Unifying Structure Reasoning and Language Pre-training for Complex Reasoning Tasks

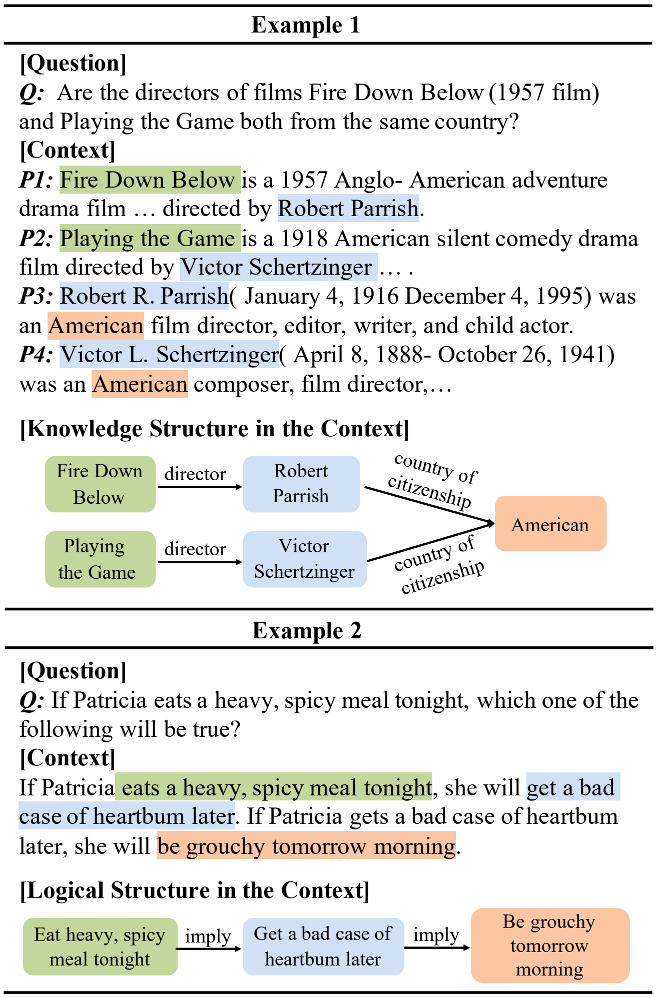
Siyuan Wang, Zhongyu Wei, Jiarong Xu, Taishan Li, Zhihao Fan

***Abstract*—Recent pre-trained language models (PLMs) equipped with foundation reasoning skills have shown remarkable performance on downstream complex tasks. However, the significant structure reasoning skill has been rarely studied, which involves modeling implicit structure information within the text and performing explicit logical reasoning over them to deduce the conclusion. This paper proposes a unified learning framework that combines explicit structure reasoning and language pre-training to endow PLMs with the structure reasoning skill. It first identifies several elementary structures within contexts to construct structured queries and performs step-by-step reasoning along the queries to identify the answer entity. The fusion of textual semantics and structure reasoning is achieved by using contextual representations learned by PLMs to initialize the representation space of structures, and performing stepwise reasoning on this semantic representation space. Experimental results on four datasets demonstrate that the proposed model achieves significant improvements in complex reasoning tasks involving diverse structures, and shows transferability to downstream tasks with limited training data and effectiveness for complex reasoning of KGs modality.**

***Index Terms*—Structure reasoning skill, language model pre-training, complex reasoning.**

I. Introduction

RECENT years have witnessed an ever-growing research on complex reasoning, which requires comprehending the given information and applying complex rules to draw inferences [1, 2, 3, 4]. As a defining property of advanced intelligence, it inspires immense potential for numerous real-world applications, such as fact checking [5, 6], math word problem solving [7, 8], natural language navigation [9, 10] and medical diagnosis [11, 12]. Existing large-scale pre-trained language models (PLMs) have shown superior performance on various downstream tasks, exhibiting the general linguistic reasoning skill for understanding contextual information learned from broad data [13, 14, 15, 16]. However, complex reasoning tasks are far more challenging and diverse, involving further foundation reasoning skills. For example, the numerical reasoning skill to abstract quantitative information from text [19, 20], and the spatial reasoning skill to perceive spatial relations between objects are two essential abilities [21, 22, 23]. Equipping PLMs with these foundation skills enables them to be adapted to a wide variety of downstream reasoning tasks, which significantly steps forward in artificial general intelligence.



Example 1

[Question]

Q: Are the directors of films Fire Down Below (1957 film) and Playing the Game both from the same country?

[Context]

P1: Fire Down Below is a 1957 Anglo- American adventure drama film … directed by Robert Parrish.

P2: Playing the Game is a 1918 American silent comedy drama film directed by Victor Schertzinger … .

P3: Robert R. Parrish(January 4, 1916 December 4, 1995) was an American film director, editor, writer, and child actor.

P4: Victor L. Schertzinger(April 8, 1888- October 26, 1941) was an American composer, film director,…

[Knowledge Structure in the Context]

Example 2

[Question]

Q: If Patricia eats a heavy, spicy meal tonight, which one of the following will be true?

[Context]

If Patricia eats a heavy, spicy meal tonight, she will get a bad case of heartbum later. If Patricia gets a bad case of heartbum later, she will be grouchy tomorrow morning.

[Logical Structure in the Context]

Eat heavy, spicy meal tonight
imply
Get a bad case of heartbum later
imply
Be grouchy tomorrow morning

Fig. 1: A multi-hop reasoning example (Example 1) from [17] and a logical reasoning example (Example 2) from [18]. The contexts embody implicit structures for reasoning. The shaded rectangles are entities and the texts on top of arrows are relations in the knowledge and logical structures, and the shaded phrases in contexts are the mentions of corresponding entities.

In addition to these foundation reasoning skills, there is also an essential ability that is rarely explored, namely the structure reasoning skill. Structure reasoning aims to model

Manuscript received March 30, 2023. (Corresponding author: Zhongyu Wei).

Siyuan Wang, Zhongyu Wei and Zhihao Fan are with the School of Data Science, Fudan University, Shanghai 200433, China. And Zhongyu Wei is also with the Research Institute of Intelligent and Complex Systems, Fudan University, Shanghai 200433, China (email: wangsy18@fudan.edu.cn; zywei@fudan.edu.cn; fanzh18@fudan.edu.cn).

Jiarong Xu is with the School of Management, Fudan University, Shanghai 200433, China (jiarongxu@fudan.edu.cn).

Taishan Li is with the School of Information Science and Technology, ShanghaiTech University, Shanghai, 201210, China (litsh@shanghaitech.edu.cn).

implicit structure information within the text and perform explicit logical reasoning over them to deduce the conclusion. In Figure 1, we show two complex reasoning examples [18, 24, 25, 26] that involve the structure reasoning skill. We can see that contexts embody rich information for answering these questions. Although the relevant information typically exists in complex and sparse forms (shaded phrases) in context, they are implicitly organized as a structure. Specifically, the contexts of Example 1 (a multi-hop reasoning example) and Example 2 (a logical reasoning example) respectively embody an intrinsic knowledge structure and logical structure. Through explicit reasoning over these structures, the correct answers can be inferred. We argue that formulating relevant information from unstructured text as topological structures and performing logical reasoning over them can help solve extensive complex questions. In this paper, we thus explore endowing the PLMs with the foundation structure reasoning skill for complex reasoning.

