A Heterogeneous Graph Attention Network for Multi-hop Machine Reading Comprehension

Peng Gao¹†, Feng Gao²†, Jian-Cheng Ni³ and Hamido Fujita⁴,⁵,⁶

¹School of Cyber Science and Engineering, Qufu Normal University, Qufu, Shandong 273165, China.
²School of Computer Science and Technology, East China Normal University, Shanghai 200062, China.
³Network and Information Center, Qufu Normal University, Qufu, Shandong 273165, China.
⁴Faculty of Information Technology, Ho Chi Minh City University of Technology, Ho Chi Minh City 70000, Vietnam.
⁵Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada 18011, Spain.
⁶Research and Regional Cooperation Division, Iwate Prefectural University, Takizawa, Iwate 020-0611, Japan.

†These authors contributed equally to this work.

Abstract

Multi-hop machine reading comprehension is a challenging task in natural language processing, which requires more reasoning ability across multiple documents. Spectral models based on graph convolutional networks grant inferring abilities and lead to competitive results. However, part of them still faces the challenge of analyzing the reasoning in a human-understandable way. Inspired by the concept of the Grandmother Cells in cognitive neuroscience, a spatial graph attention framework named ClueReader was proposed in this paper, imitating the procedure. This model is designed to assemble the semantic features in multi-level representations and automatically concentrate or alleviate information for reasoning via the attention mechanism. The name "ClueReader" is a metaphor for the pattern of the model: regard the subjects of queries as the start points of clues, take the reasoning entities as bridge points, consider the latent candidate entities as the grandmother cells,
and the clues end up in candidate entities. The proposed model allows us to visualize the reasoning graph, then analyze the importance of edges connecting two entities and the selectivity in the mention and candidate nodes, which can be easier to be comprehended empirically. The official evaluations in the open-domain multi-hop reading dataset WIKIHop and Drug-drug Interactions dataset MEDHop prove the validity of our approach and show the probability of the application of the model in the molecular biology domain.

**Keywords:** Machine Reading Comprehension, Heterogeneous Graph, Graph Neural Networks, Attention Mechanism

### 1 Introduction

Machine reading comprehension (MRC) is one of the most attractive and long-standing tasks in natural language processing (NLP). Compared with single paragraph MRC, multi-hop MRC is more challenging since multiple confusing answer candidates are contained in different passages [1]. Models designed for this task are supposed to have abilities to reasonably traverse multiple passages and discover reasoning clues following given questions. For more complexities of the multi-hop MRC task, it is required to pursue more understandable, reliable, and analyzable methodologies to improve the reading performance.

Better understanding biological brains could play a vital role in building intelligent machines [2]. Previous cognitive research in reading can benefit the challenging multi-hop MRC task. Previous research [3] suggests that when we recognize visual or text stimulation, responses from *grandmother cells* can be observed. This physiological concept introduces a hierarchical “sparse” coding scheme in the human mind. It interprets a stupefying indication that a single neuron can respond to only one stimulation out of thousands [3], which is intuitively quite similar to reading and inference in multi-hop MRC somehow:

- **Selectivity.** *The grandmother cells* concept organizes the neurons in a localist coding scheme. It activates some specific of them to make the response to stimulation, which is similar to that we store reasoning evidence maps (neurons) in our minds during reading, and recall related some of them to infer the answer with a question (stimulation) constrained.

- **Specificity.** The concept interprets that brains contain *grandmother* neurons that so specialized being dedicated to the same single person, image, or concept, which is similar to a particular MRC question resulting in the only answer among multiple reading passages and their complex reasoning evidence.

- **Class character.** The amazing selectivity is captured in *the grandmother cells*. However, it must result from computation by much larger networks and the collective operations of many functionally different low-level cells, which is similar to the human multi-hop reading that we usually gather evidence in different levels as much as possible and decide the final answer in some candidate endpoints.
To imitate the grandmother cells in multi-hop MRC, it is supposed to organize the reading evidence as level-classified neurons and perform the selections with the specific question stimulation. As for multi-hop MRC tasks, the hops between two entities could be connected as node pairs and gradually constructed into the reasoning evidence graph taking all the related entities as nodes. This reasoning evidence graph is intuitively represented in a graph structure, which can be empirically considered that it contains the implicit reasoning chains from the question starting to the end answer nodes (entities). Specifically, we generally recall a mountain of related evidence as a node, whatever the form it is (such as a paragraph, a short sentence, or a phrase) to meet the class character, and we coordinate their inter-relationship before carrying out the final results.

The appearance of graph neural networks (GNNs) arouses us that operating on the graph and manipulating the structured knowledge can support relational reasoning [4] in a sophisticated and flexible pattern, which can be considered as the fundamental implementation of the grandmother cells regarding the cells as nodes in the graph and collecting the evidence in multi-classified aspects of nodes’ representations. Further, the spatial graph attention networks (GATs) using attention mechanisms perform the selectivity in reasoning evidence graph in the grandmother cell way. This work has two main contributions:

1. A novel multi-hop MRC architecture named ClueReader is proposed to imitate the neuroscience concept the grandmother cells, which organizes reading evidence as nodes forming the reasoning graph.
2. The ClueReader performs the selectivity in the reasoning evidence for reading question answering, which is visualized inner status, and they can be further analyzed.

The rest of the article is organized as follows. Section 2 describes the related work to multi-hop MRC, and Section 3 proposes the ClueReader imitating the grandmother cells for multi-hop MRC. Experimental evaluations are conducted in Section 4, and conclusions are summarized in Section 5.

2 Related Work

2.1 Sequential Reading Models for Multi-hop MRC

The sequence models are first used for single passage MRC tasks, and most of them are based on Recurrent Neural Networks (RNNs) or their variants. As the attention mechanism is introduced into NLP tasks, their performance has been significantly improved [5, 6]. In the initial benchmarks of the QANGAROO [7], a dataset for multi-hop MRC dataset, the milestone model BiDAF [5] was firstly applied to evaluate its performance in the multi-hop MRC task. It represented the context at different levels and utilized a bi-directional attention flow mechanism to get query-aware context representation, then it was used for predictions.

