds all possible path from passages by selecting a passage that contains a head entity from the question, and then finds all entities and noun phrases from the same sentence. Afterward, it selects the next passage that contains the potential intermediate entity identified above. Finally, it is checked whether the next passage contains any of the candidate answer choices or not. The resulting will be a set of entity sequences. After obtaining all potential paths, it is time to score each path using the PathNet model based on two perspectives: 1) Context-based Path Scoring, which is based on the interaction with the question encoding, and 2) Passage-based Path Scoring, which is based on the interaction between the passage-based path encoding vector and the candidate encoding. There is an example of the process in Figure 12 which In the Rank-1 path, the model composes the implicit located in relations between (Zoo lake, Johannesburg) and (Johannesburg, Gauteng). However, this method extracts many invalid paths, then causes wasting the computing resources.

Question: (zoo lake, located in the administrative territorial entity,?) **Answer**: gauteng
Rank-1 Path: (zoo lake, Johannesburg, gauteng) Passage1: ... **Zoo Lake** is a popular lake and public park in **Johannesburg**, South Africa. It is part of the Hermann Eckstein Park and is ... Passage2: ... **Johannesburg** (also known as Jozi, Joburg and Egoli) is the largest city in South Africa and is one of the 50 largest urban areas in the world. It is the provincial capital of **Gauteng**, which is ...
Rank-2 Path: (zoo lake, South Africa, gauteng) Passage1: ... **Zoo Lake** is a popular lake and public park in Johannesburg, **South Africa**. It is ... Passage2: ... aka The Reef, is a 56-kilometre - long north - facing scarp in the **Gauteng** Province of South Africa. It consists of a ...

FIGURE 12. Two top-scoring path

ChainEx: Chen, Lin and Durrett [20] proposed a sentence-based model that does not rely on gold annotated chains or supporting facts at the training and test phases, instead, pseudo-gold reasoning chains are derived using some heuristics based on named entity recognition and coreference resolution during the training time, and it learns to extract chains from raw texts at the test time. To construct the reasoning chain (Figure 13), each sentence is considered as a node in the chain, and there is an edge between two sentences if they have the same entity. Besides, there are edges between all sentences from the same paragraph. The model starts from the question and finds all possible reasoning chains. The chain extractor is a neural network with two main components: Sentence Encoding and Chain Prediction. In the sentence encoding component, the BERT-Para model provides a representation from each paragraph jointly with the question. In the chain prediction component, an LSTM-based pointer network is used to extract the reasoning chain. The output of the chain extractor is a variable-length sequence of sentences. Finally, a BERT-based QA system is applied to the extracted chains to find the final answer.

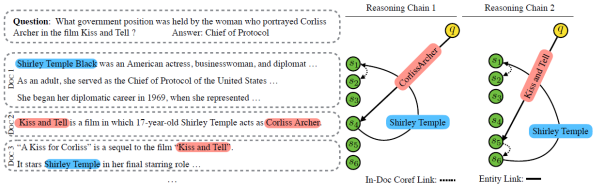


FIGURE 13. ChainEx Architecture

PEI: Huang et al. [21] proposed Prompting Explicit and Implicit knowledge (PEI) framework to which connect explicit and implicit knowledge to answer a multi-hop question. The input passages are the explicit knowledge which is incorporated with type-specific reasoning as the implicit knowledge. There are three components in PEI (Figure 14): 1) the type prompter identifies and acquires the weights associated with specific reasoning types for multi-hop questions; 2) the knowledge prompter elicits implicit knowledge through harnessing of explicit knowledge; 3) the unified prompt-based PLM integrates explicit and implicit knowledge, as well as question types, providing a comprehensive approach for multi-hop QA.

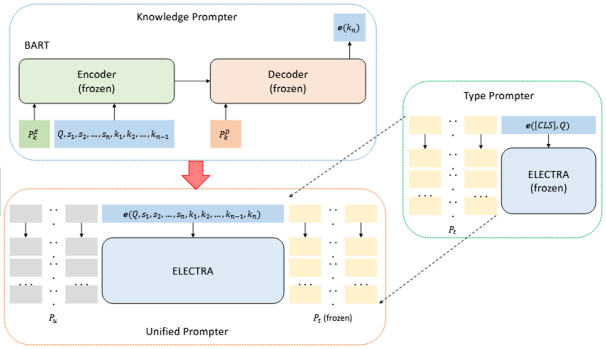


FIGURE 14. PEI Architecture

Unlike the decomposition technique, this technique focuses on multi-hop reasoning instead of simplifying the multi-hop MRC problem and has achieved great attention among studies. The main disadvantage of this technique is that they are expensive and time-consuming approaches, and require a large amount of data to train.

3.3. Graph-based. Graph-based techniques because of their natural language relationship representation ability [2] has attracted attention in multi-hop MRC. It is natural to model natural language context into graph structure and the process of multi-hop reasoning as moving among nodes. Constructing graphs from input data is one of the basic parts of this technique. Some studies construct an entity graph from the input data, which means the nodes of the graph nodes are the entities of the context and question. A lot of studies work on this kind of graph which will be reviewed in the following.