Previous research usually generates large-scale question-answer pairs that require reasoning and continue training PLMs over the synthetic data to inject foundation reasoning skills [27, 28, 29]. Such data-driven methods attempt to learn foundation skills in an implicit manner, making it uncertain whether they really gain these capabilities or just exploit data biases as shortcuts for question answering. Instead, we aim to model the structure reasoning skill more explicitly and take inspiration from geometric embedding methods for step-by-step reasoning over structures (knowledge graphs) [30, 31, 32].

To this end, we propose a unified learning framework to combine explicit structure reasoning and language pre-training. For structure reasoning, we first define several elementary structures within contexts and perform reasoning over the structure for the answer entity identification. Specifically, our model first utilizes these context-inherent basic structures to construct structured queries with corresponding answer entities, and encode them as geometric shapes in a representation space. Then the reasoning along the query structure is explicitly conducted by iteratively executing logical operations over the representation space. The goal is to push the answer representation to be close enough to the final query representation. For the fusion of textual semantics and structure reasoning, we use contextual representations learned by PLMs from contexts to initialize the representation space of structures, and perform stepwise reasoning on this semantic representation space. In this way, the pre-trained model is taught with the explicit structure reasoning capability over text and can easily generalize to different structures composed of these studied elementary ones.

We conduct experiments on two datasets for multi-hop reasoning, HotpotQA [24] and 2WikiMultiHopQA [17], as well as two logical reasoning datasets, ReClor [18] and LogiQA [26]. The results demonstrate that our model unifying the explicit structure reasoning skill into PLMs achieves significant improvements in complex reasoning tasks involving diverse structures. Further analysis shows its transfer ability to downstream tasks with limited training data and effectiveness for complex reasoning of KGs modality. The contributions of this work can be summarized as follows:

- We propose a new foundation reasoning skill, namely structure reasoning skill, to formulate implicit structure information from the text and perform logical reasoning over them, which is significant for complex reasoning.
- We propose a unified learning framework to explicitly inject the structure reasoning skill into PLMs for better generalizing to different complex reasoning tasks with only unstructured text but involving structure reasoning. It is a generic framework that can be plugged into different pre-trained language models.
- We present extensive experiments, demonstrating the effectiveness and low-resource transfer ability of our proposed framework.

II. Related Work

A. Foundation Reasoning Skills Learning

Large-scale pre-trained language models [14, 15, 33, 34] facilitate a variety of downstream NLP applications but have difficulty in challenging and diverse complex reasoning tasks. To generally improve complex reasoning, a set of foundation reasoning skills is introduced for capturing some common reasoning abilities across different tasks and are incorporated into PLMs. For example, [19, 20] propose to inject numerical reasoning skills for complex tasks requiring understanding and reasoning over quantitative information. Commonsense reasoning [35] and logical inference abilities [36] are also taught to pre-trained language models for solving problems involving commonsense and informal logic. Here, we focus on learning another significant structure reasoning skill, which aims to formulate the implicit structure from text and perform explicit reasoning over them. Even recent large language models, such as GPT-3 [34] and ChatGPT [1], are yet to master explicit structural reasoning. These models continue to have difficulties in complex reasoning tasks, such as multi-hop reasoning, and are prone to generating factual errors [37, 38]. This highlights the crucial need to develop structural reasoning.

Existing work primarily exploits pre-defined templates to synthesize substantial amounts of data requiring different reasoning skills, which are in the form of question-answer pairs or mask-out statements. They endow PLMs with these foundation reasoning skills by further refining PLMs on augmented training data. Different from this implicit teaching process, we inject the structure reasoning skills into PLMs more explicitly by stepwise performing logical operations along structures over semantic representation space.

B. Knowledge Enhanced Pre-trained Language Models

Another line of related work is on incorporating structured knowledge from external sources into PLMs to provide extra clues for reasoning [35, 39, 40, 41]. The main challenge is that the infused knowledge and the text usually have heterogeneous representation space. To integrate heterogeneous information, some works need to revise the architecture of PLMs with an additional fusion module [42, 43]. Others require hand-crafted heuristics to combine knowledge and text into a unified

[1] https://chat.openai.com/

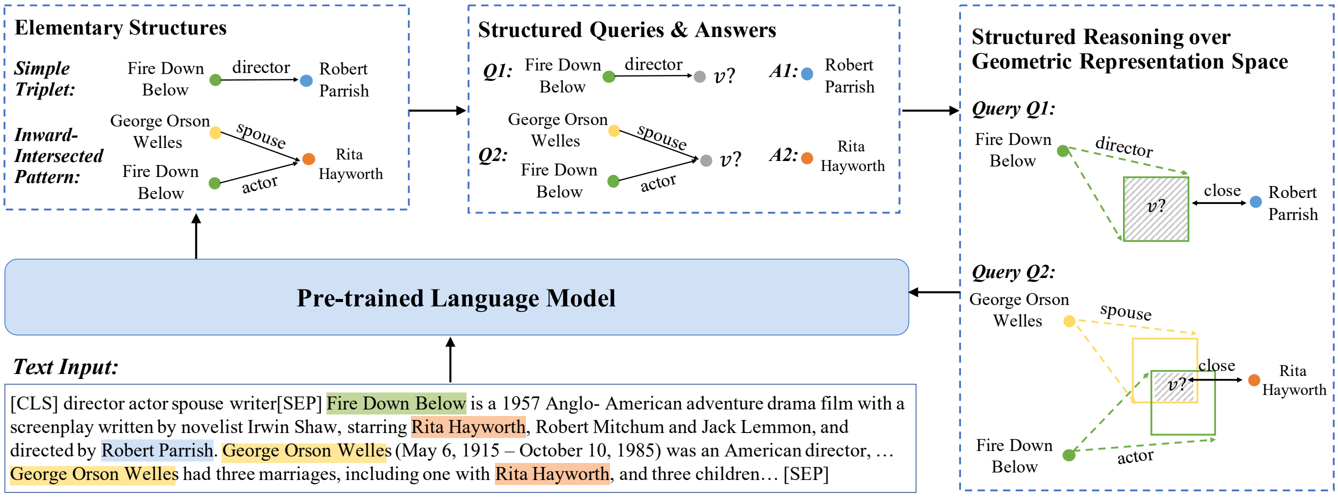

Elementary Structures
Simple Triplet:
Fire Down Below
director
Robert Parrish
Inward-Intersected Pattern:
George Orson Welles
spouse
Fire Down Below
actor
Rita Hayworth
Structured Queries & Answers
Q1:
Fire Down Below
director
v?
A1:
Robert Parrish
Q2:
George Orson Welles
spouse
Fire Down Below
actor
v?
A2:
Rita Hayworth
Structured Reasoning over Geometric Representation Space
Query Q1:
Fire Down Below
director
v?
close
Robert Parrish
Query Q2:
George Orson Welles
spouse
v?
close
Rita Hayworth
Fire Down Below
actor
Pre-trained Language Model
Text Input:
[CLS] director actor spouse writer[SEP] Fire Down Below is a 1957 Anglo- American adventure drama film with a screenplay written by novelist Irwin Shaw, starring Rita Hayworth, Robert Mitchum and Jack Lemmon, and directed by Robert Parrish. George Orson Welles (May 6, 1915 – October 10, 1985) was an American director, … George Orson Welles had three marriages, including one with Rita Hayworth, and three children… [SEP]

Fig. 2: The overall architecture of our framework unifying structure reasoning and language modeling.

data structure, such as a sequence [44], a sentence tree [45] and a word-knowledge graph[46], and convert them into input sequences for encoding. However, these methods ignore the structural information of knowledge graphs in some cases [47]. Besides, knowledge retrieval is required to identify relevant information from external sources and the inaccurate retrieval procedure introduces noisy knowledge inevitably.