Some research [8–11] argues that the independent attention mechanism (BERT-style models) applied on sequential contexts even can outperform the former RNN-based approaches in various NLP downstream tasks including MRC. When
the sequential approaches were applied to the multi-hop MRC tasks, however, they suffered from the challenge that the super-long contexts — to adapt the design of the sequential requirement, multiple passages are concatenated into one passage — resulted in the dramatically increased calculation and time consumption. The long-sequence architecture, Longformer [11], overcomes the self-attention restriction and allows the length of sequences to be up from 512 to 4,096. Then, it concatenated all the passages into a long sequential context for reading. The Longformer modified the Question Answering (QA) methodology proposed in the BERT [8]: the long sequential context consisted of a question, candidates, and passages, which were separated by special tags that were applied to the linear layers to output the predictions, while it still encounters the memory requirements leading to include the first 4,096 length sequence.

Although the approaches above have shown effectiveness, [12] indicates that model reasoning is not robust enough. We consider that there are still two main challenges that should be further addressed: (1) With the expansion of the problem scale and the reasoning complexity, the token-limited problem may appear again eventually. For instance, a fullwiki setting task in HOTPOTQA requires models predict answers in the scope of the entire WIKIPEDIA. It is hard to imagine how a huge search space is built based on a large amount of text. (2) Some models who simply concatenate text to a long context lack some logical relationships, which is unconvinced in terms of their reasoning. Thus, the approaches based on the GNNs were proposed to improve the scalability or explainability in multi-hop MRC.

2.2 Graph Neural Networks for Multi-hop MRC

Reasoning about explicitly structured data, in particular graphs, has arisen at the intersection of deep learning and structured approaches [4]. As the representative graph methodology, the Graph Convolutional Networks (GCNs) [13] were widely applied in multi-hop MRC approaches. The CogQA [14] was founded on the dual process theory [15, 16], and it divided the multi-hop reading process into two stages: the implicit extraction (System I) based on the BERT and the explicit reasoning (System II) established in GCNs. System I extracted the answer candidates and useful next-hop entities from passages for the cognitive graph construction, then System II would update the entities’ representations and predict the final answer in the GCN’s message passing way. In this procedure, the passages were selected, which were not put in the system at once. As a result, CogQA could keep its scalability when the scope of reading materials was massive. Entity-GCN [17] extracted all the text spans matching the candidates as nodes and obtained their representations from the contextualized word embedding ELMo [18], then passed them to the GCNs module for the inferring process. Based on Entity-GCN, BAG [19] added the Glove word embeddings and two manual features, i.e., named-entity recognition (NER) and part-of-speech (POS) tags, to reflect the semantic properties of tokens. And on account of the full usage of the question’s contextual information, it applied the bi-directional attention mechanism, both node2query and query2node, which aimed to obtain the query-aware nodes’ representations in the
reasoning graph for better predictions. *Path-based GCN* [20] introduced related entities in the graph more than the nodes merely match to the candidates to enhance the performance of the model. *HDE* [21] introduced the heterogeneous nodes into GCNs, which contain different granularity levels of information. Besides, *KA-DGN* [22] is proposed and designed to a dynamic graph neural network to further tackle reading over multiply scattered text snippets. Furthermore, [23] and [24] propose knowledge-aware and evidence-aware GNN reading models separately, which newly integrate dependency relations or multiple pieces of evidence from multiple paragraphs.

However, the approaches above-mentioned still suffered from explaining the reading process, especially in the graph neural networks, which aroused our interest in the selectivity of this procedure.

### 3 Methodology

In this section, we will introduce our design and implementation of the proposed model, *ClueReader*, and we illustrate its diagram in Fig. 1.

![Fig. 1 The diagram of our proposed ClueReader: a heterogeneous graph attention network for multi-hop MRC.](image)

#### 3.1 Task Formalization

The given query $q = (s, r, ?)$ is in a triple form, where $s$ is the subject entity and $r$ is the query relation (*i.e.*, predicate), and $q$ can be converted into a sequential form $q = \{q_1, q_2, ..., q_m\}$, where $m$ is the number of tokens in query. Then a set of candidates $C_q = \{c_1, c_2, ..., c_z\}$ and a series of supporting documents $S_q = \{s_1, s_2, ..., s_n\}$ containing the candidates are also provided, where $z$ is the number of the given candidates, and $n$ is the number of the given multiple documents, the subscript $q$ means the two sets are constrained by the query $q$. Besides, $S_q$ is provided in random order, and without $S_q$ the answer of the query $q$ could be multiple. Our goal is to identify the single correct answer $a^* \in C_q$ by reading $S_q$. 

---

*ClueReader 5*

---
3.2 Encoder Layer

We use the Bidirectional Long Short-Term Memory (Bi-LSTM) \cite{25,26} to encode the sequential contexts, and the pre-trained word embeddings are applied to improve the encoding performance. To simplify the expression, we consider the left to right sequence context $\vec{h}_i$, i.e., the forward pass, in Bi-LSTM, and their formulas are as follows:

\begin{align*}
  i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\
  f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
  g_t &= \tanh(W_{xc}x_t + W_{hc}h_{t-1} + W_{cc}c_{t-1} + b_c) \\
  c_t &= i_t g_t + f_t c_{t-1} \\
  o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\
  h_t &= o_t \tanh(c_t)
\end{align*}

where $i_t, f_t, o_t, c_t$ represent the inputs, forget gate, outputs, and cell states at time $t$, respectively, and $\sigma$ is the activation function. The right to left sequence context $\vec{h}_i$, i.e., the backward pass, can be calculated by the same token. Then, in the encoder layer the output of the $i$-th word is shown as follows:

\begin{equation}
  h_i = [\vec{h}_i \vec{h}_i]
\end{equation}

where the “:” means that we concatenate the forward pass and backward pass outputs to represent $h_i$. It is desirable to use three independent Bi-LSTM to encode the sequential contexts, w.r.t. the support documents $S$, candidates $C$, and query $q$. And their outputs are $H^i_s \in \mathbb{R}^{l_i \times h}$, $H^j_c \in \mathbb{R}^{l_j \times h}$ and $H_q \in \mathbb{R}^{l_q \times h}$, respectively, where $i$ is the $i$-th document, $j$ is the $j$-th candidate, $l$ is the sequence length, and $h$ is the encoder’s output dimension.

3.3 Heterogeneous Reasoning Graph Construction

The concept of the grandmother cells reveals that the brains of monkeys, like those of humans, contain neurons that are so specialized they appear to be dedicated to a single person, image, or concept. This amazing selectivity is uncovered in a single neuron, while it must result from computation by a much larger network \cite{3}.