3.3.1. *Entity-node graphs.* : One of the most basic ideas for converting text to a graph is to construct an entity graph, which means the entities within the text are node graphs and the relationships between the entities are considered edges. In this section, the studies that uses entity graph will be reviewed.

DFGN: Xiao et al. [22] proposed a fusion process of Doc2Graph and Graph2Doc for multi-hop reasoning that leads to a less noisy entity graph and more accurate answers. The process of constructing dynamic entity graph iterates in multiple rounds to achieve multi-hop reasoning. In each round, DFGN generates and reasons on a dynamic graph, where irrelevant entities are masked out while only reasoning sources are preserved, via a mask prediction module. Then the fusion block not only aggregates information from documents to the entity graph (doc2graph) but also propagate the information of the entity graph back to document representations (graph2doc). Figure 15 illustrates the Fusion block in DFGN.

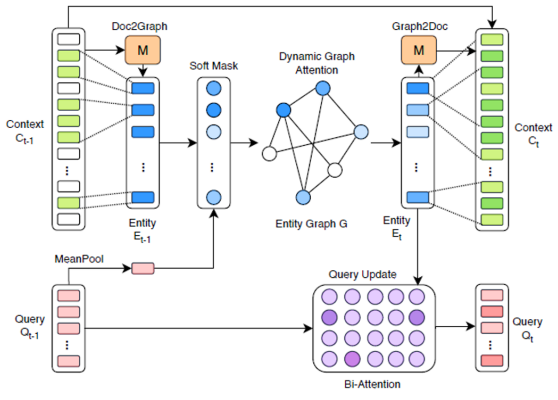


FIGURE 15. Fusion block

BAG: Cao, Fang, and Tao [23] proposed a Bi-directional Attention Entity Graph Convolutional Network (BAG) with a focus on Relational Graph Convolutional Network (RGCN). The R-GCN is employed to handle high-relational data characteristics and make use of different edge types. As it is shown in Figure 16, the model has five layers, Entity Graph Construction, Multi-level Features, GCN, Bi-directional Attention, and Output. However, as the number of inferences increases, the complexity of models will rise sharply due to the iteration of the message passing algorithm, resulting in low efficiency.

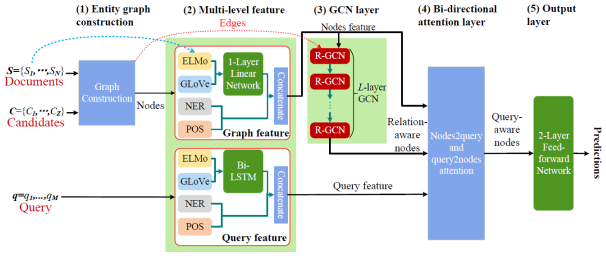


FIGURE 16. BAG Architecture

DRNQA: Li et al. [24] proposed a Dynamic Reasoning Network (DRN) approach that, unlike other models that read the question only once, uses a query reshaping mechanism that considers the question several times, which as a result enhances the ability of the reasoning over the entity graph. There are two main parts: This graph (Figure 17) is constructed from three levels: 1) Question-based level: edge between two entities if their sentences have a common entity with the question, 2) Context-based level: edge between two entities if they are from the same passages, and 3) Passage-based level: edge between two entities if their sentences have at least one entity or phrase in common. Also, the dynamic reasoning network component (Figure 18) is proposed to reason over the entity graph. The query reshaping mechanism (Figure 19) causes the important parts of the question to be read frequently.

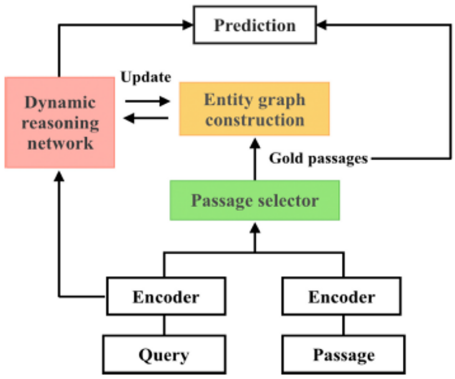


FIGURE 17. DRNQA Architecture

3.3.2. *Multiple-node graphs.* Previous studies have used entity graphs in which only entities are considered node graphs. However, due to the complex and different structures in natural language text, it seems that entity graphs can cause a lot of information loss. Due to the importance of this issue, many studies have tried to construct more complex and enriched graphs to capture more information about the context, which will be investigated in the following.