Contrary to these works incorporating external structured information, we study modeling the intrinsic structures in the text and reasoning over them. It is essential for complex reasoning tasks [18, 24, 25] where the question needs to be answered grounded on the context with complex but sparse information, which actually constitutes an implicit structure. Although some attempts can be viewed as structure modeling in text, they only consider reasoning over simple triplets [48, 49, 50, 51]. Moreover, they mainly adopt span-based masked pre-training or transform the triplets to text sequences for contrastive learning, which ignores the structured information and explicit structure reasoning over them. The crucial difference is that we propose to unify language modeling and explicit structure reasoning on both simple and complex structures within text.

III. Methodology

In this work, we propose to unify explicit structure reasoning and language pre-training in one framework for complex reasoning (see Figure 2). Given a piece of text sequence s with existing entities $\mathcal{E} = \{e_1, e_2, e_3, ...\}$ and their relations $\mathcal{R} = \{..., r_{i,j}, ...\}$ constituting triplets $\mathcal{T} = \{(e_i, r_{i,j}, e_j) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}\}$, our goal is to learn the contextual representations of s together with reasoning over complex structure information within the text. Complex structures vary drastically in different texts, which usually exhibit in the form of multi-step paths and intersected triplets, or a combination of them. Therefore, we aim to learn general reasoning skills over elementary structures to achieve generalizability to any structures. We define four types of elementary structures that can constitute almost arbitrary complex structures by free combination. It includes the simple triplet and three complex structures, i.e., the two-step path, outward-intersected pattern, and inward-intersected pattern.

In this section, we first introduce how to obtain the representations of entities and relations from text, which are the basic units of our defined structures (§ III-A). We then detailedly describe these elementary structures and their identification process from the text (§ III-B). We follow the geometric embedding-based method for explicit structure reasoning over the semantic representation space(§ III-C). Finally, we show the pre-training process to learn a structure-aware language representation for acquiring structure reasoning skills and its fine-tuning stage on downstream complex reasoning tasks (§ III-D).

A. Basic Structure Unit Representations

Taking the text sequence s with tokens $\{s_1, s_2, ..., s_{|s|}\}$ as input, the pre-trained language model outputs a sequence of contextual representations $H = \{h_1, h_2, ..., h_{|s|}\}$. We do not maintain separate embeddings for structures. Instead, we obtain representations of basic structural units from the contextual language representations, to fuse textual and structured semantics and improve the structure reasoning skill during language pre-training. In detail, for an entity e_i in the text s, we take the average hidden state of its start token $s^{e_i}_{start}$ and end token $s^{e_i}_{end}$ as its representation $\mathbf{e_i}$. If the entity occurs in the text sequence multiple times N_{e_i}, we further take the average of its multiple occurrences as Eq. 1. As the relation may not be explicitly or consecutively mentioned in the text, we additionally concatenate the relations of all existing triplets with the text sequence s as the input `[CLS]` $r_{i,j}$, `...` `[SEP]` s `[SEP]` for encoding. For each relation $r_{i,j}$ we also get the average hidden state of its start token $s^{r_{i,j}}_{start}$ and end token $s^{r_{i,j}}_{end}$ as its representation $\mathbf{r_{i,j}}$.

$$\mathbf{e_i} = \frac{1}{N_{e_i}} \sum_{1}^{N_{e_i}} \frac{1}{2}(h^{e_i}_{start} + h^{e_i}_{end}) \tag{1}$$

$$\mathbf{r_{i,j}} = \frac{1}{2}(h^{r_{i,j}}_{start} + h^{r_{i,j}}_{end}) \tag{2}$$

B. Elementary Structures Identification

After getting the basic unit representations, we propose four types of elementary structures for reasoning over both simple and complex structures in the text as shown in Figure 3. A simple triplet is comprised of two entities and one relation. Complex structures can be multi-step paths, intersected triplets, or a combination of them. We consider three elementary complex structures, including the two-step path, outward-intersected pattern and inward-intersected pattern. We aim to learn general reasoning skills over these four types of elementary structures to achieve generalizability to model any complex structures in text, even those never seen during pre-training. For example, the complex knowledge structure of the first example in Figure 1 can be modeled by combining several two-step paths and an inward-intersected pattern.

Type	**Structure**
Simple Triplet	$e_1 \xrightarrow{r_{1,2}} e_2$
Two-step Path	$e_1 \xrightarrow{r_{1,2}} e_2 \xrightarrow{r_{2,3}} e_3$
Outward-intersected Pattern	$e_1 \xrightarrow{r_{1,2}} e_2$ $e_1 \xrightarrow{r_{1,3}} e_3$
Inward-intersected Pattern	$e_2 \xrightarrow{r_{2,1}} e_1$ $e_3 \xrightarrow{r_{3,1}} e_1$

Fig. 3: The definition of different elementary structures.

Taking simple triplets extracted from the context as input, we further identify the other three complex structures from the text. For two triplets $(e_1, r_{1,2}, e_2)$ and $(e_2, r_{2,3}, e_3)$ appear in the same context, if the tail e_2 of the first triplet is the head of the other, we treat it as a *two-step path*. If two triplets $(e_1, r_{1,2}, e_2)$ and $(e_1, r_{1,3}, e_3)$ in the same context share the same head e_1 but their tails are different, we combine them as a *outward-intersected pattern*. If two triplets $(e_2, r_{2,1}, e_1)$ and $(e_3, r_{3,1}, e_1)$ have the same tail e_1 but differ in heads, we recognize them as a *inward-intersected pattern*.

C. Structure Reasoning Injection

Based on these identified structures, we can construct corresponding structural queries and perform explicit structure reasoning along them to reach answers. Inspired by [31, 32], we utilize geometric embedding learning methods for stepwise updating structural query representation and encourage the query representation to be close to the answer representation in the vector space. We use contextual language representations of structures from PLMs to initialize their geometric embeddings to integrate textual and structured semantics.

a) Structural Query Construction: We construct structural queries on top of pre-defined elementary structures by taking different entities from different structures as target answers (query node) and viewing the remaining as queries. Specifically, the tail entity is extracted as the target in the simple triplet, while the intersected entities are treated as answers for intersected patterns. For the two-step path, we remove the intermediate entity and take the last entities as the query target. The detailed strategies for structured query construction with their answer entities are shown in Figure 4.

