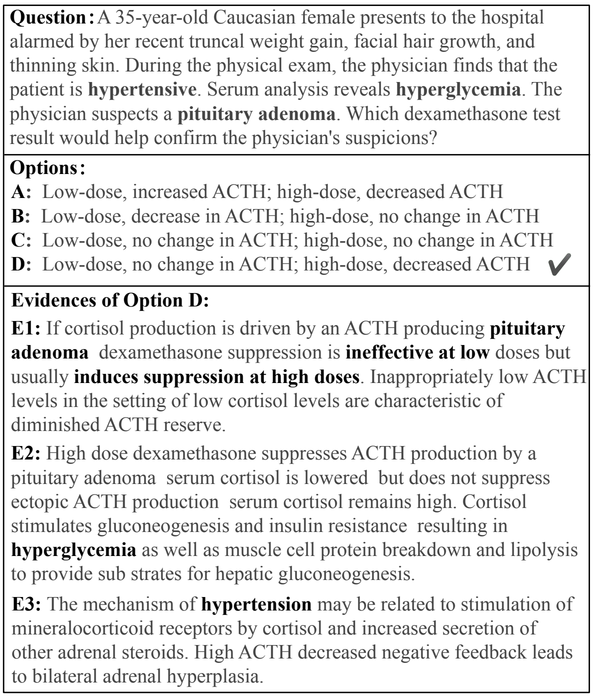
Hierarchical Representation-based Dynamic Reasoning Network for Biomedical Question Answering

Jianguo Mao[1,2†‡] **Jiyuan Zhang**[3†] **Zengfeng Zeng**[3], **Weihua Peng**[3], **Wenbin Jiang**[3*]
Xiangdong Wang[1], **Hong Liu**[1], **Yajuan Lyu**[3]
[1]Beijing Key Laboratory of Mobile Computing and Pervasive Device,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Baidu Inc., Beijing, China
{maojianguo20s,xdwang,hliu}@ict.ac.cn
{zhangjiyuan01,zengzengfeng,jiangwenbin,lvyajuan}@baidu.com

Abstract

Recently, Biomedical Question Answering (BioQA) has attracted growing attention due to its application value and technical challenges. Most existing works treat it as a semantic matching task that predicts answers by computing confidence among questions, options and evidence paragraphs, which is insufficient for scenarios that require complex reasoning based on a deep understanding of biomedical evidences. We propose a novel model termed **H**ierarchical Representation-based **D**ynamic **R**easoning **N**etwork (HDRN) to tackle this problem. It first constructs the hierarchical representations for biomedical evidences to learn semantics within and among evidences. It then performs dynamic reasoning based on the hierarchical representations of evidences to solve complex biomedical problems. Against the existing state-of-the-art model, the proposed model significantly improves more than **4.5%**, **3%** and **1.3%** on three mainstream BioQA datasets, PubMedQA, MedQA-USMLE and NLPEC. The ablation study demonstrates the superiority of each improvement of our model. https://github.com/mikeblueskydl/HDRN



Question: A 35-year-old Caucasian female presents to the hospital alarmed by her recent truncal weight gain, facial hair growth, and thinning skin. During the physical exam, the physician finds that the patient is **hypertensive**. Serum analysis reveals **hyperglycemia**. The physician suspects a **pituitary adenoma**. Which dexamethasone test result would help confirm the physician's suspicions?

Options:
A: Low-dose, increased ACTH; high-dose, decreased ACTH
B: Low-dose, decrease in ACTH; high-dose, no change in ACTH
C: Low-dose, no change in ACTH; high-dose, no change in ACTH
D: Low-dose, no change in ACTH; high-dose, decreased ACTH ✓

Evidences of Option D:

E1: If cortisol production is driven by an ACTH producing **pituitary adenoma** dexamethasone suppression is **ineffective at low** doses but usually **induces suppression at high doses**. Inappropriately low ACTH levels in the setting of low cortisol levels are characteristic of diminished ACTH reserve.

E2: High dose dexamethasone suppresses ACTH production by a pituitary adenoma serum cortisol is lowered but does not suppress ectopic ACTH production serum cortisol remains high. Cortisol stimulates gluconeogenesis and insulin resistance resulting in **hyperglycemia** as well as muscle cell protein breakdown and lipolysis to provide sub strates for hepatic gluconeogenesis.

E3: The mechanism of **hypertension** may be related to stimulation of mineralocorticoid receptors by cortisol and increased secretion of other adrenal steroids. High ACTH decreased negative feedback leads to bilateral adrenal hyperplasia.

Figure 1: **An example from MedQA-USMLE dataset**. (✓: correct answer option).