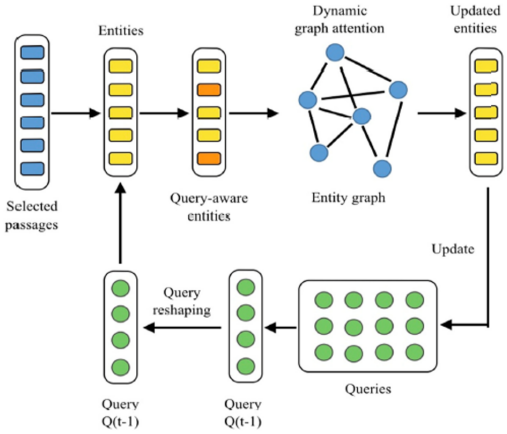


FIGURE 18. DRN component

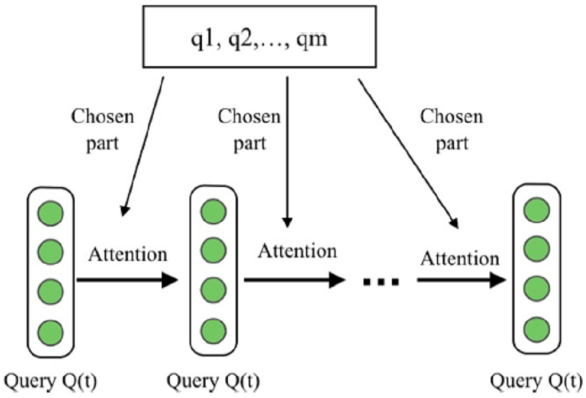


FIGURE 19. Query Reshaping component

HDE: Tu et al. [25] proposed a more complex graph named Heterogeneous Document-Entity (HDE) graph with different types of nodes and edges (Figure 20). This graph can cover different granularity levels of information in context and also enables rich information interaction among different types of nodes for accurate reasoning. As Figure 21 shows,this model can be categorized into three parts: initializing HDE graph nodes with co-attention and self-attention-based context encoding, and reasoning over HDE graph with GNN-based message passing algorithms and score accumulation from updated HDE graph nodes representations. However, it has been shown that if the number of inferences increases, the complexity of models will rise sharply due to the iteration of cumbersome message passing algorithm, resulting in low efficiency.

HGNN: Wang et al. [26] proposed a hierarchical graph neural network with a focus on compositional QA. The nodes can be normal tokens, question tokens, sentence tokens, and special html image tokens. As you can see in figure (left), the input of the BERT sequence encoder is a question q, two sentences $(s_1 + s_1)$ and a special image node h1 (some special tokens is used to indicate the question <SEP>, sentence <EOS> and special html image

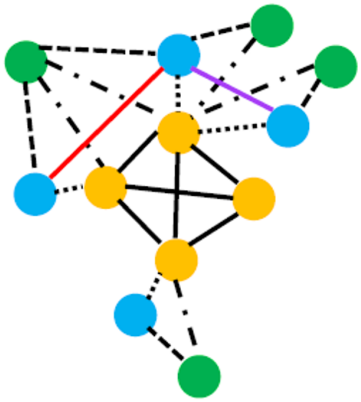


FIGURE 20. HDE graph

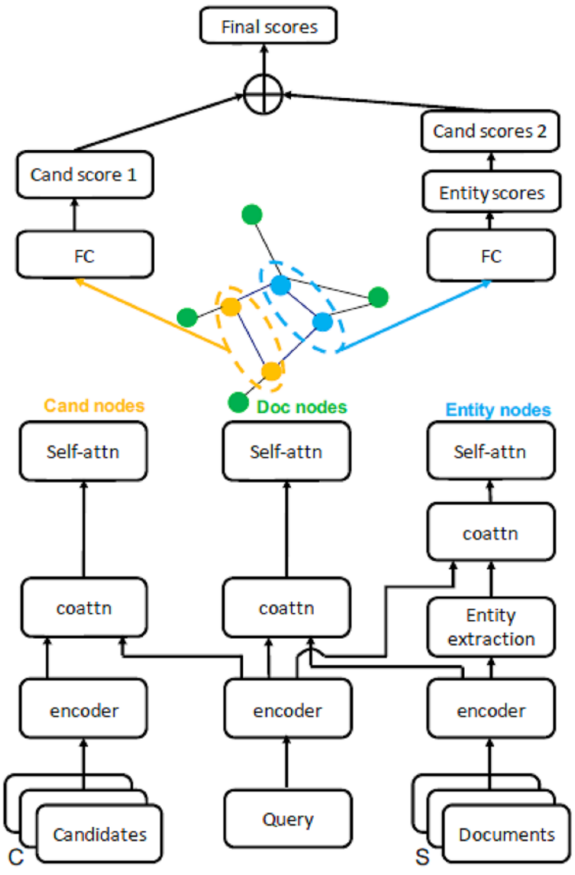


FIGURE 21. HDE Architecture

element <html>. The hierarchical graph neural networks blocks are shown in Figure 22 (middle). The attention-based Hierarchical Graph Neural Network (HGNN) uses three types of connections: 1) intra-sentence connection which means the connection of words

within a sentence, 2) inter-sentence connection which means the connection of common tokens (e.g., sentence tokens or question tokens), and 3) global connection which means the connection between the question tokens and all the words in the document. The final prediction is made based on the sentence nodes and special html nodes. Finally, the connection mask matrix is used to connect the different tokens in the graph and predict the final answer.