Type	**Query**	**Answer**
Simple Triplet	$e_1 \xrightarrow{r_{1,2}} v?$	e_2
Two-step Path	$e_1 \xrightarrow{r_{1,2}} \ \xrightarrow{r_{2,3}} v?$	e_3
Outward-intersected Pattern	$v? \xrightarrow{r_{1,2}} e_2$ $v? \xrightarrow{r_{1,3}} e_3$	e_1
Inward-intersected Pattern	$e_2 \xrightarrow{r_{2,1}} v?$ $e_3 \xrightarrow{r_{3,1}} v?$	e_1

Fig. 4: Structured query construction for different structures and $v?$ is the query node.

b) Geometric Representation Learning: As a structure reasoning method for answering complex queries over knowledge graphs, it models a set of entities into a geometric region in the vector space. Correspondingly, it learns the logical operators in queries including relation projection and intersection as operations over geometric regions, which will result in new regions. The structure reasoning can be conducted from the start entities by explicitly executing the logical operators along the structural query to update the query representation, and encourage the answer entities to be inside or close to the final region of the query.

Specifically, we use boxes (hyper-rectangles) as the geometric regions to represent the structural queries. A box embedding $\mathbf{b}$ is composed of a center vector Cen($\mathbf{b}$) and an offset vector Off($\mathbf{b}$) as $\mathbf{b} = (\text{Cen}(\mathbf{b}), \text{Off}(\mathbf{b}))$, and model a set of entities in the following box:

$$\{\mathbf{e} : \text{Cen}(\mathbf{b}) - \text{Off}(\mathbf{b}) \preceq \mathbf{e} \preceq \text{Cen}(\mathbf{b}) + \text{Off}(\mathbf{b})\} \tag{3}$$

where Cen($\mathbf{b}$) and Off($\mathbf{b}$) are respectively the center and positive offset of the box. Each entity e can then be represented as a zero-offset box $\mathbf{e} = (\text{Cen}(\mathbf{e}), \mathbf{0})$ with only one element as the center. The relation r is also embedded as $\mathbf{r} = (\text{Cen}(\mathbf{r}), \text{Off}(\mathbf{r}))$ and the relation projection operator is modeled as $\mathbf{b} + \mathbf{r}$ to respectively sum the centers and offsets, and obtain a transformed box embedding. The intersection operator over multiple box embeddings $\{\mathbf{b_1}, \mathbf{b_2}, ...\mathbf{b_n}\}$ is modeled as $\mathbf{b}_\cap = (\text{Cen}(\mathbf{b}_\cap), \text{Off}(\mathbf{b}_\cap))$, where Cen($\mathbf{b}_\cap$) and Off($\mathbf{b}_\cap$) are computed as Eq. 4 and 5. a_i is the attention weight over the box center $\mathbf{b_i}$ and DeepSets($\cdot$) is the permutation-invariant function over sets [52].

$$\text{Cen}(\mathbf{b}_\cap) = \sum_{i}^{n} a_i \odot \text{Cen}(\mathbf{b_i}) \tag{4}$$

$$\begin{aligned}\text{Off}(\mathbf{b}_\cap) = \text{Min}(\{\text{Off}(\mathbf{b_i}), i \in 1, ..., n\}) \\ \odot\, \sigma(\text{DeepSets}(\{\mathbf{b_1}, i \in 1, ..., n\}))\end{aligned} \tag{5}$$

c) Structure Reasoning: According to the constructed structural queries, we then perform structure reasoning using the geometric representation learning method and make the answer embeddings to be as similar as possible to the query box embedding. We take the contextual semantic representations of entities and relations in Eq. 1 and 2 as their center vectors Cen($\mathbf{e}$) and Cen($\mathbf{r}$). We do not learn an adaptive offset for different relations as in [31] but train a shared one to reduce the extra learning burden and fuse the structure reasoning into language pre-training more smoothly. The loss of structure reasoning for a query-answer pair using negative sampling is calculated as follows:

$$\begin{aligned}\mathcal{L}_{QA} = &-\log\sigma(\gamma - d(\mathbf{a},\mathbf{q})) \\ &-\sum_{k=1}^{K}\frac{1}{K}\log\sigma(d(\mathbf{a}_{\mathbf{k}}^{'},\mathbf{q}) - \gamma)\end{aligned} \tag{6}$$

where $\mathbf{q}$ is the derived box embedding of the query q and $\mathbf{a}$ is the answer entity embedding of a. $\mathbf{a}_{\mathbf{k}}^{'}$ is the embedding of a negative answer span $a_k^{'}$ randomly sampled from the same text. γ is the fixed margin and the function $d(\cdot,\cdot)$ measures the distance between an entity and a box.

The detailed distance calculation between an entity embedding $\mathbf{e}$ and a query box embedding $\mathbf{b} = (\text{Cen}(\mathbf{b}), \text{Off}(\mathbf{b}))$ consist of two parts [31] as following:

$$d(\mathbf{e},\mathbf{b}) = d_{out}(\mathbf{e},\mathbf{b}) + \alpha\ d_{in}(\mathbf{e},\mathbf{b}) \tag{7}$$

where $d_{out}(\mathbf{e},\mathbf{b})$ and $d_{in}(\mathbf{e},\mathbf{b})$ are respectively the outer distance between the entity and the closest box corner and the inner distance between the box center and one box corner. α is a scalar coefficient balancing these two distances.

$$\begin{aligned}d_{out}(\mathbf{e},\mathbf{b}) =& || \max(\mathbf{e} - \mathbf{b}_{\max}, \mathbf{0}) + \\ & \max(\mathbf{b}_{\min} - \mathbf{e}, \mathbf{0})|| &(8)\\ d_{in}(\mathbf{e},\mathbf{b}) =& ||\text{Cen}(\mathbf{b}) - \\ & \min(\mathbf{b}_{\max} - \max(\mathbf{b}_{\min}, \mathbf{e}))|| &(9)\\ \mathbf{b}_{\max} =& \text{Cen}(\mathbf{b}) + \text{Off}(\mathbf{b}) &(10)\\ \mathbf{b}_{\min} =& \text{Cen}(\mathbf{b}) - \text{Off}(\mathbf{b}) &(11)\end{aligned}$$

D. Pre-training & Fine-tuning

In order to inject the structure reasoning skill into PLMs, we jointly optimize structure reasoning and language modeling. We follow the basic-level masking strategy of RoBERTa [14] for masked language modeling. Although the basic-level MLM might mask entities, the positions of all entities within the text remain known. The contextual representations for these positions, even when masked, still denote corresponding entities, since MLM's objective is to utilize these contextual representations to recover the masked tokens. Therefore, the fact that MLM might mask entities doesn't affect our implementation in III-A to obtain basic structure unit representations from text.

The overall pre-training objective is the combination of the structure reasoning loss $\mathcal{L}_{SR}$ and the masked language modeling loss $\mathcal{L}_{MLM}$ as Eq. 12.

$$\mathcal{L} = \mathcal{L}_{SR} + \mathcal{L}_{MLM} \tag{12}$$

For each text, we do not model all existing elementary structures. We have attempted to model different numbers of complex structures which all show similar performance on downstream tasks. However, modeling more complex structures simultaneously will increase the computational burden and slow down the optimization speed. Therefore, we randomly choose a simple triplet and one of the other three complex structures to construct both a simple query and a complex query for structure reasoning over text. The $\mathcal{L}_{SR}$ is computed as follows:

$$\mathcal{L}_{SR} = \lambda_1\mathcal{L}_{QA}^{\text{simple}} + \lambda_2\mathcal{L}_{QA}^{\text{complex}} \tag{13}$$

where $\mathcal{L}_{QA}^{\text{simple}}$ and $\mathcal{L}_{QA}^{\text{complex}}$ are respectively the structure reasoning loss over simple and complex structures queries. λ_1 and λ_2 are weighted hyper-parameters to balance the losses.