1 Introduction

Machine reading comprehension (MRC) (Rajpurkar et al., 2016, 2018; Yang et al., 2018) tasks are often used to evaluate the intelligence degree of a system, and many recent large-scale pre-trained language models (Lan et al., 2019; Zaheer et al., 2020) have surpassed the human performance on open-domain MRC. In recent years, Biomedical Question Answering (BioQA) (Tsatsaronis et al., 2012; Wang et al., 2018; Tang et al., 2019; Li et al., 2020; Dai et al., 2022) has attracted growing attention due to its great application value and technical challenges. As Figure 1 shows, compared with open-domain MRC tasks, BioQA raises higher demands for understanding professional biomedical evidences and relies more on complex reasoning based on semantics within and among evidences to predict answers, which is also difficult for humans. The pass rate of the human examinee is less than 14.2% in the National Licensed Pharmacist Examination in China (Li et al., 2020). While in open-domain MRC, the human performance can reach 86.8% Exact Match (Rajpurkar et al., 2018).

Due to the extremely high cost of collection and annotation of the biomedical data, BioASQ (Tsatsaronis et al., 2012) was for a long time the only authoritative benchmark for the development of BioQA systems. Recently, more high-quality

[*]Corresponding author.
[†]Equal contribution.
[‡]This work was done during internship at Baidu Inc.

Proceedings of the 29th International Conference on Computational Linguistics, pages 1480–1489
October 12–17, 2022.

datasets have further contributed to the development of the field, such as NLPEC (Li et al., 2020), MedQA-USMLE (Zhang et al., 2018), and PubMedQA (Jin et al., 2019). Some researchers have explored pre-trained language models to solve this task (Huang et al., 2019; Beltagy et al., 2019a; Lee et al., 2020; Dai et al., 2022). Meanwhile, some researchers have introduced external biomedical knowledge to aid the model in answering questions. (Zhang et al., 2018; Yue et al., 2020). Despite the success, previous works mainly explore a better language model or external domain knowledge, which is insufficient to deal with BioQA in complex scenarios which require complex reasoning based on the semantics within and among evidences to answer the question. Intuitively, it is essential to explore a better representation learning method for biomedical evidences and a better reasoning mechanism for complex biomedical questions.

To tackle this problem, we propose a novel model, termed Hierarchical Representation-based Dynamic Reasoning Network (HDRN), to achieve this goal in two main parts. First, constructing hierarchical representations to learn semantics within and among the biomedical evidences needed to reason the answer. To this end, we first use a shared pre-trained language model to obtain the intra-level representations of the question and evidences separately. Then, we construct coarse to fine-grained inter-level representations of evidences to learn semantics among them. Second, conducting multi-step dynamic reasoning based on the hierarchical representations to predict the answer. At each step, it adaptively aggregates critical information from hierarchical representations according to current state and conducts single-step reasoning to update the state. All intermediate reasoning results are dynamically integrated to predict the answer.

To sum up, the contributions of our work are as follows:

- We propose HDRN, a novel neural network used for semantic representation learning and reasoning for BioQA.
- We design a hierarchical representation learning method to learn semantics within and among the biomedical evidences.
- We design a novel reasoning mechanism that iteratively performs multi-step dynamic reasoning to solve complex biomedical questions.
- We achieve state-of-the-art performances on three BioQA datasets, and the experiment results demonstrate the superiority of each component of the proposed model.

2 Related Work

Biomedical Question Answering BioQA is an emerging and challenging task. Given a question, it requires intelligent systems to understand the complex biomedical domain expertise and reason the answers. Meanwhile, collecting and annotating data requires experts with a medical background to complete, which is difficult and costly. BioASQ (Tsatsaronis et al., 2012) is a benchmark for biomedical semantic indexing and question answering for a long time. Recently, many works have made efforts to construct more high-quality and challenging BioQA datasets. The datasets can be mainly divided into two categories, the first category is constructed based on biomedical domain publications or electronic medical records, including emrQA (Pampari et al., 2018), MedQA-USMLE (Zhang et al., 2018), and PubMedQA (Jin et al., 2019). The second category is constructed based on biomedical examinations from different countries, including Head-QA (Vilares and Gómez-Rodríguez, 2019) and NLPEC (Li et al., 2020). To tackle this task, most of previous works have optimized the language model by pre-training on biomedical domain-related corpus, and obtained great success (Peng et al., 2019; Alsentzer et al., 2019; Jin et al., 2019; Beltagy et al., 2019b; Lee et al., 2019; raj Kanakarajan et al., 2021; Yasunaga et al., 2022). In addition, Yue et al. (2020) explored the external clinical domain knowledge to enhance the generalization of the model. Yasunaga et al. (2021) explored joint reasoning over text and knowledge graph for BioQA. Dai et al. (2022) solved the parameter competition problem via a Mixture-of-Expert.

Unlike previous works, our proposed HDRN model has two distinctive characteristics: (1) It explores a hierarchical representation learning method that can better learn semantics within and among evidences. The effectiveness of the similar idea of hierarchical representation learning method has also been validated in the vision (Lan et al., 2014), recommendation (Jiang et al., 2018) domains, our proposed method is better adapted to the characteristics of professional biomedical evidences. (2) It explores a novel dynamic reasoning mechanism that adaptively aggregates critical information from hierarchical representations for multi-step reasoning and dynamically integrates intermediate reasoning results to predict answers. Although the concept of multi-step reasoning mechanism has been mentioned in other natural language processing tasks (Haug et al., 2018; Liu et al., 2020; Zhao et al., 2021) or domains (Song et al., 2018; Gan et al., 2019; Le et al., 2021), our proposed mechanism has more flexible and powerful information convergence and reasoning capabilities for biomedical question answering.