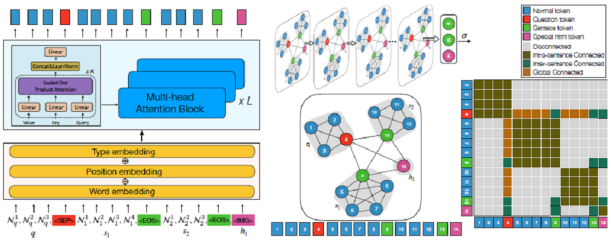


FIGURE 22. Architecture of HGNN

AMGN: Li et al. [27] proposed a model named Asynchronous Multi-grained Graph Network (AMGN) by constructing a multi-grained graph using the entity and sentence to reflect the relation level of the information. Also, an algorithm for asynchronous message propagation according to the relationship levels to update the graph to mimic human multi-hop reading logic is proposed. Besides, a question reformulation mechanism (RNN-based) is proposed to iteratively update the latent question representation with sentence nodes. These sentence nodes are directly used for supporting fact prediction. As it has been shown in Figure 23, the whole model consists of four main components: Paragraph-selector for reducing search space, Encoder to encode the context and question, Construction & Reasoning for multi-grained graph construction and multi-step asynchronous node update, and Multi-task Prediction to predict the final answer.

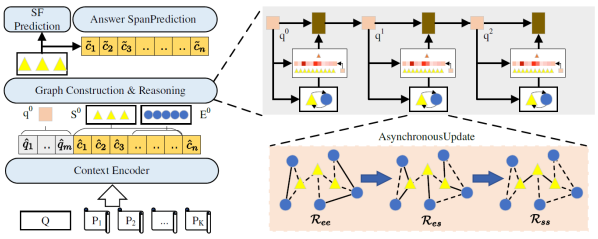


FIGURE 23. Architecture of AMGN model

ClueReader: Feng et al. [28] proposes a model with a heterogenous graph and also based on the grandmother cells concept [5] to imitate the process of human brain. This concept explains that, we generally recall a mountain of related evidence whatever the form it is (such as a paragraph, a short sentence, or a phrase), and coordinate theirs inter relationships before we carry out the final results. However, most of the studies on multi-hop MRC cannot gather the semantic features in multi-angle representations, which causes incomplete conclusions. Then, a spatial graph attention framework named

ClueReader has been proposed. This model is designed to assemble the semantic features in multi-angle representations and automatically concentrate or alleviate the information for reasoning. As you can see in Figure 24 after constructing a graph from different kinds of nodes, a GAT layer performs the message passing to find the final answer.

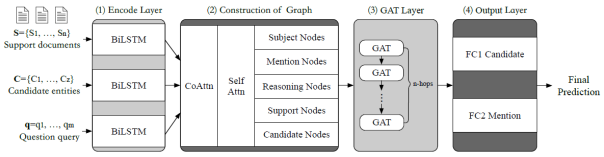


FIGURE 24. Architecture of ClueReader model

IP-LQR: Tang et al. [29] proposed a model named the latent query reformulation method (IP-LQR) to consider the phrase as nodes to construct the graph. Then they proposed the latent query reformulation method (IP-LQR), which incorporates phrases in the latent query reformulation to improve the cognitive ability of the system. They also design a semantic-augmented fusion method based on the phrase graph, which is then used to propagate the information (Figure 25). Also, a similarity evaluation strategy is designed to calculate the weights of edges in the graph, also the fusion layer is used as an information aggregation to latently update the original question's representation. Finally, a re-attention mechanism is used to help locate the gold answer based on the new representation.

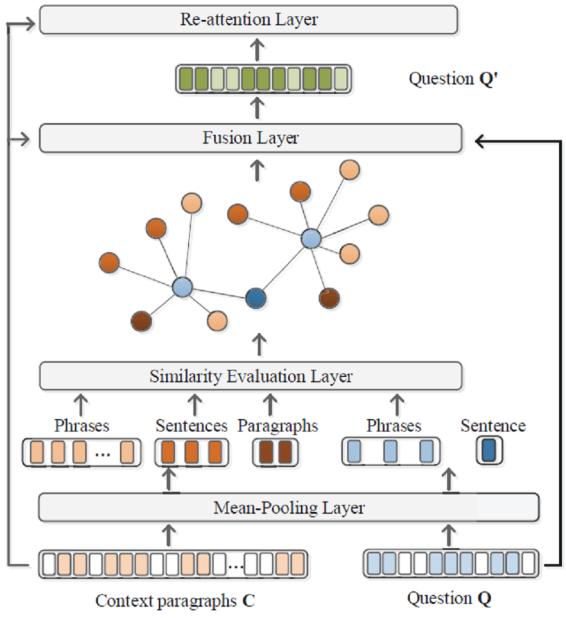


FIGURE 25. Architecture of IP-LQR model

MKGN: Ying et al. [30] proposed a model named Multi-Dimensional Knowledge Enhanced Graph Network (MKGN) utilizes dependency relations and commonsense knowledge for the reasoning process and proposed Multidimensional Knowledge enhanced Graph Network (MKGN). It uses specific knowledge to deal with information gaps and enhance representations of both questions and contexts in the reasoning process by using different dimensional knowledge, i.e., named entities, dependency relations and commonsense. There are two main components (Figure 26): 1) Knowledge Extractor extracts various knowledge from the contexts and the question and formulates them and 2) Knowledge Enhancer enhances representations of questions and contexts with each kind of knowledge generated by the knowledge extractor. Besides, they use the sequential and parallel manner kinds of fusion architectures. To stimulate a sequential reasoning process, they fuse entity information, dependency relations, and commonsense one by one, also according to the fact that humans exploit multiple knowledge at the same time when making inferences they consider a parallel architecture for multi-dimensional knowledge utilization as well.