With the structure reasoning skill injected during language pre-training, our model learns structure-aware language representations and can be directly fine-tuned on downstream language tasks requiring complex structure reasoning.

IV. Experiments

In the experiments, we evaluate our model on the textual complex reasoning tasks, and further analyze the contribution of each structural component and the transfer ability to downstream tasks with limited training data. Moreover, the capabilities of our model in complex reasoning over KGs and knowledge probing are also validated.

A. Experimental Setup

a) Pre-training phase: We utilize Wikipedia documents with entities and their relations constructed by [50] to obtain structures for structure reasoning enhanced language pre-training. Specifically, it utilizes spaCy to perform named entity recognition. It then aligns entity mentions to Wikidata items for annotating relations. According to their human evaluation, the F1 scores of entity and relation extraction are respectively 84.7% and 25.4%. We concatenate all documents and splits them into sequences of the same length for masked language modeling so that each sequence may cover multiple documents. Then we can utilize triplets from different documents involved in one sequence to construct cross-document structures and queries involving complex reasoning.

Our primary goal is to demonstrate the general effectiveness of endowing PLMs with structure reasoning ability across different tasks and datasets, and fairly compare its performance with other (knowledge) structure-injected methods. Therefore, we employ the widely adopted RoBERTa-base [14] as the backbone of our language model. We set the learning rate as 5e-5. The hyper-parameters λ_1 and λ_2 that balance structure reasoning losses over different structures are set as 1 and 0.1, respectively. The fixed margin γ is set as 24. The coefficient to balance outer and inner distances between an entity and a query box is set as $\alpha = 0.02$.

b) Fine-tuning Period: We evaluate our model on different tasks requiring reasoning over text with implicit complex structures, including multi-hop reasoning on HotpotQA [24] (the distractor setting) and 2WikiMultiHopQA [17] datasets, and logical reasoning on ReClor [18] and LogiQA [26] datasets. For datasets without publicly available test sets, we randomly split the development set into two half-sections for validation and testing respectively. We evaluate our model on the validation set of each dataset to choose parameters for testing. All models are implemented using Huggingface [53]. For both HotpotQA and 2WikiMultiHopQA, the batch size is set to 96 while for ReClor and LogiQA, we use a batch size of 24. The Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ is taken as the optimizer and we use a linear learning rate scheduler with 10% warmup proportion. The proposed systems and other comparison models are trained on NVIDIA Tesla V100 GPUs.

B. Overall Performance

a) Comparison Models: We compare our framework with the baseline model and several other structure-injected models, which can be divided into four categories: (1) *RoBERTa* is the baseline model without any structure modeling. (2) *CoLAKE* [46] and *KEPLER* [54]: external knowledge structure enhanced language models. *CoLAKE* aggregates surrounding triplets and converts them into input sequences for entity masked pre-training. *KEPLER* incorporates unnecessarily relevant triplets and learns a unified structure-textual representation. (3) *ERICA* [50] models the structure information in text including entity and relations through contrastive textual discrimination to improve language modeling. (4) *Ours* is our proposed model which learns reasoning skills over intrinsic complex structures. We also compare a variant *Ours(PTransE)* utilizing the path-based TransE method [55] for structure reasoning which embeds the query as single point instead of the box in vector space and only models simple triplets and multiple-step paths.

Since our model is the first to introduce the structure reasoning skills to capture the intrinsic complex structures within the text, the most comparable work is *ERICA* which models simpler in-text relational structures. Another line of related work is on modeling structures of external knowledge into PLMs. To integrate heterogeneous information from text and external knowledge, some works revise the architecture of PLMs with an additional fusion module, which makes their direct application to downstream QA datasets challenging. Thus We focus on those models still following RoBERTa architecture and only modify the input formatting or training paradigms, and select CoLAKE and KEPLER for a fair comparison. Given that all these comparison models adopt the RoBERTa architecture, we directly load our pre-trained checkpoints and fine-tune them on the QA datasets in the same manner as RoBERTa.

Additionally, we also compare our framework with ChatGPT by randomly sampling 200 test instances from each dataset. As with all other models, we feed the context and question to ChatGPT for each instance and prompt it to generate the corresponding answer. For multi-hop reasoning datasets in an extractive question-answering format, we utilize the following prompt to obtain an answer span.

System: *You are a helpful assistant.*
User:
Given: *Meghan Elizabeth Trainor (born December 22, 1993) is an American singer-songwriter. Interested in music from a young age, she wrote, recorded, and produced three independently released albums between ...*
Question: *Which award the performer of song 'Watch Me Do' got?*
Your answer should be extracted directly from the Given context.
Answer:

For logical reasoning datasets in multiple-choice question-answering format, we also provide four options and use the following prompt to get one of them as the answer.

System: *You are a helpful assistant.*
User:
Given: *The computer anti-virus company calls the viruses that have been captured and processed as known viruses, otherwise it is an unknown virus...*
Question: *Which of the following, if true, can weaken the above argument to the greatest extent?*
A. *Viruses that are truly innovative and ...*
B. *99% of new viruses are written after imitating ...*
C. *Computer viruses are written by humans. They are ...*
D. *Every time an anti-virus company claims to ...*
Your answer should begin with the letter corresponding to your choice (i.e. A, B, C and D).
Answer:

b) Multi-hop Reasoning: Multi-hop reasoning aims to aggregate multiple pieces of documents and formulate cross-document structures to answer a complex question. We adopt two textual multi-hop reasoning datasets, HotpotQA and 2WikiMultiHopQA, that involve various reasoning steps (2$\sim$4) and structures. HotpotQA and 2WikiMultiHopQA respectively consist of 90,447 / 7,405 / 7,405 and 167,454 / 3,702 / 3,703 samples in training, development and test sets. They both need to identify an answer span to the question from the context and predict the supporting facts to explain the reasoning. We concatenate the question and context and take them as input for fine-tuning. The models are trained on HotpotQA for 10 epochs with the learning rate 7e-5, and on 2WikiMultiHopQA for 5 epochs with the learning rate 3e-5.

The experimental results are shown in Table I. We have the following findings.

- Our model outperforms almost all other models on both HotpotQA and 2WikiMultiHopQA datasets. Even when compared to the powerful ChatGPT, our model achieves comparable performance (Noting that ChatGPT can not predict the answer phrase exactly). This demonstrates the necessity of explicit structure reasoning over text during language modeling.
- Compared to *CoLAKE* and *KEPLER* utilizing external knowledge structures to enhance language modeling, our models and *ERICA* achieve a considerable improvement.

TABLE I: Experimental results of different pre-trained models on HotpotQA and 2WikiMultiHopQA datasets. Results marked with * indicate the percentage of sampled questions that can be correctly answered by ChatGPT.