3 Background

BioQA is a classification task that uses accuracy as the evaluation metric. Specifically, given a natural language question Q and evidences C, it requires the intelligent system to predict the correct option $\hat{o}$ from the candidate set Ω_o based on the understanding of the evidences C. θ is set of the model parameters.

$$\hat{o} = \underset{o \in \Omega_o}{\operatorname{argmax}} P(o|Q, C; \theta) \tag{1}$$

As Figure 2 (a) shows, most previous works treat BioQA as a semantic matching task. They first concatenate all information together, including questions and evidences, then encode them using a language model, and finally predict the best option by computing the semantic matching score with a multi-layer perceptron. The previous works accomplish both representation learning and reasoning using a single model. There are two main drawbacks: (i). Lack of deep understanding of biomedical evidences. The hierarchical information among evidences is easily lost when all the information is mixed together for encoding. (ii). Lack of strong reasoning capability. Single-step implicit reasoning does not cope well with complex questions and evidences. In terms of human experience, the reasoning is usually an iterative process that requires multi-step to solve complex questions.

As Figure 2 (b) shows, to address the problem, we propose HDRN, a network for representing learning and reasoning for BioQA. It first constructs hierarchical representations to obtain a deep understanding of the biomedical evidences, and then performs multi-step dynamic reasoning to solve complex questions.

4 Method

In this section, we describe the detail of the proposed method. The overall architecture is shown in Figure 3, which consists of two components: (a) Dynamic Reasoning Mechanism, which can better solve complex biomedical questions by conducting multi-step dynamic reasoning. (b) Hierarchical Representation Learning, which can better understand the biomedical evidences by learning the semantics within and among the them.

4.1 Hierarchical Representation Learning

Each BioQA instance usually contains multiple evidences related to the question, and the semantic information within and among evidences are essential for reasoning. We propose a hierarchical representation learning method to obtain a deep understanding of the evidences. Specifically, it consists of three levels of representations, **(1) Intra-level Representations**: learning the semantic information of each evidence. **(2) Coarse-grained Inter-level Representations**: learning the correlations among evidences based on token representations of all evidences. **(3) Fine-grained Inter-level Representations**: learning more abstract correlations among evidences based on sentence representations of all evidences.

Intra-level Representations Given a question Q, and the set of M evidence sentences $C = \{c_m\}_{m=1}^{M}$, we use a pre-trained language model(PLM) as the language encoder to extract intra-level representations. E.g., given a sentence S with T tokens, where $S \in \mathbb{R}^{T\times 1}$, we first add a special token $[cls]$ at the beginning as input. After encoding by the language encoder, we obtain the representations $R^S \in \mathbb{R}^{(T+1)\times d^l}$, where d^l is the output feature dimension of language encoder. Specifically, for the set of evidence sentences C, we first concatenate each evidence c_m with question Q, and take the representations of all tokens as their intra-level representations $R^C = \{R^{C_m}\}_{m=1}^{M}$, where $R^{C_m} \in \mathbb{R}^{(T+1)\times d^l}$.

$$R^{C_m} = \text{PLM}([Q; c_m]) \tag{2}$$

where ; is the concatenate operation.

Question Q has only one sentence that does not contain hierarchical information, so we choose the representations of $[cls]$ token as the representations of the question $R^Q \in \mathbb{R}^{1\times d^l}$, which is sufficient to represent the semantics of the whole sentence