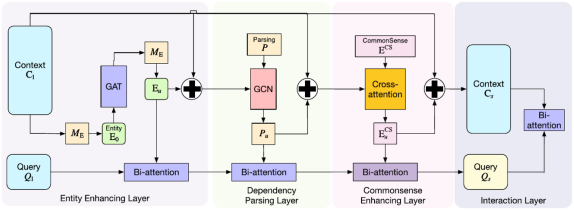


FIGURE 26. Architecture of MKGN model

FOCAL REASONER: As the entity-aware knowledge is insufficient for tasks that require knowledge of temporary facts or events, Ouyang et al. [31] proposed a model named FOCAL REASONER which instead of focusing on entity-aware knowledge cover both commonsense and temporary knowledge clues hierarchically. Then, a general formalism of knowledge units as a super-graph is proposed by extracting backbone constituents of the sentence, such as the subject-verb-object formed "facts". FOCAL REASONER consists of three stages (Figure 27). Firstly, it extracts fact units from raw texts via syntactic processing and constructs a super graph. Then, it performs reasoning over the super graph along with a logical fact regularization. Finally, it aggregates the learned representation to decode the right answer.

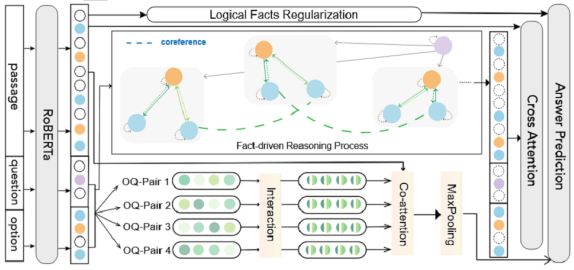


FIGURE 27. Architecture of OCAL REASONER model

CNT: Ghafori et al. [32] proposed a content-based reasoning approach based on graph-based machine reading comprehension methods. In this regards, relevant paragraphs are selected in a two-step process then, an incoherent graph is constructed. Lastly, to overcome the challenge of interpretability in the question-answering system, a transformer and the predicted answer are utilized. Figure 28 shows the framework of this model which is consist of modules for paragraph selection, heterogeneous graph construction, question answering, and supporting fact detection. To leverage the advantages of graph structures, a lightweight heterogeneous graph is designed to reduce computational costs without compromising efficiency, while facilitating reasoning for question answering. This graph has three types of nodes: Entities or Noun Phrases, Sentences and Questions, Paragraphs.

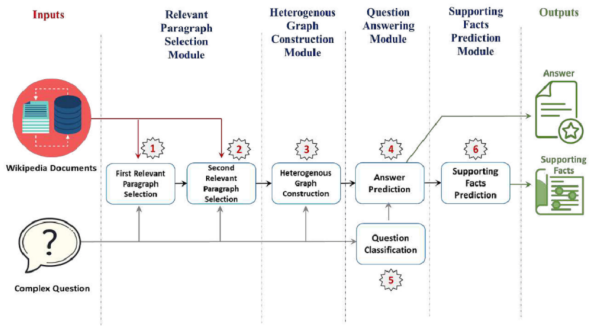


FIGURE 28. Architecture of CNT model

Recently, graph-based models have become very popular and recognized as the main solution for multi-hop MRC because of the nature of modeling such a process into graph structure and the good results. But there is some drawback to this technique, the first one is the expensive computational process of the graph-based methods [52], and the second problem is that the graph-based models often can't cover all the inherent structure of documents and loss valuable structural information by modeling documents into graphs [7]. As the last part of this section, Figure 29 has been prepared to summarize the techniques and models.

4. COMPARISON

After reviewing the details of each model and categorizing them, in this section, a fine-grained comparison of models and techniques will be presented. In this regard, we first focus on the frequency of usage and popularity of each technique in recent years.

4.1. Techniques frequency. The frequency of each technique among reviewed studies is shown in Figure 30. As you can see, the number of studies of the Graph-based techniques is the most, and after that the Recurrent reasoning-based technique has achieved good attention. But the number of studies cannot be enough to have a proper investigation, and it is needed to show the growth trend of each technique in different years. In this regard, Figure 31 shows the growth trend of each technique from 2018 to 2024. The graph-based technique has achieved the most attention in 2019, 2021, and 2024 that proves that popularity trend of this technique in different years as well. The first graph-free study

FIGURE 29. Multi-hop MRC models

has been proposed in 2020 and immediately this question was raised that whether the graph was really necessary due to the expensive computational? After that, some other studies followed this question and it can be said that can affect the popularity trend of the graph-based technique in future. However, graph-based technique still can be considered the most popular technique in multi-hop MRC.