We heuristically consider this procedure in multi-hop reading could be summarized as three steps: (I) the query (or the question) locates the related neurons in low-level, and then they pass the stimulation to the higher-level neurons to trigger the computation; (II) the higher-level neurons would begin to respond to increasingly broader portions of other neurons for reasoning, and to avoid the broadcast storm, the informative selectivity takes place in this step; (III) at the top level, some independent neurons are supposed to be responsible for the computations that happened in step II, and we regard them as the grandmother cells and expect they could provide the proper results that are corresponding with the query.

We attempt to imitate the grandmother cells in our reading procedure, then we present our reasoning graph as consistent as possible to the three steps mentioned.
above. The graph is denoted as $G = \{\mathcal{V}, \mathcal{E}\}$, and we define five different types of nodes $\mathcal{V}$ which are similar to the neurons, and ten kinds of edges among the nodes $\mathcal{E}$.

### 3.3.1 Reasoning Graph Nodes Construction

- **Subject Nodes** — As the form of query $q$, the subject entity $s$ of a question is given in $q = (s, r, ?)$, and all the text spans matching to $s$, we extract them from documents and regard them as the low-level neurons and the start points to open up the reading clues triggering the further computations. The subject nodes are denoted as $\mathcal{V}_{\text{sub}}$.

- **Reasoning Nodes** — In light of the requirement of the multi-hop MRC, there are some gaps between the subject entities and candidates. To build bridges between the two and make the reasoning clues as complete as possible, we replenish those clues with the named recognition entities and nominal phrases from the documents containing the question subjects and answer candidates. The reasoning nodes are marked as $\mathcal{V}_{\text{rea}}$.

- **Mention Nodes** — A series of candidate entities are given in $\mathcal{C}$, and we traverse the documents and extract the text spans corresponding to each candidate. The mention nodes are presented as $\mathcal{V}_{\text{men}}$.

- **Support Nodes** — As described in [3], we consider that multi-angle representations may contribute to the reading process, thus the support documents containing the above nodes are introduced to $G$ as information resources. These support nodes are notated as $\mathcal{V}_{\text{sup}}$.

- **Candidate Nodes** — The imitation of the grandmother cell, the high-level neurons, we use candidate node to converge the representation from $\mathcal{V}_{\text{men}}$ which are consistent with $\mathcal{C}_q$. We look upon them as the end of reading clues and the crucial output to provide the final prediction, and they are denoted as $\mathcal{V}_{\text{can}}$.

### 3.3.2 Reasoning Graph Edges Definitions

- **$\mathcal{E}_{\text{sup2can}}$** — If the content of $v^i_{\text{sup}}$ contains the text of $v^j_{\text{can}}$, the two nodes will be connected as $e^{ij}_{\text{sup2can}}$.

- **$\mathcal{E}_{\text{can2can}}$** — All the $v_{\text{can}}$ nodes will be fully connected.

- **$\mathcal{E}_{\text{sup2men}}$** — If the content of $v^i_{\text{sup}}$ contains the text of $v^j_{\text{men}}$, the two nodes will be connected as $e^{ij}_{\text{sup2men}}$.

- **$\mathcal{E}_{\text{can2men}}$** — If $v^j_{\text{men}}$ is an appearance of $v^i_{\text{can}}$, the two nodes will be connected as $e^{ij}_{\text{can2men}}$.

- **$\mathcal{E}_{\text{sub2rea}}$** — If $v^j_{\text{sub}}$ and the extracted $v^j_{\text{rea}}$ appear in the same document, the two nodes will be connected as $e^{ij}_{\text{sub2rea}}$.

- **$\mathcal{E}_{\text{rea2rea}}$** — If two reasoning nodes, $v^j_{\text{rea}}$ and $v^j_{\text{rea}}$, are extracted from the same document or represent the same entity, the two nodes will be connected as $e^{ij}_{\text{rea2rea}}$.

- **$\mathcal{E}_{\text{rea2men}}$** — If the extracted $v^j_{\text{rea}}$ and $v^j_{\text{men}}$ appear in the same document, the two nodes will be connected as $e^{ij}_{\text{rea2men}}$. 
ClueReader

- $E_{edgesin}$ — If two mention nodes, $v_{men}^i$ and $v_{men}^j$, exist in the same document, the two nodes will be connected as $e_{edgesin}^{ij}$.
- $E_{edgesout}$ — If two mention nodes corresponding to one candidate, $v_{men}^i$ and $v_{men}^j$, appear in different documents, the two nodes will be connected as $e_{edgesout}^{ij}$.
- $E_{complete}$ — If two nodes are not connected by the above rules, draw an edge between them.

After the node extraction and the edge connection, the constructed reasoning graph can be illustrated in Fig. 2. In our reasoning graph, the clue-reading path can be represented by $V_{sub} \leftrightarrow V_{rea} \leftrightarrow V_{men} \leftrightarrow V_{can}$, whose edges are covered by $E_{sub2rea}$, $E_{rea2men}$, and $E_{can2men}$. $E_{edgesout}$ and $E_{rea2rea}$ give the model the ability to transfer information across documents and edges in $E_{sup2sub}$, $E_{sup2can}$, and $E_{sup2men}$ are responsible to supplement the multi-angle textual information from the documents. Furthermore, the $E_{can2men}$ could gather all the information of the mentioned nodes corresponding to the candidates and then pass their representations to the output layer to realize the imitation of the grandmother cells.