Model	Answer		Supporting Fact		Joint	
	EM	F1	EM	F1	EM	F1
			HotpotQA			
RoBERTa [14]	64.82	78.68	61.18	86.61	42.66	70.12
CoLAKE [46]	63.31	77.93	61.56	86.73	42.41	69.63
KEPLER [54]	63.20	77.29	61.74	86.82	42.41	69.07
ERICA [50]	64.34	78.52	61.50	86.64	43.63	70.13
Ours(PTransE)	**65.25**	79.07	**62.23**	87.05	43.41	70.72
Ours	64.85	**79.14**	62.07	**87.06**	**43.82**	**70.87**
ChatGPT	69.00*	-	-	-	-	-
			2WikiMultiHopQA			
RoBERTa [14]	62.29	66.25	80.30	90.68	55.20	62.56
CoLAKE [46]	62.50	66.80	80.01	90.53	55.25	62.99
KEPLER [54]	62.25	66.83	79.87	90.50	54.83	63.00
ERICA [50]	66.16	71.22	79.82	90.66	57.76	67.14
Ours(PTransE)	66.28	71.14	80.30	90.52	58.57	67.13
Ours	**66.30**	71.11	**80.41**	**90.70**	**59.00**	**67.32**
ChatGPT	50.50*	-	-	-	-	-

This suggests that instead of introducing external knowledge, it is more important to directly model the intrinsic structures in text for complex multi-hop reasoning.

- To illustrate the utility of box embedding method, we compare *Ours* with *Ours(PTransE)*. The results show that *Ours* performs better than *Ours(PTransE)* in the joint performance of answer prediction and supporting fact prediction, which shows that the box embedding method is more effective than path-based TransE for explicit complex structure reasoning.

c) Logical Reasoning: Logical reasoning aims to uncover the logical structures within the text and perform inference over them to deduce the answer. We evaluate two benchmarks covering diverse logical structures, ReClor and LogiQA, which are collected from standardized exams including GMAT and LSAT and National Civil Servants Examination of China respectively. Reclor and LogiQA respectively contain 4,638 / 250 / 250 and 7,376 / 651 / 651 data points for training, validation and testing. More specifically, we conduct multiple-choice question answering on ReClor and LogiQA by taking a context, a question and four options as the input and outputting the most plausible option as the answer. We concatenate the context, the question and each option as an input sequence for encoding, resulting in four formulated sequences, and then choose the one with the highest score. The maximum sequence length is set as 256, the number of training epochs is 10 and the learning rate is 1e-5.

The results of logical reasoning are presented in Table II. Our model performs the best across almost all models, including powerful ChatGPT. This verifies that explicitly incorporating structure reasoning over text during pre-training indeed helps improve logical reasoning performance. On the contrary, *CoLAKE* and *KEPLER* that incorporate external knowledge structures into language pre-training would damage the logical reasoning performance of language models in some cases.

TABLE II: Evaluation accuracies (%) of different pre-trained models on both validation and test sets of ReClor and LogiQA.

Model	ReClor		LogiQA	
	Val.	Test	Val.	Test
RoBERTa	52.8	52.8	34.56	32.26
CoLAKE	51.6	50.4	34.72	31.18
KEPLER	46.8	44.0	31.18	27.96
ERICA	55.6	55.2	31.18	26.27
Ours(PTransE)	52.8	56.4	31.34	32.87
Ours	**56.8**	**57.6**	**35.18**	**33.03**
ChatGPT	-	56.0	-	34.00

C. Ablation Study

a) Different Elementary Structures: To demonstrate the impact of different elementary structures on our model, we conduct an ablation study by learning structure reasoning respectively over the simple triplet, two-step path, inward-intersected pattern and outward-intersected pattern into *RoBERTa*. Table III presents the joint EM and joint F1 on both validation and test sets on the 2WikiMultiHopQA dataset. The results demonstrate that the simple triplet and other three complex structures can improve multi-hop reasoning performance. Moreover, our model combining these four structures performs the best, which indicates that these four types of elementary structures are complementary for complex reasoning.

b) Different Pre-training Objectives: We also provide an additional study to show how much improvement are respectively from structure reasoning loss L_{SR} and masked language modeling loss L_{MLM} on 2WikiMultiHopQA dataset. As shown in Table IV, our model simultaneously optimizing structure reasoning and masked language modeling can achieve best performance.

TABLE III: Ablation study on 2WikiMultiHopQA. EM and F1 respectively denote joint EM and joint F1 scores. *outward.pattern* and *inward.pattern* mean the outward-intersected pattern and inward-intersected pattern.

Model	Val.		Test	
	EM	F1	EM	F1
RoBERTa	55.53	62.92	55.20	62.56
w/ simple triplet	58.76	67.49	58.64	66.90
w/ two-step path	57.51	65.90	56.95	65.33
w/ outward.pattern	55.90	63.42	55.87	63.08
w/ inward.pattern	57.47	66.87	57.16	66.40
Ours	59.08	67.75	59.00	67.32

TABLE IV: Ablation study of various pre-training losses on 2WikiMultiHopQA.

Model	Val.		Test	
	EM	F1	EM	F1
RoBERTa	55.53	62.92	55.20	62.56
w/ L_{MLM}	57.70	66.52	57.25	65.90
w/ L_{SR}	58.43	67.53	57.52	66.74
Ours	59.08	67.75	59.00	67.32

D. Further Analysis

a) Transfer Ability under Low-Resource Setting: It is challenging to transfer a pre-trained model to complex reasoning tasks in low-resource settings, where only limited training data are available. To illustrate the generalization capability of our model to the complex reasoning tasks under low-resource settings, we conduct experiments on 2WikiMultiHopQA with limited 10% training data (see Table V). Compared to training on the whole training set, we find that all models show a significant performance drop when only 10% training data is available. This suggests that fine-tuning with limited training data is nontrivial. Furthermore, we see that our model still outperforms the others by a considerable margin, exhibiting a better low-resource transfer ability. This is because the foundation structure reasoning skill learned by our model can help generalize to new reasoning tasks requiring structure reasoning over text.

TABLE V: Results on the test set of 2WikiMultiHopQA with respectively 10% and 100% training data.

Model	10%		100%	
	Joint EM	Joint F1	Joint EM	Joint F1
RoBERTa	39.28	49.27	55.20	62.56
CoLAKE	37.69	47.77	55.25	62.99
KEPLER	39.84	49.49	54.83	63.00
ERICA	40.67	50.27	57.76	67.14
Ours	41.03	51.21	59.00	67.32

b) Complex KG Reasoning: We further conduct experiments to verify the effectiveness of our model with explicit structure reasoning skills for complex reasoning on KGs. The complex KG reasoning task aims to answer complex structural queries (first-order logic queries) on incomplete knowledge graphs with one or a set of entities. We adopt the FB15k dataset [56] derived from Freebase for evaluation. We follow the setting of [31] to generate 5 types of complex structural queries (i.e., 1p, 2p, 3p, 2i, 3i) for training and 9 structural query types (i.e., 1p, 2p, 3p, 2i, 3i, ip, pi, 2u, up) for evaluation, so that we can evaluate queries that are both seen and unseen during training time. For a structural query, we perform explicit structure reasoning on it via step-by-step updating the geometric embedding of the query and deduce the final query box embedding. The entities whose embeddings are close enough to the query embedding will be predicted as the final answers. We use different pre-trained models to initialize the embeddings of entities and relations as described in Sec § III-C and compare their reasoning performance. We fine-tune for 100,000 steps with the learning rate 1e-4.