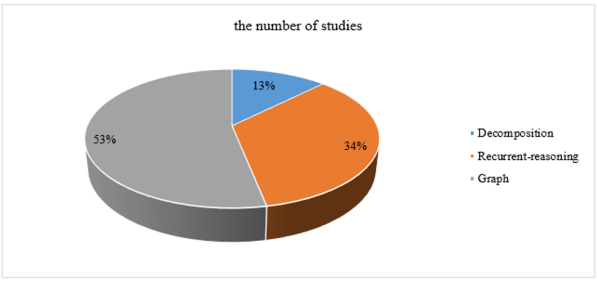
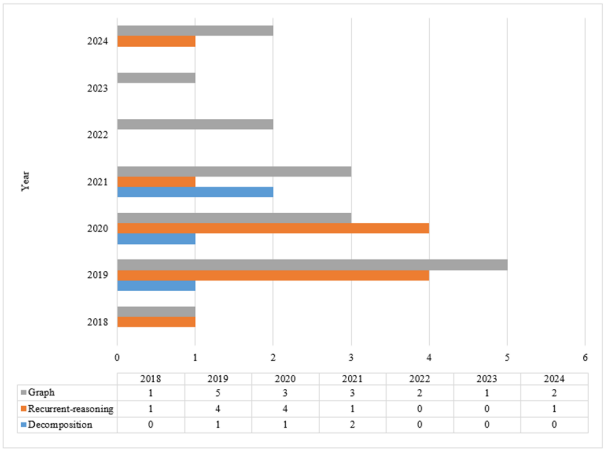


FIGURE 30. Technique frequency of revised Multi-hop model

4.2. Technique characteristic. In this section, the aim is to evaluate the performance and effectiveness of the models in each technique categories. In this regard two datasets that are the most ferequent among the reviewd models (HotpotQA [3] and WikiHop [6]) have used. The aim of the comparisons presented in this section is not only to compare models but also to compare the performance of the multi-hop techniques.

HOTPOTQA: HotpotQA [3] contains around 113k questions and Span-extraction answers extracted from Wikipedia articles, and each question often requires combining information from two or more documents. First, the models that have used the HotpotQA [3] dataset are investigated whose results are shown in Table 19. Since the answer type of the HotpotQA dataset is Span-extraction, the evaluation metrics for this dataset are F1



	2018	2019	2020	2021	2022	2023	2024
Graph	1	5	3	3	2	1	2
Recurrent-reasoning	1	4	4	1	0	0	1
Decomposition	0	1	1	2	0	0	0

FIGURE 31. Frequency of Multi-hop techniques in recent years

score and EM. F1 measures the harmonic mean of precision and recall, while EM (Exact Match) measures the percentage of predictions that exactly match the correct answers. For a better comparison of the results, the year of publications, the category, the evaluation metrics, and the underlying datasets along with their version have been presented in this table. As it can be seen in Table 1, CNT [32] and PEI [21] have achieved the best results, which indicates that Graph-based and Recurrent Reasoning-based techniques are the most effective techniques to solve the multi-hop MRC problem on the HotpotQA dataset. Since the HotpotQA dataset is the most common multi-hop MRC dataset, the results on this dataset are very important.

TABLE 1. Results of revised models on the HotpotQA dataset

Paper	Year	Category	EM	F1 score
DFGN [22]	2019	Graph-based	56.31	69.69
Self-assembling MNM [7]	2019	Decomposition	49.58	62.71
HGN [26]	2020	Graph-based	69.22	81.19
ChainEX [20]	2020	Graph-based	61.20	74.11
DRNQA [24]	2020	Graph-based	58.49	72.42
TAP [18]	2020	Recurrent Reasoning-based	66.64	79.82
CGDe-FGIn [4]	2021	Recurrent Reasoning-based	50.89	65.41
MKGN [30]	2022	Graph-based	58.84	72.90
PEI [21]	2024	Recurrent Reasoning-based	72.89	85.32
CNT [32]	2024	Graph-based	78.71	86.51

WikiHop: WikiHop [6] has about 51k multiple-choice questions, where the model must select the correct answer from a set of candidate options by reasoning across different Wikipedia passages. Accuracy is the evaluation metric on this dataset which measures the percentage of questions for which the model predicts the correct answer. For each model, the year of publication, the category, and results are shown in Table 2. As you can see in this table, the most common techniques in this dataset are Graph-based and Reasoning-based techniques. ChainEX [20]which has used a Graph-based technique has achieved the best result.

TABLE 2. Results of revised models on the WikiHop dataset

Paper	Year	Category	Accuracy	
			EM	F1 score
BAG [23]	2019	Graph-based	69	66.5
HDE [25]	2019	Graph-based	74.3	70.9
PathNet [16]	2019	Recurrent Reasoning-based	69.6	67.4
ChainEX [20]	2020	Graph-based	76.5	72.2
ClueReader [28]	2023	Graph-based	72.0	66.5

5. CONCLUSION

In this study, we focused on the multi-hop MRC approaches. Then, after presenting the multi-hop MRC problem definition, techniques had been explained based on 34 studies from 2018 to 2024. In addition to categorize the approaches based on the main technique, they also were reviewed in detail including the architecture, superiority, and motivations.