Fig. 2 The reasoning graph in ClueReader. Different nodes are filled with different colors, and the edges can also be distinguished by the types of lines. Subject nodes are gray, reasoning nodes are orange, mention nodes are green, candidate nodes are blue, and document nodes are red. To illustrate the graph clearly, the $E_{complete}$ are omitted. The nodes in the light yellow square are all selected to input to the two MLP obtaining the prediction score distribution.
3.4 Heterogeneous Graph Attention Network for Multi-hop Reading

3.4.1 Query-aware Contextual Information

Following HDE \cite{21}, we use the co-attention \cite{27} to combine the query contextual information and documents. And it is also applied to the other semantic representations that require reasoning consistent with the query. To represent the query-aware support documents, it can be calculated as follows:

$$A_{qs}^i = H_s^i (H_q)^\top \in \mathbb{R}^{l_s \times l_q}$$  \hspace{1cm} (8)

where $A_{qs}$ is similarity matrix for two sequences, between the $i$-th support document $H_s^i \in \mathbb{R}^{l_s \times h}$ and query $H_q \in \mathbb{R}^{l_q \times h}$, and $h$ is the dimension of context. Then, the query-aware representation of documents $S_{ca}$ is computed as follows:

$$K_q = \text{softmax} (A_{qs}^\top) H_s \in \mathbb{R}^{l_s \times h}$$  \hspace{1cm} (9)

$$K_s = \text{softmax} (A_{qs}) H_q \in \mathbb{R}^{l_q \times h}$$  \hspace{1cm} (10)

$$D_s = \text{BiLSTM} (\text{softmax} (A_{qs}) K_q) \in \mathbb{R}^{l_s \times h}$$  \hspace{1cm} (11)

$$S_{ca} = [K_s D_s] \in \mathbb{R}^{l_s \times 2h}$$  \hspace{1cm} (12)

In order to project the sequence into a fixed dimension and output the representation $N_{sup}$ of $V_{sup}$ for graph optimization, the self-attention \cite{27} is introduced to summarize the contextual information:

$$j_s = \text{softmax} (\text{MLP} (S_{ca})) \in \mathbb{R}^{l_s \times 1}$$  \hspace{1cm} (13)

$$N_{sup} = j_s^\top S_{ca} \in \mathbb{R}^{1 \times 2h}$$  \hspace{1cm} (14)

Besides the query-aware support documents, the co-attention and self-attention are used to other sequential representations to generate their query-aware node representations.

3.4.2 Message Passing in the Heterogeneous Graph Attention Network

We present the messaging passing in the heterogeneous graph attention network for reading within multiple relations in diverse nodes. The input of this module is a graph $G = \{\mathcal{V}, \mathcal{E}\}$ and node representations $\mathcal{N}$. Initially, a shared weight matrix $W_n$ is applied to $\mathcal{N}$, then the attention coefficients and nodes attention coefficients are computed as:

$$e_{ij} = \text{MLP} (W_n n_i W_n n_j)$$  \hspace{1cm} (15)

$$\alpha_{ij} = \text{softmax}_j (e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_j} \exp(e_{ik})}$$  \hspace{1cm} (16)

where $e_{ij}$ is the attention coefficients indicating the importance of node $j$’s features to node, and $\alpha_{ij}$ is normalized across all node $i$’s structure neighbors $\mathcal{N}_i$. The attention
mechanism is the charge of selectivity with node interdependence, which enables us to show how the nodes take effect during the reasoning.

Considering different types of edge defined in 3.3.2, we model the relational edges basing on vanilla graph attention networks [28]:

\[
\mathbf{n}_i^{l+1} = \frac{1}{k} \sum_{k=1}^{K} \sigma \left( \sum_{j \in \mathcal{N}_i} \sum_{r \in \mathcal{R}_{ij}} \frac{1}{|\mathcal{N}_r|} \alpha_{r_{ij}}^{k,i} \mathbf{W}^{k,i}_{r_{ij}} \mathbf{n}_j^l \right)
\]

(17)

where \(n_i^l \in \mathbb{R}^{1 \times 2h}\) is the hidden state of node \(i\) in the \(l\)-th layer, all the GAT layers are parameter shared, \(k\) is the \(k\)-th head following [9, 28], \(\mathcal{R}\) is the set of all types of edges in \(\mathcal{E}\), \(\mathcal{N}_i\) is the set of all the neighbors of node \(i\), and \(\alpha_{r_{ij}}^{k,i}\) is normalized attention coefficients computed by the \(k\)-th attention mechanism with relation \(r\), which is presented in [28] and “\(\parallel\)” indicates the concatenation from \(K\)-head attention mechanism.

The message passing is a key component of our model. To echo the selectivity of the grandmother cells, we use the attention mechanism to perform selection (i.e., activate or deactivate) on key node pairs in our reasoning graph, and we empirically regard this process as the reading reasoning in the graph.

### 3.4.3 Gating Mechanism

previous study [13] has shown the GNNs are suffering from the smoothing problem when they are calculated by stacking many layers, thus, we overcome this issue by applying a question-aware gating [20] and a general gating mechanism [29] to optimize the procedure.

\[
\mathcal{H}_q = \text{BiLSTM}(H_q)
\]

(18)

\[
w_{ij} = \sigma \left( \mathbf{W}_q^T \left[ n_i^l \parallel \mathcal{H}_{q_{ij}} \right] \right)
\]

(19)

\[
\alpha_{i,j}^{\text{gate}} = \frac{\exp(w_{ij})}{\sum_{k=1}^{m} \exp(w_{ik})}
\]

(20)

\[
p_{ij}^l = \sum_{j=1}^{m} \alpha_{i,j}^{\text{gate}} \mathcal{H}_{q_{ij}}
\]

(21)

\[
\beta_{i}^l = \sigma(\mathbf{W}_s^T [p_{i}^l n_i^l])
\]

(22)

\[
n_{i}^{l'} = \beta_{i}^l \odot \tanh \left( p_{i}^l \right) + \left( 1 - \beta_{i}^l \right) \odot n_{i}^l
\]

(23)

where \(\mathcal{H}_q\) is the query representation given by a dedicated Bi-LSTM encoder to keep consistency with the dimension of node features \(\mathcal{N}\), \(j\) is the \(j\)-th query token, \(m\) is the query length, \(\sigma\) is sigmoid function, and the \(\odot\) indicates element-wise multiplication. Then the general gating mechanism is introduced as follows:

\[
\mathbf{x}_i^l = \sigma(\mathbf{MLP}[n_{i}^{l'} n_i^l])
\]

(24)
\[ n_i^{l+1} = x_i^l \odot \text{tanh} \left( n_i'^l \right) + \left( 1 - x_i^l \right) \odot n_i^l \quad (25) \]

### 3.5 Output Layer

After the node representation updating, we use two two-layer multilayer perceptrons, \( \text{MLP}_{\text{can}} \) and \( \text{MLP}_{\text{men}} \), to transform the node features to the prediction scores. All the \( \mathcal{N}_{\text{can}} \) and \( \mathcal{N}_{\text{men}} \) nodes from \( G \) are calculated by Equation (26) outputting the prediction score distribution \( a \):

\[ a = \gamma \times \text{MLP}_{\text{can}}(\mathcal{N}_{\text{can}}) + (1 - \gamma) \times \max(\text{MLP}_{\text{men}}(\mathcal{N}_{\text{men}})) \quad (26) \]

where \( \max(\cdot) \) takes the maximum mention node score over \( \text{MLP}_{\text{men}} \), then the two parts are summed with the effect of a harmonic \( \gamma \) as the final prediction score distribution.