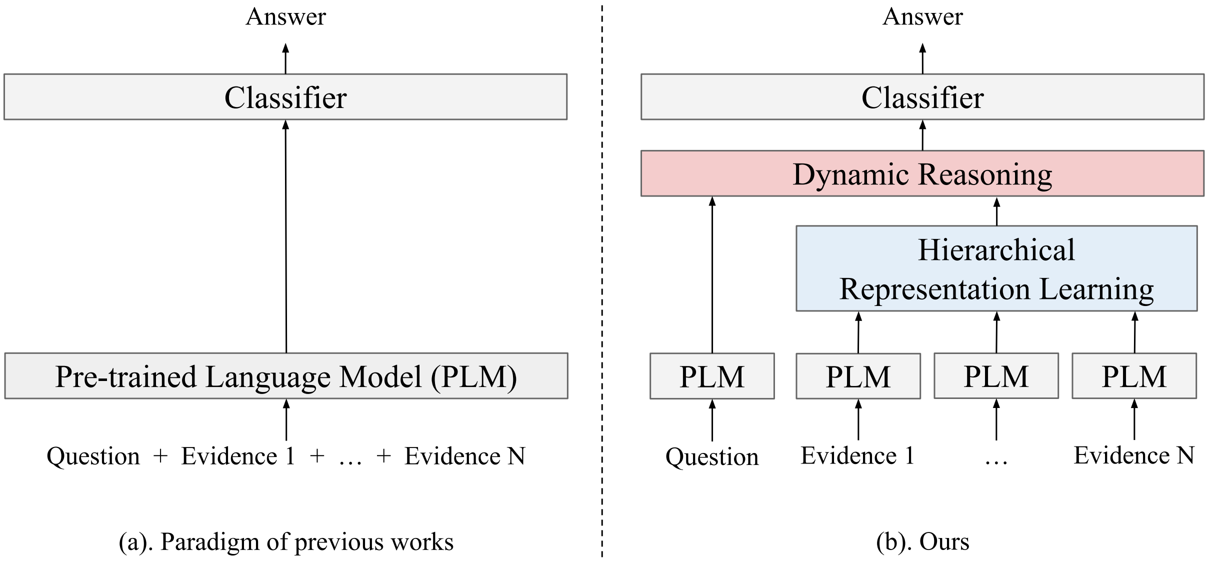

Answer
Classifier
Pre-trained Language Model (PLM)
Question + Evidence 1 + … + Evidence N
(a). Paradigm of previous works
Answer
Classifier
Dynamic Reasoning
Hierarchical Representation Learning
PLM
PLM
PLM
PLM
Question
Evidence 1
…
Evidence N
(b). Ours

Figure 2: **Comparison of the model architecture between our method and previous works**. Unlike previous works that adopt a unified model to conduct representation learning and single-step implicit reasoning, our method constructs hierarchical representations to understand biomedical evidences and conduct multi-step dynamic reasoning to solve complex questions.

and makes the subsequent reasoning process more elegant.

$$R^Q = \mathrm{PLM}_{\mathrm{CLS}}([Q]) \tag{3}$$

Coarse-grained Inter-level Representations We construct the coarse-grained inter-level representations to learn the correlations among evidences paragraphs that are important for reasoning. Specifically, we concatenate the intra-level representations of M evidence paragraphs $R^C = \{R^{C_m}\}_{m=1}^{M}$ into a sequence

$$R^C_{concat} = [R^{C_1}; R^{C_2}; ...; R^{C_M}] \tag{4}$$

where $R^C_{concat} \in \mathbb{R}^{M(T+1)\times d^l}$.

Then, we use Scaled Dot-Product Attention (Vaswani et al., 2017) to update the representations of each token according to the representations of other tokens in all evidence paragraphs to learn the semantic relationships among all evidences.

$$R'^C_{concat} = \mathrm{Attention}(R^C_{concat}, R^C_{concat}, R^C_{concat}) \tag{5}$$

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^\top}{d^l})V \tag{6}$$

Then, we apply a linear projection layer and a residual connection on the updated representations R'^C_{concat}

$$R'^C_{concat} = R^C_{concat} + \mathrm{Linear}(R'^C_{concat}) \tag{7}$$

Finally, we obtain the coarse-grained inter-level representations $R^C_{cInter} = \{R^{C_m}_{cInter}\}_{m=1}^{M}$ by slicing the R'^C_{concat} according to the length of the evidence paragraphs.

Fine-grained Inter-level Representations In order to learn more abstract correlations among evidences, we construct the fine-grained inter-level representations. Specifically, we concatenate the coarse-grained inter-level representations of the $[cls]$ token for each evidence paragraphs into a vector sequence

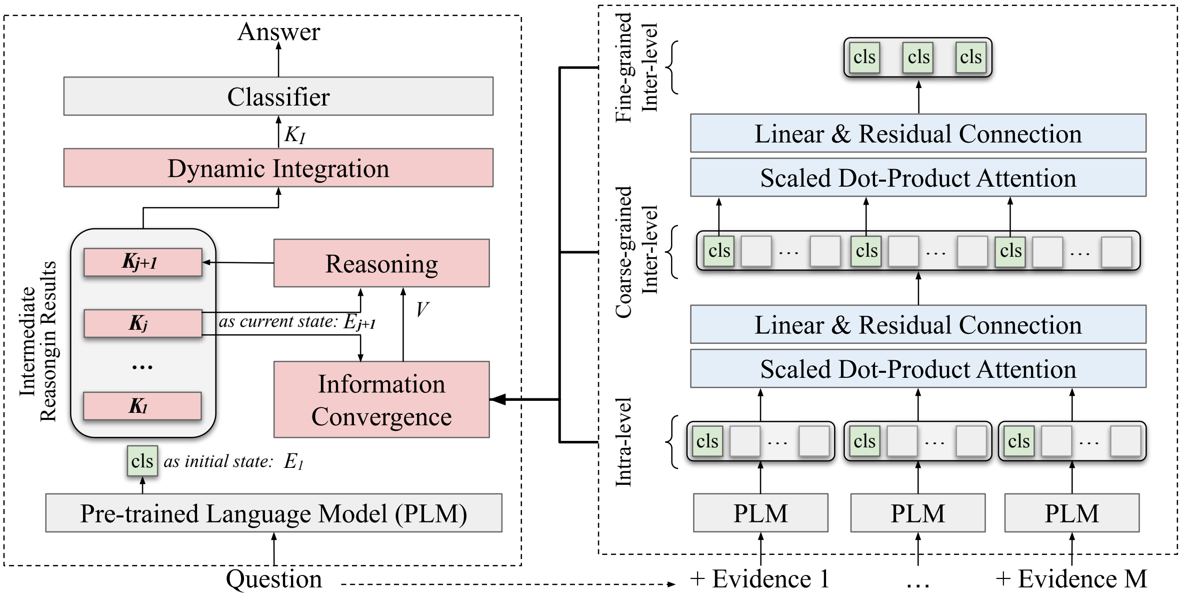
$$R^{C^{cls}}_{cInter} = [R^{C^{cls}_1}_{cInter}; ...; R^{C^{cls}_M}_{cInter}] \tag{8}$$