6. OPEN ISSUE

Large Language Models (LLMs) such as ChatGPT and PaLM have significantly impacted the field of multi-hop question answering (QA) datasets. The emergence of LLMs, with their enhanced abilities in understanding context, performing complex reasoning, and generating coherent text, has led to both advancements and new challenges in the context of these datasets. One key impact is the surprisingly strong performance of LLMs on existing multi-hop QA datasets, often achieving state-of-the-art results with minimal or no fine-tuning. Then existing multi-hop QA datasets is not sufficient to evaluate the true reasoning abilities of LLMs. Moreover, LLMs are also being used as tools in the creation of multi-hop QA datasets. Their ability to generate complex questions and plausible distractor answers can aid in the development of more challenging and diverse benchmarks. In conclusion, LLMs have demonstrated impressive capabilities on multi-hop QA datasets, pushing the boundaries of what's achievable in this task. However, their performance has also highlighted the shortcomings of existing benchmarks and driven the community towards developing more robust and insightful evaluation methods that can truly assess the multi-step reasoning abilities of these powerful models. The interaction between LLMs and multi-hop QA datasets is an ongoing and evolving process, with each influencing the development of the other.

The lower number of free-form answering and cloze-style multi-hop MRC datasets is primarily due to the increased complexity and cost of annotation and evaluation, which has historically led to a greater focus on other types of MRC tasks. The lack of free-form and close-style datasets has made it impossible to present models for these two tasks, while both tasks have a good potential to use in form of multi-hop MRC. Among all the types of MRC tasks described in Section 2, the free-form task is the most similar task to the real-world case but due to its complexity, there is no proper dataset for it. Since this type of answer is more similar to real-world scenarios, focusing on presenting free-form datasets to encourage models can improve the application of MRC systems.

Another open challenge in multi-hop MRC research lies in the limitations of existing benchmark datasets. Issues such as artifacts and shortcut reasoning often allow models to exploit superficial patterns or spurious correlations instead of engaging in genuine multi-step reasoning. As a result, model performance may be overestimated, raising concerns about the validity of evaluation. These problems are not dataset-specific but have been

reported across widely used benchmarks, most notably HotpotQA and WikiHop, where annotation artifacts and shallow lexical overlaps can unintentionally reveal the correct answer. Addressing these dataset-related issues is essential to ensure that future models are assessed based on their true reasoning capabilities rather than their ability to leverage unintended cues.

References

1. P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," Proc. 2016 Conf. Empir. Methods Nat. Lang. Process. Austin, Texas Assoc. Comput. Linguist., pp. 2383–2392, 2016. doi: 10.18653/v1/D16-1237.
2. D.Chen, J. Bolton, and C. D. Manning, "A thorough examination of the cnn/daily mail reading comprehension task," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 2358–2367. doi: 10.18653/v1/P16-1229.
3. Z. Yang 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, 2018, pp. 2369–2380. doi: 10.18653/v1/D18-1262.
4. X. Cao and Y. Liu, "Coarse-grained decomposition and fine-grained interaction for multi-hop question answering," J. Intell. Inf. Syst., pp. 1–21, 2021. doi: 10.1007/s10844-021-00685-6.
5. M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1601–1611. doi: 10.18653/v1/P17-1149. 6, pp. 317–328, 2018. doi: 10.1162/tacl-a-00018.
6. J. Welbl, P. Stenetorp, and S. Riedel, "Constructing datasets for multi-hop reading comprehension across documents," Trans. Assoc. Comput. Linguist., vol. 6, pp. 287–302, 2018. doi: 10.1162/tacl-a-00016.
7. 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, 2019, pp. 4474–4484. doi: 10.18653/v1/D19-1447.
8. S. Liu, X. Zhang, S. Zhang, H. Wang, and W. Zhang, "Neural machine reading comprehension: Methods and trends," Appl. Sci., vol. 9, no. 18, p. 3698, 2019. doi: 10.3390/app9183698.
9. R. Baradaran, R. Ghiasi, and H. Amirkhani, "A survey on machine reading comprehension systems," Nat. Lang. Eng., pp. 1–50, 2020. doi: 10.1017/S0969994X2000007X.
10. M. Thayaparan, M. Valentino, and A. Freitas, "A Survey on Explainability in Machine Reading Comprehension," arXiv Prepr. arXiv2010.00389, 2020. doi: 10.48550/arXiv.2010.00389.
11. Z. Zhang, H. Zhao, and R. Wang, "Machine reading comprehension: The role of contextualized language models and beyond," arXiv Prepr. arXiv2005.06249, 2020. doi: 10.48550/arXiv.2005.06249.
12. Y. Bai and Daisy Zhe Wang, "More Than Reading Comprehension: A Survey on Datasets and Metrics of Textual Question Answering," arXiv Prepr. arXiv2109.12264 , 2021. doi: 10.48550/arXiv.2109.12264.
13. C. Zeng, S. Li, Q. Li, J. Hu, and J. Hu, "A Survey on Machine Reading Comprehension—Tasks, Evaluation Metrics and Benchmark Datasets," Appl. Sci., vol. 10, no. 21, p. 7640, 2020. doi: 10.3390/app10217640.
14. R. Fu, H. Wang, X. Zhang, J. Zhou, and Y. Yan, "Decomposing Complex Questions Makes Multi-Hop QA Easier and More Interpretable," 2021. doi: 10.18653/v1/2021.findings-emnlp.17.
15. E. Perez, P. Lewis, W. T. Yih, K. Cho, and D. Kiela, "Unsupervised question decomposition for question answering," 2020. doi: 10.18653/v1/2020.emnlp-main.713.
16. S. Kundu, T. Khot, A. Sabharwal, and P. Clark, "Exploiting explicit paths for multi-hop reading comprehension," Proc. 57th Annu. Meet. Assoc. Comput. Linguist., pp. 2737–2747, 2018. doi: 10.18653/v1/P18-1254.
17. K. Nishida 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, 2019, pp. 2335–2345. doi: 10.18653/v1/P19-1225.
18. G. P. S. Bhargav et al., "Translucent answer predictions in multi-hop reading comprehension," in Proceedings of the AAAI conference on artificial intelligence, 2020, vol. 34, no. 05, pp. 7700–7707. doi: 10.1609/aaai.v34i05.6322.