### 4 Experiments

In this section, we present the performance of our model in the QANGAROO [7] dataset and the evaluation in detail. Then, the ablation study and the visualization will introduce the benefit of the model. Finally, the case study shows the relationship between the answers output from the models and human reading results.

#### 4.1 Dataset for Experiments

QANGAROO is a multi-hop MRC dataset containing two independent datasets WIKIHOP and MEDHOP, which are the open-domain field and molecular biology field, respectively. Both WIKIHOP and MEDHOP are divided into three subsets: train set, development set, and the undisclosed test set, which is used for official evaluation. The dataset sizes are shown in Table 1.

| Dataset     | Train | Development | Test  | Total   |
|-------------|-------|-------------|-------|---------|
| WikiHop     | 43,738| 5,129       | 2,451 | 51,318  |
| MedHop      | 1,620 | 342         | 546   | 2,508   |

The WIKIHOP is created from the WIKIPEDIA (as the document corpus), and the WIKIDATA (as structured knowledge triples) [7]. A sample from the dataset is shown as Fig. 3(a). In this sample, the query, \( (\text{located_in_the_administrative_territorial_entity}, \ \text{hampton_wick_war_memorial}, \ ?) \), requires us to answer the administrative territory of the Hampton Wick War Memorial. To predict it, a named recognition entity Hampton Wick is extracted from the seventh support document, and it links to the same tokens in the zeroth support document where the correct candidate answer appears as well. The reasoning clue, Hampton Wick War Memorial ↔ Hampton Wick # 1 ↔ Hampton Wick # 2 ↔
ClueReader

London Borough of Richmond upon Thames, presents our model’s procedure for the multi-hop MRC task.

To validate whether the dataset can be consistent with the formalization of the multi-hop MRC, the dataset founder asked human annotators to evaluate the samples in the WIKIHop development set and test set. For each sample in the two sets, at least 3 annotators participated in the evaluation, and they were required to answer three questions [7]:

• whether they knew the fact before;
• whether the fact follows from the texts (with options “follows”, “likely”, and “not follows”);
• whether multiple documents are required to answer the question

All the samples in the test set are human-selected and are labeled by a majority of annotators with “follows” and “multiple documents required”. Annotators merely note the samples in the development set without the selection.

The MEDHop dataset is constructed using the DRUGBANK as certain knowledge. Then the creators extract the research paper abstracts from MEDLINE as corpus, and the dataset aims to predict the drug-drug interactions (DDIs) after reading the texts. The promise of applying the multi-hop methods in this prediction is finding and combining individual observations that can suggest previously unobserved DDIs from inferring and reasoning the prior public knowledge in contents rather than some costly experiments. The only query type is interacts_with. A sample given in [7] is illustrated as Fig. 3(b) and note that accession numbers replace the medical proper nouns (e.g., DB00007, DB06825, DB00316) rather than the names of drugs and human proteins (e.g., Leuprolide, Triptorelin, Acetaminophen) in practice.

4.2 Experiments Settings

We exploit NLTK [30] toolkit to tokenize the support documents and candidates, then split the query \( q = \{s, r, ?\} \) into relation \( r \) and subject entity \( s \). All the text spans matching with candidates \( C_q \) are extracted as mention nodes \( \mathcal{V}_{men} \), and the SPACY\(^1\) is used to extract the named entities and noun phrases from texts as reasoning nodes \( \mathcal{V}_{rea} \). We concatenate GloVe [31] and n-gram character embeddings [32] to obtain 400-dimensional word embeddings, which are input to encoder layer. The out-of-vocabulary words are presented with random vectors. The word embedding is set to be fixed in WIKIHop’s experiment and trainable in MEDHop’s. We implement our ClueReader model with PyTorch and PyTorch Geometric [33]. The NetworkX [34] is utilized to visualize the reading graph, the weights of node pairs, and the selections of nodes.

4.3 Results and Analyses

In Table 2 we present the performances of ClueReader in the development and test sets of WIKIHop and MEDHop, and compare them with previously published models mainly based on GNNs. Our model has improved the accuracy of the HDE based

---

\(^1\)https://spacy.io
Question:
(located_in_the_administrative_territorial_entity, hampton_wick_war_memorial, ?)

---

Document #7:
The Hampton Wick War Memorial is situated on the Hampton Wick side of Kingston Bridge, in the south-west suburbs of London, England…

Document #0:
Hampton Wick, formerly a village, is a Thames-side area of the London Borough of Richmond upon Thames, which… It is buffered by a public green space, one of the Royal Parks of London from Hampton and Hampton Hill…

Candidates:
{hampton, hampton_hill, ..., london_borough_of_richmond_upon_thames, scotland}

Answer:
london_borough_of_richmond_upon_thames

(a) A sample from the WikiHop.

Question:
((Leuprolide, interacts_with, ?))

Document #1:
Leuprolide... elicited a long-lasting potentiation of excitatory postsynaptic currents... GnRH receptor-induced synaptic potentiation was blocked... by Progonadoliberin-1, a specific GnRH receptor antagonist...

Document #2:
... our research to study the distribution, co-localization of Urofollitropin and its receptor[... and co-localization of Urofollitropin and GnRH receptor...

Support #3:
Analyses of gene expression demonstrated a dynamic response to the Progonadoliberin-1 superagonist Triptorelin.

Candidates:
{Triptorelin, Urofollitropin}

Answer:
Triptorelin

(b) A sample from the MedHop.