where $R^{C^{cls}_m}_{cInter} \in \mathbb{R}^{1\times d^l}$.

Then we use the same attention, linear projection layer and residual connection as learning the coarse-grained inter-level representations to obtain the fine-grained inter-level representations $R^C_{fInter} \in \mathbb{R}^{M\times d^l}$

$$R'^{C^{cls}}_{cInter} = \mathrm{Attention}(R^{C^{cls}}_{cInter}, R^{C^{cls}}_{cInter}, R^{C^{cls}}_{cInter}) \tag{9}$$

$$R^C_{fInter} = R^{C^{cls}}_{cInter} + \mathrm{Linear}(R'^{C^{cls}}_{cInter}) \tag{10}$$



(a) Dynamic Reasoning Mechanism

(b) Hierarchical Representation Learning

Figure 3: **Overall architecture of the proposed HDRN**. It contains two components: (a) Dynamic Reasoning Mechanism (described in section 4.2) and (b) Hierarchical Representation Learning (described in section 4.1).

4.2 Dynamic Reasoning Mechanism

In general, reasoning is an iterative process that requires constantly updating the current state according to the state-related information and gradually reasoning out the answer. Inspired by the nature of human reasoning mechanism, we design the Dynamic Reasoning Mechanism to imitate the process. It iteratively performs multi-step dynamic reasoning to predict the answer. At each step of reasoning, it adaptively aggregates hierarchical representations (**Information Convergence**) according to the current state and performs single-step reasoning to obtain the intermediate result and update the state. Each intermediate reasoning result focuses on different parts of the hierarchical information. Thus, we integrate them dynamically to predict the answer better. (**Dynamic Integration**). We set the initial state E_1 to the question representations R^Q at the first reasoning step. When performing the jth step of reasoning, the state is E_j:

Information Convergence We get state-related information $\{H^C, H^C_{cInter}, H^C_{fInter}\}$ from hierarchical representations $\{R^C, R^C_{cInter}, R^C_{fInter}\}$ through Scaled Dot-Product Attention according to the current state E_j.

$$H^C = \text{softmax}(\frac{E_j R^{C^\top}}{d^l})R^C \tag{11}$$

$$H^C_{cInter} = \text{softmax}(\frac{E_j R^{C^\top}_{cInter}}{d^l})R^C_{cInter} \tag{12}$$

$$H^C_{fInter} = \text{softmax}(\frac{E_j R^{C^\top}_{fInter}}{d^l})R^C_{fInter} \tag{13}$$

Then, we apply a linear layer with softmax as the activation function on the state-related information to calculate the distribution of weight D and get the weighted sum as the reasoning-related information V.

$$D = \text{softmax}(\{H^C, H^C_{cInter}, H^C_{fInter}\}) \tag{14}$$

$$V = D \cdot [H^C; H^C_{cInter}; H^C_{fInter}] \tag{15}$$

Reasoning We perform single-step reasoning according to the current state E_j and the reasoning-related information V to obtain the intermediate reasoning results K_j.

$$K_j = E_j + \text{ReLU}(W_1 V + b_1)W_2 + b_2 \tag{16}$$

where W_1 and W_2 are weight matrices and b_1 and b_2 are biases.

Methods	**Accuracy (%)** **Test**
BlueBERT (Peng et al., 2019)	48.4
ClinicalBERT (Alsentzer et al., 2019)	49.0
PubMedBERT (Jin et al., 2019)	55.8
SciBERT (Beltagy et al., 2019b)	57.3
BioBERT (Lee et al., 2019)	60.2
BioELECTRA (raj Kanakarajan et al., 2021)	64.0
UNIFIEDQA-v2 (Khashabi et al., 2022)	64.2
BioLink-BERT (Yasunaga et al., 2022)	72.1
HDRN (Ours)	**76.6**

Table 1: **Performance comparison on the PubMedQA.**

Methods	**Accuracy (%)** **Test**
BERT (Devlin et al., 2018)	34.3
BioRoBERTa (Gururangan et al., 2020)	36.1
BioBERT (Lee et al., 2019)	36.7
PubMedBERT (Jin et al., 2019)	38.1
QAGNN (Yasunaga et al., 2021)	38.0
GreaseLM (Zhang et al., 2022)	38.5
MoE-BQA (Dai et al., 2022)	41.6
BioLink-BERT (Yasunaga et al., 2022)	44.6
HDRN (Ours)	**47.6**

Table 2: **Performance comparison on the MedQA-USMLE.**

At the next step, we set the state E_{j+1} to the K_j. The above Information Convergence and Reasoning process is performed again based on the state E_{j+1}. We obtain J intermediate reasoning results $K = \{K_1, ..., K_J\}$ after repeating the above process J times.