We report H@3 results on different structural queries of FB15k in Table VII. We find that box embedding method (*BoxE*) with our pre-trained model achieves better performance on average. Our model help improves the performance on the structured queries (1p, 2p, 2i) that we have modeled during pre-training and other complex ones composed of them (3p, 3i, ip, pi). Even on the structural queries that are unseen during the training period (ip, pi, 2u, up), our model can achieve better or comparable performance. These observations demonstrate that our model with explicit structure reasoning is more beneficial to initialize geometric embeddings for complex KG reasoning tasks, and generalize across different knowledge structures.

c) Knowledge Probing: To illustrate that explicit structure reasoning integrated language model can learn the structured knowledge, we conduct knowledge probing experiments on LAMA and LAMA-UHN probes. The task requires the pre-trained model to directly predict masked spans in factual descriptions without fine-tuning. The experimental results are shown in Table VI. We do not compare ERICA as it can not probe any of this factual knowledge. We can see that although our model performs slightly worse than *CoLAKE* which incorporates external surrounding knowledge in PLMs, it outperforms *RoBERTa*, *KEPLER* and *ERICA*. This suggests that the structure reasoning over text can also help learn a certain amount of knowledge.

TABLE VI: Knowledge probing results (P@1) on LAMA and LAMA-UHN. * indicates that the results are from [46] and [54]. Others are from our implementation.

Model	LAMA		LAMA-UHN	
	Google-RE	T-REx	Google-RE	T-REx
*RoBERTa**	5.3	24.7	2.2	17.0
*CoLAKE**	9.5	28.8	4.9	20.4
*KEPLER**	7.3	24.6	3.3	16.5
RoBERTa	4.7	18.9	1.8	14.5
Ours	8.3	28.8	3.1	19.2

TABLE VII: H@3 results of box embedding (BoxE) method using different pre-trained models for knowledge embedding initialization on different structured queries of FB15k dataset. ‘p’, ‘i’, and ‘u’ respectively represent ‘relation projection’, ‘triplet intersection’, and ‘triplet union’.

Model	Avg	1p	2p	3p	2i	3i	ip	pi	2u	up
BoxE	0.497	0.797	0.421	0.313	0.606	0.721	0.221	0.429	0.628	0.338
BoxE(RoBERTa)	0.512	0.812	0.436	0.320	0.628	0.740	0.232	0.441	0.652	0.346
BoxE(CoLAKE)	0.512	0.812	0.437	0.319	0.624	0.742	0.232	0.444	0.652	0.346
BoxE(KEPLER)	0.500	0.799	0.425	0.318	0.614	0.727	0.222	0.424	0.627	0.341
BoxE(ERICA)	0.511	0.812	0.435	**0.321**	0.624	0.741	0.230	0.442	0.648	0.344
BoxE(Ours(PTransE))	0.512	0.812	0.437	0.318	0.628	0.742	0.231	0.444	0.653	0.343
BoxE(Ours)	**0.514**	**0.813**	**0.439**	0.320	**0.629**	**0.745**	**0.233**	**0.444**	**0.653**	**0.346**

TABLE VIII: A Case of the geometric property change in semantic representations before and after implementing our structured reasoning framework.

Head ($\mathbf{h}$)	Relation ($\mathbf{r}$)	Tail ($\mathbf{t}$)	Distance(*RoBERTa*)	Distance(*Ours*)
Fire Down Below	director	Robert Parrish	1.0732	0.9995
Fire Down Below	actor	Rita Hayworth	1.2129	0.9995
George Orson Welles	spouse	Rita Hayworth	1.1797	1.0498
Fire Down Below	writer	Irwin Shaw	1.1406	0.9976

E. Case Study

In Table VIII, we compare the geometric property of semantic representations before and after our structure reasoning training using the case from Figure 2 to illustrate its interpretable effectiveness. We measure the distance between each structural query representation and its corresponding answer entity representation, contrasting results from the baseline model (*RoBERTa*) and our model (*Ours*). Since the baseline model does not learn an offset vector, we do not calculate the box embedding distance as in Eq. 7. For a fair comparison, we follow TransE [56] to calculate the squared Euclidean distance between the query $\mathbf{h}+\mathbf{r}$ and answer $\mathbf{t}$ for simple triplets as follows:

$$d(\mathbf{h}+\mathbf{r},\mathbf{t}) = ||\mathbf{h}||_2^2 + ||\mathbf{r}||_2^2 + ||\mathbf{t}||_2^2 - 2(\mathbf{h}^T\mathbf{t} + \mathbf{r}^T(\mathbf{t}-\mathbf{h})) \tag{14}$$

where $\mathbf{h}$, $\mathbf{t}$, and $\mathbf{r}$ are respectively head entity, tail entity, and relation of the triplet.

In our representation space, answer entities align more closely with structural queries than they do in RoBERTa's representation space. This demonstrates that incorporating structural reasoning during language pre-training can endow semantic representations with inherent geometric structures.

V. Conclusion

In this paper, we propose to inject the foundation structure reasoning skill into PLMs for complex reasoning tasks to model implicit structures within contexts and perform explicit reasoning over them. To accomplish this objective, we present a unified framework combining structure reasoning and language modeling. It extracts four types of elementary structures from contexts to construct structured queries and adopts the stepwise embedding method for explicit structure reasoning along the constructed queries to find the answer. We utilize contextual language representations to initialize structure representations for fusing textual and structure semantics. Experimental results show the general effectiveness of our model on complex reasoning tasks.

In the future, we will take different complexity of knowledge structures into consideration and model a more comprehensive skill over more structures. Besides, more foundation reasoning skills can be explored, such as abductive reasoning and external knowledge retrieval, and be combined with the structure reasoning skill for more general reasoning tasks.

References

[1] N. B. Songer, B. Kelcey, and A. W. Gotwals, “How and when does complex reasoning occur? empirically driven development of a learning progression focused on complex reasoning about biodiversity,” *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, vol. 46, no. 6, pp. 610–631, 2009.

[2] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick, “Neuroscience-inspired artificial intelligence,” *Neuron*, vol. 95, no. 2, pp. 245–258, 2017.

[3] S. Wang, Z. Liu, W. Zhong, M. Zhou, Z. Wei, Z. Chen, and N. Duan, “From lsat: The progress and challenges of complex reasoning,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2201–2216, 2022.

[4] K. Czechowski, T. Odrzygóźdź, M. Zbysiński, M. Zawalski, K. Olejnik, Y. Wu, Ł. Kuciński, and P. Miłoś, “Subgoal search for complex reasoning tasks,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 624–638, 2021.

[5] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "Fever: a large-scale dataset for fact extraction and verification," *arXiv preprint arXiv:1803.05355*, 2018.

[6] W. Ostrowski, A. Arora, P. Atanasova, and I. Augenstein, "Multi-hop fact checking of political claims," *arXiv preprint arXiv:2009.06401*, 2020.

[7] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner, "Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs," *arXiv preprint arXiv:1903.00161*, 2019.

[8] A. Amini, S. Gabriel, P. Lin, R. Koncel-Kedziorski, Y. Choi, and H. Hajishirzi, "Mathqa: Towards interpretable math word problem solving with operation-based formalisms," *arXiv preprint arXiv:1905.13319*, 2019.

[9] H. Chen, A. Suhr, D. Misra, N. Snavely, and Y. Artzi, "Touchdown: Natural language navigation and spatial reasoning in visual street environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 538–12 547.

[10] H. Kim, A. Zala, G. Burri, H. Tan, and M. Bansal, "Arramon: A joint navigation-assembly instruction interpretation task in dynamic environments," *arXiv preprint arXiv:2011.07660*, 2020.