Fig. 3 The subject entities, reasoning entities, mention entities, and candidate entities are set to gray, orange, green and blue color, respectively. And the occurrence of the correct answer is drawn on a square frame outside.
Table 2  The performance of the proposed ClueReader in the development and test sets of the WikiHop and MedHop, and the comparisons to other published approaches on the leaderboard.

| Single models                  | WikiHop Accuracy (%) | MedHop Accuracy (%) |
|--------------------------------|----------------------|---------------------|
|                                | Dev  | Test | Dev  | Test |
| Coref-GRU [36]                 | 56.0 | 59.3 | -    | -    |
| MHQA-GRN [35]                  | 62.8 | 65.4 | -    | -    |
| Entity-GCN [17]                | 64.8 | 67.6 | -    | -    |
| HDE [21]                       | **68.1** | 70.9 | -    | -    |
| BAG [19]                       | 66.5 | 69.0 | -    | -    |
| Path-based GCN (Glove) [20]    | 64.5 | -    | -    | -    |
| Document-cue [7]               | -    | 36.7 | -    | 44.9 |
| FastQA [7]                     | -    | 25.7 | -    | 23.1 |
| TF-IDF [7]                     | -    | 25.6 | -    | 9.0  |
| BiDAF [7]                      | -    | 42.9 | -    | **47.8** |
| ClueReader                     | 66.5 | **72.0** | 48.2 | 46.0 |

ClueReader

on the heterogeneous GCNs in the test set from 70.9% to 72.0% and Path-based GCN (with GloVe word embedding setting) in dev set from 64.5% to 66.9%, while Path-based GCN using GloVe and ELMo surpassed our model by 0.5% in the test set, which confirms the conclusion that the initial representations of nodes is extremely critical [20]. However, limited by the architecture and computing resources, we have not used powerful contextual word embeddings like the ELMo and the BERT in our model, which can be further addressed. Compared to the other GNN-based reading models [17, 19, 35] and the sequential models [7, 36], our model has the higher accuracy. And we notice that we are the first GNN-based model applied to the MedHop, although the accuracy is lower by 1.8% than the BiDAF, we argue that the possible reason due to the failure in extracting the reason node of the SPACY toolkit, which means the bridge entities are not complete.

In order to analyze the scalability of our model, we divide the development set into 6 groups according to the number of support documents and then report the accuracy in each group. The grouped accuracies of the WikiHop is illustrated as Fig. 4, and our model shows the competitive results 73.59% and 63.57% in the group of “1-10” and “11-20”, which the total number of samples in these two groups is 4,039 accounting for 95% of the development set. The lowest accuracy of 55.74% appears in the group of “41-50”. However, it raises to 62.5% in “51-62”, which shows the scalability of our model is effective. Besides, the grouped accuracies in the MedHop are shown in Fig. 5, and the accuracy of each group is neck-and-neck. The highest and second-highest accuracies, i.e., 60.00% and 51.85%, are in “31-40” and “21-30”, which the lowest and second-lowest accuracies, i.e., 0% and 35.59%, are in “1-10” and “51-62”. In particular, the result in the “51-64” of the MedHop is against the “51-62” of the WikiHop, which inspires that we are supposed to concentrate on the difference between the open-domain and molecular textual contexts. In sum, the results in the different number of support documents show that our model has contributed to the scalability of the multi-hop MRC tasks.
Fig. 4  The statistics of the model’s performance on the different number of support documents in the WikiHop development set.

Fig. 5  The statistics of the model’s performance on the different number of support documents in the MedHop development set.
Table 3  The Performance on the WikiHop Development Subsets

| Annotation                  | Accuracy (%) |
|-----------------------------|--------------|
| follows fact                | 74.9         |
| requires multiple document  | 74.0         |
| requires single document    | 74.0         |
| likely follows fact         |              |
| requires multiple document  | 71.4         |
| requires single document    | 71.4         |
| “not follows” is not given  | 71.5         |

As mentioned above, the WikiHop development set has annotated the consistency between facts and documents, and whether multiple documents are required to reason the question, we perform our models in five categories as the following strategies. In each category, all three annotators annotated: (1) “requires multiple documents” and “follows fact”; (2) “requires single document” and “follows fact”; (3) “requires multiple documents” and “likely follows fact”; (4) “requires single document” and “likely follows fact”; (5) “not follows” is not given. The performance of our model is presented in Table 3. We observe that our model presents the best performance of 74.9% in the samples which follow the facts and require multiple passages. This phenomenon proves the model’s effectiveness in pure multi-hop MRC tasks. Then, the second-best result of 74.0% presents in the samples following the facts and requiring the single document, which supports that our model is also effective in single-passage MRC tasks. Further, we figure that authenticity can seriously impact the accuracy of our prediction. Those categories may not follow the fact suffer from the worse results, i.e., 71.4%, 71.4%, 71.5%, respectively, in the groups of “likely follows the fact (one-document and multi-documents)” and “not follows” is not given. The same analysis is infeasible in the development set of MedHop since the document complexity, and the number of documents per sample is significantly larger compared to the WikiHop [7].

4.4 Ablation Study

We propose five types of nodes in $G$, in order to analyze how they took effect in reasoning, we remove the edges with specific connections and make the nodes be isolated to evaluate the performance in the subset of the WikiHop’s development set, i.e., “not follows” is not annotated. And we also test the model without the message passing in the graph $G$. The ablated performance is shown as Table 4.

In WikiHop, it reveals that the proposed heterogeneous graph attention network is the most effective component of our architecture. Without its contribution, the accuracy would decrease by 18.76%. When we block the nodes by groups, we argue that the support nodes contribute 9.43% absolutely, the mentioned nodes dedicate 8.11%, then the candidate nodes contribute 5.58%. And as to reason nodes and subject nodes, we consider the small quantities contained in the graph leading to low status in contributions. However, we observe considerably different performances between the WikiHop and MedHop. As the results show in Table 4, the most effective part of the model changed into the mentioned nodes. When we block the mentioned nodes in the graph, the accuracy decreased significantly by 43.28%, but
Table 4 Ablation Performance on the QA$\text{NGAROO}$ Development Subset

| Model    | WikiHop  | Δ | MedHop  | Δ |
|----------|----------|---|---------|---|
| Full Model | 71.45    | - | 48.25   | - |
| w/o GAT   | 52.69    | 18.76 | 37.72   | 10.53 |
| w/o $N_{sub}$ | 70.95 | 0.5 | 47.37 | 0.88 |
| w/o $N_{men}$ | 63.34 | 8.11 | 4.97 | 43.28 |
| w/o $N_{rea}$ | 70.77 | 0.68 | 47.37 | 0.88 |
| w/o $N_{sup}$ | 62.02 | 9.43 | 48.54 | -0.29 |
| w/o $N_{can}$ | 65.87 | 5.58 | 44.77 | 3.48 |