Dynamic Integration After J steps of reasoning, we obtain all intermediate reasoning results $K = \{K_1, ..., K_J\}$. We integrate them by applying a nonlinear transformation with the softmax function.

$$K_I = \text{softmax}(W_I K + b_I)K \tag{17}$$

4.3 Classifier

Given the integrated reasoning results K_I, we use a linear layer as a classifier to obtain the logits l_k for the answer options. Then we calculate the probability distribution of each answer option by applying a softmax function. We use cross-entropy loss as our model loss $\mathcal{L}$ to update the model parameters.

$$l_k = \text{classifier}(K_I) \tag{18}$$

$$\hat{y} = \text{softmax}(l_k), \mathcal{L} = \text{CrossEntropy}(\hat{y}) \tag{19}$$

5 Experiments

5.1 Datasets

As mentioned in Section 2, there are mainly two categories of datasets. In our work, we select three widely used and challenging datasets in two categories to evaluate the model performance. Among them, the MedQA-USMLE and PubMedQA are in english, and the NLPEC is in chinese. We use accuracy to measure the model performance.

Methods	**Accuracy (%)** **Test**
BiDAF (Seo et al., 2016)	43.6
Co-Matching (Wang et al., 2018)	45.8
SeaReader (Zhang et al., 2018)	48.4
Multi-Matching (Tang et al., 2019)	48.7
BERT-base (Devlin et al., 2018)	52.2
ERNIE (Sun et al., 2019)	53.4
RoBERTa-wwm-ext-large (Cui et al., 2021)	57.9
KMQA (Li et al., 2020)	61.8
MoE-BQA (Dai et al., 2022)	62.2
HDRN (Ours)	**63.5**

Table 3: **Performance comparison on the NLPEC.**

PubMedQA PubMedQA is a large scale English BioQA dataset collected from PubMed abstracts. It contains a total of 273.5k QA examples, of which 1k expert-annotated, 211.3k artificially generated, and 61.2k unlabeled. The number of examples for the train/dev/test set is 272,950/50/500. Each instance consists of a question, a context which is the abstract from PubMed without its conclusion, and a long answer which is the conclusion of the context. It requires answering the question with yes/no/maybe based on the reasoning over context.

MedQA-USMLE MedQA-USMLE is a large scale multilingual BioQA dataset collected from the National Medical Board Examinations in the USA, Mainland China, and Taiwan. Most previous work only used English subset for training and evaluation, so we also use English subset for fair comparison. The English subset contains 12k QA examples in total. The number of examples for the train/dev/test set is 10,178/1,272/1,273. Each example consists of a question, four candidate options with the correct one annotated. It requires predicting the correct option corresponding to the given question.

NLPEC NLPEC is a large scale Chinese BioQA dataset containing 21.7k multiple-choice questions with human-annotated answers collected from the National Licensed Pharmacist Examination in China. The number of examples for the train/dev/test set is 18,703/2,500/547. Each question has five candidate options and evidences retrieved from the official exam guidebook that contains the information needed to answer the question. It requires predicting the correct option corresponding to the given question.

5.2 Implementation Details

For all three datasets, we use the official dataset splits to train and test our model. We set the number of evidences to 3. The feature dimension of language encoder is set to 1024. We conduct our experiments on NVIDIA A100 GPUs with 40GB memroy.

PubMedQA We use BioLink-BERT's (Yasunaga et al., 2022) parameters as initialization parameters for the language model. We use 450 annotated examples and 10k randomly selected artificially generated QA examples for model training. Our model does not use long answers, which is more challenging. The number of reasoning steps is set to 3. We set the batch size to 32, and use AdamW with β_1=0.9 and β_2=0.999 as the optimizer. We set the learning rate to 3e-5. The maximum number of epochs is set to 23.

MedQA-USMLE We use BioLink-BERT's (Yasunaga et al., 2022) parameters as initialization parameters for the language model. We use BM25 to retrieve six sentences with highest scores for each option from official guided books provided by the datasets as evidences. The number of reasoning steps is set to 2. We set the batch size to 32, and use AdamW with β_1=0.9 and β_2=0.98 as the optimizer. We set the learning rate to 3e-5. The maximum number of epochs is set to 6.

NLPEC We use RoBERTa-wwm-ext-large's (Cui et al., 2021) parameters as initialization parameters for the language model. The number of reasoning steps is set to 3. We set the batch size to 16, and use AdamW with β_1=0.9 and β_2=0.999 as the optimizer. We set the learning rate to 3e-5. The maximum number of epochs is set to 35.