[11] S. Datta and K. Roberts, "A hybrid deep learning approach for spatial trigger extraction from radiology reports," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2020. NIH Public Access, 2020, p. 50.

[12] B. van Aken, J.-M. Papaioannou, M. Mayrdorfer, K. Budde, F. A. Gers, and A. Loeser, "Clinical outcome prediction from admission notes using self-supervised knowledge integration," *arXiv preprint arXiv:2102.04110*, 2021.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[15] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *Proc. of ICLR*, 2019.

[16] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," *arXiv preprint arXiv:2006.03654*, 2020.

[17] X. Ho, A.-K. Duong Nguyen, S. Sugawara, and A. Aizawa, "Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6609–6625. [Online]. Available: https://aclanthology.org/2020.coling-main.580

[18] W. Yu, Z. Jiang, Y. Dong, and J. Feng, "Reclor: A reading comprehension dataset requiring logical reasoning," *arXiv preprint arXiv:2002.04326*, 2020.

[19] M. Geva, A. Gupta, and J. Berant, "Injecting numerical reasoning skills into language models," *arXiv preprint arXiv:2004.04487*, 2020.

[20] D. Petrak, N. S. Moosavi, and I. Gurevych, "Improving the numerical reasoning skills of pretrained language models," *arXiv preprint arXiv:2205.06733*, 2022.

[21] K. C. Moen, M. R. Beck, S. M. Saltzmann, T. M. Cowan, L. M. Burleigh, L. G. Butler, J. Ramanujam, A. S. Cohen, and S. G. Greening, "Strengthening spatial reasoning: Elucidating the attentional and neural mechanisms associated with mental rotation skill development," *Cognitive Research: Principles and Implications*, vol. 5, no. 1, pp. 1–23, 2020.

[22] R. Mirzaee, H. R. Faghihi, Q. Ning, and P. Kordjmashidi, "Spartqa:: A textual question answering benchmark for spatial reasoning," *arXiv preprint arXiv:2104.05832*, 2021.

[23] S. Subramanian, W. Merrill, T. Darrell, M. Gardner, S. Singh, and A. Rohrbach, "Reclip: A strong zero-shot baseline for referring expression comprehension," *arXiv preprint arXiv:2204.05991*, 2022.

[24] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," in *Proc. of EMNLP*, 2018.

[25] J. Welbl, P. Stenetorp, and S. Riedel, "Constructing datasets for multi-hop reading comprehension across documents," *TACL*, 2018.

[26] J. Liu, L. Cui, H. Liu, D. Huang, Y. Wang, and Y. Zhang, "Logiqa: A challenge dataset for machine reading comprehension with logical reasoning," *arXiv preprint arXiv:2007.08124*, 2020.

[27] S. Li, J. Chen, and D. Yu, "Teaching pretrained models with commonsense reasoning: A preliminary kb-based approach," *arXiv preprint arXiv:1909.09743*, 2019.

[28] O. Yoran, A. Talmor, and J. Berant, "Turning tables: Generating examples from semi-structured tables for endowing language models with reasoning skills," *arXiv preprint arXiv:2107.07261*, 2021.

[29] F. Zhang, D. Tang, Y. Dai, C. Zhou, S. Wu, and S. Shi, "Skillnet-nlu: A sparsely activated model for general-purpose natural language understanding," *arXiv e-prints*, pp. arXiv–2203, 2022.

[30] W. Hamilton, P. Bajaj, M. Zitnik, D. Jurafsky, and J. Leskovec, "Embedding logical queries on knowledge graphs," *Advances in neural information processing systems*, vol. 31, 2018.

[31] H. Ren, W. Hu, and J. Leskovec, "Query2box: Reasoning over knowledge graphs in vector space using box embeddings," *arXiv preprint arXiv:2002.05969*, 2020.

[32] H. Ren and J. Leskovec, "Beta embeddings for multi-hop logical reasoning in knowledge graphs," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 716–19 726, 2020.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova,

"BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[34] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[35] B. Y. Lin, W. Zhou, M. Shen, P. Zhou, C. Bhagavatula, Y. Choi, and X. Ren, "Commongen: A constrained text generation challenge for generative commonsense reasoning," *arXiv preprint arXiv:1911.03705*, 2019.

[36] X. Pi, W. Zhong, Y. Gao, N. Duan, and J.-G. Lou, "Logigan: Learning logical reasoning via adversarial pre-training," *arXiv preprint arXiv:2205.08794*, 2022.

[37] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[38] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, and H. Chen, "Reasoning with language model prompting: A survey," *arXiv preprint arXiv:2212.09597*, 2022.

[39] A. Talmor, Y. Elazar, Y. Goldberg, and J. Berant, "olmpics-on what language model pre-training captures," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 743–758, 2020.

[40] A. Yang, Q. Wang, J. Liu, K. Liu, Y. Lyu, H. Wu, Q. She, and S. Li, "Enhancing pre-trained language representations with rich knowledge for machine reading comprehension," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2346–2357.

[41] R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, G. Cao, D. Jiang, M. Zhou *et al.*, "K-adapter: Infusing knowledge into pre-trained models with adapters," *arXiv preprint arXiv:2002.01808*, 2020.

[42] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1441–1451. [Online]. Available: https://aclanthology.org/P19-1139

[43] D. Yu, C. Zhu, Y. Yang, and M. Zeng, "Jaket: Joint pre-training of knowledge graph and language understanding," *arXiv preprint arXiv:2010.00796*, 2020.

[44] J. Guan, F. Huang, Z. Zhao, X. Zhu, and M. Huang, "A knowledge-enhanced pretraining model for commonsense story generation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 93–108, 2020.

[45] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, "K-bert: Enabling language representation with knowledge graph," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, 2020, pp. 2901–2908.

[46] T. Sun, Y. Shao, X. Qiu, Q. Guo, Y. Hu, X. Huang, and Z. Zhang, "Colake: Contextualized language and knowledge embedding," *arXiv preprint arXiv:2010.00309*, 2020.

[47] Y. Lu, H. Lu, G. Fu, and Q. Liu, "Kelm: knowledge enhanced pre-trained language representations with message passing on hierarchical relational graphs," *arXiv preprint arXiv:2109.04223*, 2021.

[48] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu, "Ernie: Enhanced representation through knowledge integration," *arXiv preprint arXiv:1904.09223*, 2019.

[49] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020.

[50] Y. Qin, Y. Lin, R. Takanobu, Z. Liu, P. Li, H. Ji, M. Huang, M. Sun, and J. Zhou, "Erica: Improving entity and relation understanding for pre-trained language models via contrastive learning," *arXiv preprint arXiv:2012.15022*, 2020.

[51] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu *et al.*, "Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation," *arXiv preprint arXiv:2107.02137*, 2021.

[52] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," *Advances in neural information processing systems*, vol. 30, 2017.

[53] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

[54] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang, "Kepler: A unified model for knowledge embedding and pre-trained language representation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 176–194, 2021.

[55] Y. Lin, Z. Liu, H. Luan, M. Sun, S. Rao, and S. Liu, "Modeling relation paths for representation learning of knowledge bases," *arXiv preprint arXiv:1506.00379*, 2015.

[56] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," *Advances in neural information processing systems*, vol. 26, 2013.