Table 5 The ablation studies of the hyper parameters of GAT layers and weights of the grandmother cells in reasoning graph predictions.

| Hyperparameters | value | Acc. of WikiHop | Acc. of MedHop |
|-----------------|-------|-----------------|----------------|
| $l$             | 3     | 57.8            | 42.4           |
|                 | 4     | 58.5            | 43.3           |
|                 | 5     | **66.5**        | **48.2**       |
|                 | 6     | 64.2            | 45.0           |
| $\gamma$        | 0     | 59.7            | 42.7           |
|                 | 0.5   | 66.1            | 44.2           |
|                 | 1.0   | **66.5**        | **48.2**       |
|                 | 1.5   | 59.1            | 43.3           |

the graph reasoning should not be underestimated, which contributes 10.53% to accuracy. Meanwhile, we find that support nodes have negative effects on the prediction by 0.29% decrement, which is diametrically opposite compared to its performance on the WIKIHOP development subset.

In Table 5, we present the models’ performances with different hyperparameters, especially the number of stacked GAT layer (a.k.a the number of hops) and the weight of the grandmother cells. The number of GAT layers controls that apply how many parameter-sharing GAT layers are in the reasoning graph. In WIKIHOP, we obtain the highest accuracy (66.5%) when we stack the graph with 5 layers, and the model with 3 or 4 GAT layers has poorer performance (57.8% or 58.5% respectively). Besides, when the number of GAT layers is 6, the accuracy drops 2.3% compared with the best performance. Furthermore, as the final prediction illustrated in Equation (26), $\gamma$ coordinates the mention nodes and the candidate nodes (the grandmother cells), we present the models’ performances with different $\gamma$ settings in Table 5. The best performance appears when $\gamma$ is set to 1. However, if we give it too much weight, i.e., $\gamma = 1.5$, the accuracy decreases by 7.4%, which is even worse than when we set $\gamma$ to 0 (59.7%), which persuades us that we should not ignore the effect of much larger networks behind the grandmother cells. In MEDHOP, we observe similar phenomena with different hyperparameter settings. When the number of hops is 5, and $\gamma$ is 1, the model presents the best performance at about 48.2%. We suspect that when we stack a few of GAT layers, the nodes’ messages cannot pass sufficiently among the
Fig. 6 The visualizations of the reasoning graphs on the WIKIHop development samples that are correctly answered. A thicker edge corresponds to a higher attention weight, the darker green nodes or the darker blue nodes has higher values output among their same type of nodes.

reasoning graph. Otherwise, when stacking too many GAT layers, the graph over-smoothing problem would lead to an accuracy drop. We also empirically consider that the models with higher $\gamma$ may lose semantic information from context and lead to the prediction accuracy dropping, which also fits the grandmother cells theory that
Fig. 7  The visualizations of the reasoning graphs on the MEDHOP development samples that are correctly answered. A thicker edge corresponds to a higher attention weight, and the darker green nodes or the darker blue nodes has higher values output among their same type of nodes.

before the final predicting determination, a huge background network calculation should be done implicitly.
4.5 Visualization

Compared to those spectral GNNs reading approaches, our proposed heterogeneous reasoning graph \textit{ClueReader} is based on the non-spectral approach, which allows us to analyze how the nodes interact with each other in various relations and how the connections take effects among nodes. We visualize the predictions in our heterogeneous reasoning graph on the \textit{WikiHOP} and \textit{MedHop} in Fig. 6 and Fig. 7, respectively, and the different types of nodes are drawn by different colors (subject nodes are gray, reasoning nodes are orange, mention nodes are green, candidate nodes are blue, and document nodes are red) and their edges are demonstrated as different thickness lines to reflect the selections of node pairs. The thicker edges are, the more important they learned from the training process. Considering the answer determination should not only infer by the weight edges but also from the output layer projects the representation of the node to $\mathbb{R}^{1 \times 2h}$ and accumulate score from $\mathcal{N}_{\text{can}}$ and $\mathcal{N}_{\text{men}}$, we use the transparency of the nodes to respond to the outputs: the darker the nodes are, the higher the values output from the output layer. Due to the output values being quite different, some mention nodes and candidate nodes are almost transparent. The weight graph provides the evidence during the reading process and the analysis of DDIs, it passes the messages in accordance with the \textit{grandmother cells} concept that not only one node becomes effective, but the cluster behind it plays a synergistic effect. We learn more about our model in visualization. For instance, the node transparency differentiation in \textit{MedHop} is significantly lower than \textit{WikiHOP}, which indicates that drug features are not sufficiently learned, leading to the convergence of node features and increasing the difficulty of classification prediction. This issue can be further addressed.
To better understand the predictions of the model and make a contribution for the further study, we generate HTML files of samples as shown in Fig. 8, try to analyze whether the text spans containing in the max-score nodes can make sense with human’s answer after reading.

5 Conclusion

We present the ClueReader, a heterogeneous graph attention network for multi-hop machine reading comprehension, which is inspired by the grandmother cells concept from cognitive neuroscience. The network contains several clue reading paths from the subject of the question and ends with the given candidates’ entities. We take the reasoning nodes and mention nodes to complete the process and use document nodes to add supernumerary semantic information. We apply our methodology in QANGAROO, a multi-hop MRC dataset, and the official evaluation supports the effectiveness of our model in open-domain QA and molecular biology domain usage.

Several potential issues could be further addressed, such as introducing intermediate supervision signals during the semi-supervised graph learning, the enhancement of using external knowledge, and dedicated word embedding methodology in the medical context, which are possible to improve the model performance in multi-hop MRC tasks.

Supplementary information. Please visit our website (https://github.com/cluereader/cluereader.github.io) for more visualization samples in HTML files.

Acknowledgments. The authors would like to thank the UCL machine reading group that created the QANGAROO dataset and their help in evaluating our model.