19. Y. Cong, Y. Wu, X. Liang, J. Pei, and Z. Qin, "PH-model: enhancing multi-passage machine reading comprehension with passage reranking and hierarchical information," Appl. Intell., pp. 1–13, 2021. doi: 10.1007/s10489-021-02830-2.
20. J. Chen, S. Lin, and G. Durrett, "Multi-hop question answering via reasoning chains," Int. Conf. Learn. Represent., 2020. doi: 10.48550/arXiv.1911.01168.
21. G. Huang, Y. Long, C. Luo, J. Shen, and X. Sun, "Prompting Explicit and Implicit Knowledge for Multi-hop Question Answering Based on Human Reading Process," arXiv Prepr.arXiv2402.19350, 2024. doi: 10.48550/arXiv.2402.19350.
22. Y. Xiao et al., "Dynamically fused graph network for multi-hop reasoning," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 6140–6150. doi: 10.18653/v1/P19-1555.
23. Y. Cao, M. Fang, and D. Tao, "Bag: Bi-directional attention entity graph convolutional network for multi-hop reasoning question answering," 2019 Conf. North Am. Chapter Assoc. Comput. Linguist., pp. 357–362, 2019. doi: 10.18653/v1/N19-1037.
24. X. Li, Y. Liu, S. Ju, and Z. Xie, "Dynamic Reasoning Network for Multi-hop Question Answering," in CCF International Conference on Natural Language Processing and Chinese Computing, 2020, pp. 29–40. doi: 10.1007/978-3-030-66487-6-3.
25. M. Tu, G. Wang, J. Huang, Y. Tang, X. He, and B. Zhou, "Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2704–2713. doi: 10.18653/v1/P19-1261.
26. B. Wang, T. Yao, W. Chen, J. Xu, and X. Wang, "ComQA: Compositional Question Answering via Hierarchical Graph Neural Networks," 2021. doi: 10.48550/arXiv.2104.03222.
27. R. Li, L. Wang, S. Wang, and Z. Jiang, "Asynchronous Multi-grained Graph Network for Interpretable Multi-hop Reading Comprehension," IJCAI Int. Jt. Conf. Artif. Intell., 2021, doi: 10.24963/ijcai.2021/531.
28. P. Gao, F. Gao, P. Wang, J.-C. Ni, F. Wang, and H. Fujita, "ClueReader: Heterogeneous Graph Attention Network for Multi-Hop Machine Reading Comprehension," Electronics, vol. 12, no. 14, p. 3183, 2023. doi: 10.3390/electronics12143183.
29. J. Tang, S. Hu, Z. Chen, H. Xu, and Z. Tan, "Incorporating Phrases in Latent Query Reformulation for Multi-Hop Question Answering," Mathematics, vol. 10, no. 4, 2022, doi: 10.3390/math10040646.
30. Ying Zhang, Fandong Meng, Jinchao Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou, "MKGN: A Multi-Dimensional Knowledge Enhanced Graph Network for Multi-Hop Question and Answering," IEICE Trans. Inf. Syst., 2022. doi: 10.1587/transinf.2021EDP7209.
31. S. Ouyang, Z. Zhang, and H. Zhao, "Fact-Driven Logical Reasoning for Machine Reading Comprehension," in Proceedings of the AAAI Conference on Artificial Intelligence, 2024, vol. 38, no. 17, pp. 18851–18859. doi: 10.1609/aaai.v38i17.31139.
32. A. Ghafouri, H. Naderi, and B. M. Bidgoli, "A Content-based Reasoning Method for Multi-hop Question Answering using Graph Neural Networks," 2024. doi: 10.1007/978-3-031-60475-1-29.
33. Rasouli, M. and Kiani, V. (2024). A survey on deep learning methods for text-based emotion classification: Advances, challenges, and opportunities. Soft Computing Journal, 13(1), 40-57. doi: 10.22052/scj.2023.248812.1126
34. Khalaf Beigi, O. , Bashiri Mosavi, S. A. and Gharloghi, S. (2023). Applying Character-Level Neural Network-Based Sentiment Analysis Model on Persian Comments of the Social Media-Online Store Platforms. Soft Computing Journal, 11(2), 118-133. doi: 10.22052/scj.2023.248311.1094
35. Rasoulzadeh Darabad, R. , Asghari, S. A. , Binesh Marvasti, M. R. and Shahbakhti, K. (2024). Dynamic Multi Level Spatially Aware Clustering with Weighted Cluster Head Selection. Soft Computing Journal, (), -. doi: 10.22052/scj.2024.254853.1241