5.3 Comparison with State-of-the-Arts

As shown in Table 1, 2 and 3, the proposed method achieves the new state-of-the-art and reaches 76.6% / 47.6%/ 63.5% accuracy on PubMedQA / MedQA-USMLE/ NLPEC datasets with 4.5% / 3.0%/ 1.3% improvement over the previous state-of-the-art method.

PubMedQA Table 1 shows the results on PubMedQA dataset. The first to eighth lines show the accuracy of the previous state-of-the-art methods on the test set. These methods follow a semantic matching paradigm to solve this task and achieve competitive results. They first use language models to encode questions, options, and evidences, and then perform single-step implicit reasoning to predict the answer. Most of the previous works optimize the performance of the pre-trained language models on BioQA by pre-training with biomedical domain-related corpus. These pre-trained language models can all benefit from the better semantic representation and reasoning capabilities of our method. Specifically, our method gains 4.5% improvement compared with the baseline BioLink-BERT (Yasunaga et al., 2022) on test set.

MedQA-USMLE Table 2 shows the results on MedQA-USMLE dataset. The first to eighth lines shows the accuracy of the previous state-of-the-art methods on the test set. These models also follow the semantic matching paradigm to solve this task. Furthermore, by introducing external knowledge and conducting joint reasoning over text and graph (Yasunaga et al., 2021), the performance is further improved. Our method gains 3.0% improvement on test set compared with the baseline BioLink-BERT.

NLPEC Table 3 shows the results on NLPEC dataset. The first to ninth lines show the accuracy of the previous state-of-the-art methods on the test set. These models also follow the semantic matching paradigm to solve this task. Furthermore, by retrieving external biomedical knowledge (Li et al., 2020), the semantic matching capability of the model can be enhanced. By introducing a mixture of experts (Dai et al., 2022) to alleviate the parameter competition problem, where each expert handles a specific type of question, providing better single-step reasoning capabilities. Our method benefits from the proposed better semantic representation learning method and more powerful dynamic reasoning mechanisms. Specifically, our method gains 5.6% improvement compared with the base-

Models		Test Accuracy (%)	
	PubMedQA	MedQA	NLPEC
HDRN (Ours)	**76.6**	**47.6**	**63.5**
w/o Hierarchical Representation Learning	75.2 (1.4 ↓)	46.9 (0.7 ↓)	62.4 (1.1 ↓)
w/o Dynamic Reasoning Mechanism	75.0 (1.6 ↓)	46.8 (0.8 ↓)	62.3 (1.2 ↓)
Hierarchical Representation Learning			
w/o Fine-grained Inter-level	75.6 (1.0 ↓)	46.7 (0.9 ↓)	63.3 (0.2 ↓)
w/o Coarse-grained Inter-level	75.6 (1.0 ↓)	46.9 (0.7 ↓)	62.6 (0.9 ↓)
w/o Intra-level	75.8 (0.8 ↓)	46.4 (1.2 ↓)	62.8 (0.7 ↓)
Dynamic Reasoning Mechanism			
w/o Information Convergence	75.7 (0.9 ↓)	46.9 (0.7 ↓)	62.9 (0.6 ↓)
w/o Dynamic Integration	75.6 (1.0 ↓)	46.8 (0.8 ↓)	62.5 (1.0 ↓)

Table 4: **Ablation study on three BioQA datasets.**

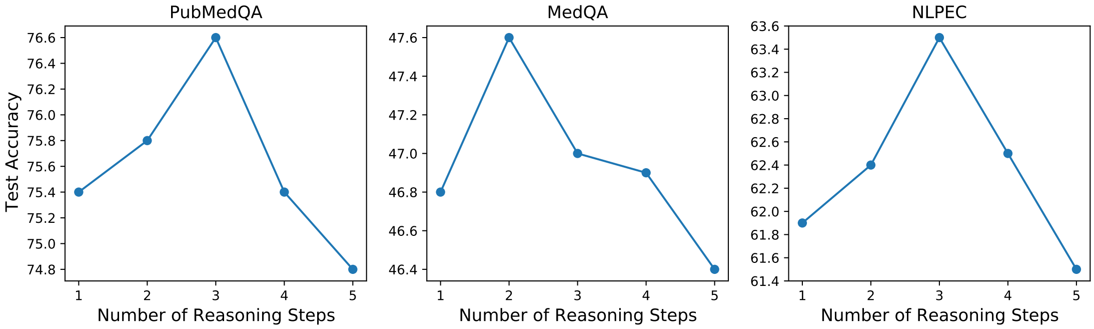


Figure 4: **Effect of the Number of Reasoning Steps.** The optimal number of reasoning steps for the PubMedQA, MedQA-USMLE, and NLPEC are 3, 2, and 3 respectively.

line RoBERTa-wwm-ext-large (Cui et al., 2021) on test set and gains 1.3% improvement compared with the latest work Moe-BQA (Dai et al., 2022).