References

[1] Wang, Y., Liu, K., Liu, J., He, W., Lyu, Y., Wu, H., Li, S., Wang, H.: Multi-passage machine reading comprehension with cross-passage answer verification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1918–1927. Association for Computational Linguistics, Melbourne, Australia (2018)

[2] Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M.: Neuroscience-inspired artificial intelligence. Neuron 95(2), 245–258 (2017)

[3] Dehaene, S.: Reading in the Brain: The New Science of How We Read. Penguin Publishing Group, ??? (2009)

[4] Battaglia, P.W., Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V.F., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gülçehre, Ç., Song, H.F., Ballard, A.J., Gilmer, J., Dahl, G.E., Vaswani, A., Allen, K.R., Nash, C., Langston, V., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., Li, Y., Pascanu, R.: Relational inductive biases,
deep learning, and graph networks. CoRR abs/1806.01261 (2018) https://arxiv.org/abs/1806.01261

[5] Seo, M.J., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, ??? (2017)

[6] Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., Hu, G.: Attention-over-attention neural networks for reading comprehension. In: Barzilay, R., Kan, M. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pp. 593–602. Association for Computational Linguistics, ??? (2017)

[7] Welbl, J., Stenetorp, P., Riedel, S.: Constructing datasets for multi-hop reading comprehension across documents. Trans. Assoc. Comput. Linguistics 6, 287–302 (2018)

[8] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, ??? (2019)

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pp. 5998–6008 (2017)

[10] 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. CoRR abs/1907.11692 (2019) https://arxiv.org/abs/1907.11692

[11] Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. CoRR abs/2004.05150 (2020) https://arxiv.org/abs/2004.05150

[12] Razeghi, Y., IV, R.L.L., Gardner, M., Singh, S.: Impact of pretraining term frequencies on few-shot reasoning. CoRR abs/2202.07206 (2022)

[13] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, ??? (2017)
[14] Ding, M., Zhou, C., Chen, Q., Yang, H., Tang, J.: Cognitive graph for multi-hop reading comprehension at scale. In: Korhonen, A., Traum, D.R., Márquez, L. (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp. 2694–2703. Association for Computational Linguistics, ??? (2019). https://doi.org/10.18653/v1/p19-1259

[15] Evans, J.S.B.: Heuristic and analytic processes in reasoning. British Journal of Psychology 75(4), 451–468 (1984)

[16] Sloman, S.A.: The empirical case for two systems of reasoning. Psychological bulletin 119(1), 3 (1996)

[17] Cao, N.D., Aziz, W., Titov, I.: Question answering by reasoning across documents with graph convolutional networks. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 2306–2317. Association for Computational Linguistics, ??? (2019). https://doi.org/10.18653/v1/n19-1240

[18] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Walker, M.A., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pp. 2227–2237. Association for Computational Linguistics, ??? (2018). https://doi.org/10.18653/v1/n18-1202

[19] Cao, Y., Fang, M., Tao, D.: BAG: bi-directional attention entity graph convolutional network for multi-hop reasoning question answering. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 357–362. Association for Computational Linguistics, ??? (2019). https://doi.org/10.18653/v1/n19-1032

[20] Tang, Z., Shen, Y., Ma, X., Xu, W., Yu, J., Lu, W.: Multi-hop reading comprehension across documents with path-based graph convolutional network. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, pp. 3905–3911. ijcai.org, ??? (2020). https://doi.org/10.24963/ijcai.2020/540

[21] Tu, M., Wang, G., Huang, J., Tang, Y., He, X., Zhou, B.: 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, pp. 2704–2713. Association for Computational
Linguistics, Florence, Italy (2019)

[22] Jia, M., Liao, L., Wang, W., Li, F., Chen, Z., Li, J., Huang, H.: Keywords-aware dynamic graph neural network for multi-hop reading comprehension. Neurocomputing 501, 25–40 (2022)

[23] ZHANG, Y., MENG, F., ZHANG, J., CHEN, Y., XU, J., ZHOU, J.: Mkgn: A multi-dimensional knowledge enhanced graph network for multi-hop question answering. IEICE Transactions on Information and Systems E105.D(4), 807–819 (2022)

[24] Song, L., Wang, Z., Yu, M., Zhang, Y., Florian, R., Gildea, D.: Evidence integration for multi-hop reading comprehension with graph neural networks. IEEE Trans. on Knowl. and Data Eng. 34(2), 631–639 (2022)

[25] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9, 1735–80 (1997)

[26] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers. The Association for Computer Linguistics, ??? (2016)

[27] Zhong, V., Xiong, C., Keskar, N.S., Socher, R.: Coarse-grain fine-grain coat-tention network for multi-evidence question answering. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, ??? (2019)

[28] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: International Conference on Learning Representations (2018)

[29] Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E.: Neural message passing for quantum chemistry. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 1263–1272. PMLR, ??? (2017)

[30] Bird, S.: NLTK: the natural language toolkit. In: Calzolari, N., Cardie, C., Isabelle, P. (eds.) ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006. The Association for Computer Linguistics, ??? (2006). https://doi.org/10.3115/1225403.1225421
[31] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A Meeting of SIGDAT, a Special Interest Group of The ACL, pp. 1532–1543. ACL, ??? (2014). https://doi.org/10.3115/v1/d14-1162

[32] Hashimoto, K., Xiong, C., Tsuruoka, Y., Socher, R.: A joint many-task model: Growing a neural network for multiple NLP tasks. In: Palmer, M., Hwa, R., Riedel, S. (eds.) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pp. 1923–1933. Association for Computational Linguistics, ??? (2017). https://doi.org/10.18653/v1/d17-1206

[33] Fey, M., Lenssen, J.E.: Fast graph representation learning with pytorch geometric. CoRR abs/1903.02428 (2019) https://arxiv.org/abs/1903.02428

[34] Schult, D.A.: Exploring network structure, dynamics, and function using networkx. In: In Proceedings of the 7th Python in Science Conference (SciPy, pp. 11–15 (2008)

[35] Song, L., Wang, Z., Yu, M., Zhang, Y., Florian, R., Gildea, D.: Exploring graph-structured passage representation for multi-hop reading comprehension with graph neural networks. CoRR abs/1809.02040 (2018) https://arxiv.org/abs/1809.02040

[36] Dhingra, B., Jin, Q., Yang, Z., Cohen, W.W., Salakhutdinov, R.: Neural models for reasoning over multiple mentions using coreference. In: Walker, M.A., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers), pp. 42–48. Association for Computational Linguistics, ??? (2018). https://doi.org/10.18653/v1/n18-2007