5.4 Ablation Study

Table 4 shows the results of the ablation study on three BioQA datasets, which demonstrate the superiority of each component. From the experimental results, if there is no Dynamic Reasoning Mechanism, the model performs single-step reasoning based on the output of the language model as in most existing works, and the model performance decreases. If there is no Hierarchical Representation Learning, all information is concatenated together for representation learning, which is the same as the previous works shown in Figure 2 (a), and the model performance further decreases. In addition, we conduct further ablation experiments to analyze the effectiveness and superiority of two key improvements. For Dynamic Reasoning, if there is no Information Convergence, the model cannot dynamically select hierarchical representations for reasoning according to the current state, and the model performance decreases, if there is no Dynamic Integration, the model only uses the result of the last step of reasoning to predict the answer, losing the key information in the reasoning process, and the performance decreases. For Hierarchical Representations, we remove different levels of representations separately to explore the effectiveness of each level representation, and the experimental results show that removing different levels of representation degrades the performance to different degrees.

Effect of the Number of Reasoning Steps Figure 4 shows the effect of the number of reasoning steps. When the number of steps is 1, the process is the same as the classical single-step implicit reasoning paradigm. Empirically, the number of reasoning steps is related to the problem complexity, and the performance gradually increases as we gradually increase the number of reasoning steps. However, when the number of reasoning steps is too large, the performance degrades due to the mismatch between the reasoning process and the problem complexity. The optimal number of reasoning steps varies slightly for different data distributions. The optimal number of reasoning steps for the PubMedQA, MedQA-USMLE, and NLPEC are 3, 2, and 3 respectively.

6 Conclusion

This paper proposes HDRN, a novel model for representation learning and reasoning for biomedical question answering. First, we construct hierarchical representations to obtain a deep understanding of the biomedical evidences. Then, we perform multi-step dynamic reasoning to solve complex biomedical questions. We evaluate our model on three BioQA datasets and achieve new state-of-the-art performances.

Acknowledgements

Our work is supported by the National Key Research and Development Program of China No.2020AAA0109400 and National Natural Science Foundation of China (62276250). This work is also supported by Baidu and ICTCAS Joint Project. We would like to thank the anonymous reviewers for their valuable feedback.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019a. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019b. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Damai Dai, Wenbin Jiang, Jiyuan Zhang, Weihua Peng, Yajuan Lyu, Zhifang Sui, Baobao Chang, and Yong Zhu. 2022. Mixture of experts for biomedical question answering. *arXiv preprint arXiv:2204.07469*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhe Gan, Yu Cheng, Ahmed El Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. *arXiv preprint arXiv:1902.00579*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Till Haug, Octavian-Eugen Ganea, and Paulina Grnarova. 2018. Neural multi-step reasoning for question answering on semi-structured tables. In *European conference on information retrieval*, pages 611–617. Springer.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Zhuoren Jiang, Yue Yin, Liangcai Gao, Yao Lu, and Xiaozhong Liu. 2018. Cross-language citation recommendation via hierarchical representation learning on heterogeneous graph. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 635–644.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*.

Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. 2014. A hierarchical representation for future action prediction. In *European Conference on Computer Vision*, pages 689–704. Springer.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Hung Le, Chinnadhurai Sankar, Seungwhan Moon, Ahmad Beirami, Alborz Geramifard, and Satwik Kottur. 2021. Dvd: A diagnostic dataset for multi-step reasoning in video grounded dialogue. *arXiv preprint arXiv:2101.00151*.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Dongfang Li, Baotian Hu, Qingcai Chen, Weihua Peng, and Anqi Wang. 2020. Towards medical machine reading comprehension with structural knowledge and plain text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1427–1438.

Ye Liu, Shaika Chowdhury, Chenwei Zhang, Cornelia Caragea, and Philip S Yu. 2020. Interpretable multi-step reasoning with knowledge extraction on complex healthcare question answering. *arXiv preprint arXiv:2008.02434*.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. Bioelectra: pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Xiaomeng Song, Yucheng Shi, Xin Chen, and Yahong Han. 2018. Explore multi-step reasoning in video question answering. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 239–247.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Min Tang, Jiaran Cai, and Hankz Hankui Zhuo. 2019. Multi-matching network for multiple choice reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7088–7095.

George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. 2012. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. Citeseer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

David Vilares and Carlos Gómez-Rodríguez. 2019. Head-qa: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966.

Shuohang Wang, Mo Yu, Shiyu Chang, and Jing Jiang. 2018. A co-matching model for multi-choice reading comprehension. *arXiv preprint arXiv:1806.04068*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*.

Xiang Yue, Bernal Jiménez Gutiérrez, and Huan Sun. 2020. Clinical reading comprehension: A thorough analysis of the emrqa dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4474–4486.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models for question answering. *arXiv preprint arXiv:2201.08860*.

Chen Zhao, Chenyan Xiong, Jordan Boyd-Graber, and Hal Daumé III. 2021. Multi-step reasoning over unstructured text with beam dense retrieval. *arXiv preprint arXiv:2104.05883*.