Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 6689740, 9 pages
https://doi.org/10.1155/2021/6689740

Research Article

Taking a Closed-Book Examination: Decoupling KB-Based Inference by Virtual Hypothesis for Answering Real-World Questions

Xiao Zhang [1] **and Guorui Zhao** [2]

[1]*School of Mechanical Electronic and Information Engineering, China University of Mining and Technology, Beijing 100083, China*
[2]*Coal Mining Research Institute, China Coal Technology and Engineering Group, Beijing 100013, China*

Correspondence should be addressed to Guorui Zhao; 147466405@qq.com

Received 24 October 2020; Revised 18 January 2021; Accepted 30 January 2021; Published 22 February 2021

Academic Editor: Qiangqiang Yuan

Complex question answering in real world is a comprehensive and challenging task due to its demand for deeper question understanding and deeper inference. Information retrieval is a common solution and easy to implement, but it cannot answer questions which need long-distance dependencies across multiple documents. Knowledge base (KB) organizes information as a graph, and KB-based inference can employ logic formulas or knowledge embeddings to capture such long-distance semantic associations. However, KB-based inference has not been applied to real-world question answering well, because there are gaps among natural language, complex semantic structure, and appropriate hypothesis for inference. We propose decoupling KB-based inference by transforming a question into a high-level triplet in the KB, which makes it possible to apply KB-based inference methods to answer complex questions. In addition, we create a specialized question answering dataset only for inference, and our method is proved to be effective by conducting experiments on both AI2 Science Questions dataset and ours.

1. Introduction

Teaching machines to answer complex questions like human beings is a very challenging task at the intersection of nature language processing (NLP), information retrieval (IR), and artificial intelligence (AI), which mainly needs three techniques, i.e., question understanding, answer retrieval, and inference. There are three subtasks specifically to evaluate the corresponding techniques: Question Answering over Knowledge Base (KBQA) is a typical task to evaluate question understanding; Text Retrieval Question Answering (TREC QA) and Reading Comprehension (RC) are good tasks to evaluate answer retrieval and answer selection; Link Prediction and Knowledge Base Completion (KBC) are traditional tasks to evaluate inference.

After achieving progress in these subtasks, researchers begin to turn their passion to more comprehensive and complex question answering (QA) tasks. Allen Institute for Artificial Intelligence (AI2) proposes a science test which is a real-world examination for elementary students and middle school students, and it is even viewed as a standardized measure of AI. An example question in a science test is shown as follows. Q1: Peach trees have sweet-smelling blossoms and produce rich fruit. What is the main purpose of the flowers of a peach tree? (Answer is A.)

(A) To attract bees for pollination
(B) To create flower arrangements
(C) To protect the tree from disease
(D) To feed migratory birds

Such complex questions are hardly solved by any single technique mentioned above, and it is also difficult to effectively combine these techniques, which makes the task far from solved.

Retrieval can get effect instantly on a question whose answer (or some keywords inside) is near the question in background corpus, and the method achieved the best performance on 8th Grade Science Challenge [1]. However, those methods can do nothing to a question whose answer does not occur in the same document with the question, and they cannot capture long-distance dependencies across documents which furnish evidence for choosing the answer. This limitation makes retrieval have little space to improve with fixed corpus. According to [2], 77% of questions need inference, and retrieval is not viewed as true artificial intelligence [1].

On the other hand, knowledge bases (KBs) are graph-structured background data which contain vast long-distance semantic associations. Inference on KB is expected to capture such long-distance semantics as evidence by logic formulas [3, 4] and knowledge graph embedding [5, 6], and it has been proven to be effective in link prediction on KB and KBC task. However, KB-based inference does not fit real-world QA well; there are two reasons:

(1) There is a gap between a natural language question and the semantic structure in KBs. Semantic parsing is used to transform a natural language to the semantic structure on the KB; e.g., for Q1, the question is expected to be Purpose (Flower, X), and its correct answer is expected to be Attract (Flower, Bee) $\wedge$ Do (Be, Pollination). However, it is difficult to obtain such precise structures, because the quality of semantic parsing is far from satisfactory [7].

(2) Even if the question can be precisely transformed into a structure by semantic parsing, complex structure in KBs is not appropriate as a hypothesis for inference. In inference, a hypothesis is a candidate proposition which needs to be proven by evidence; e.g., we substitute X in Purpose (Flower, X) with the candidate answer (as Figure 1(a) shows), and the recursive structure can be viewed as the hypothesis for inference. However, formulas used to infer such a complex structure should have a form like $R1\,(X, Y) \wedge R2\,(Y, Z) \wedge \ldots \Rightarrow$ Purpose (X, (Attract (X, Y) $\wedge$ Do (Y, Z))), which is far from frequent in KBs and difficult to be found by formula learners.

An intuitive solution to complex hypothesis is to unfold the recursive structure and divide it into several atomic hypotheses; e.g., the structure in Figure 1(a) is unfolded into four triplets by establishing relationships between Flower and the two entities in the answer, shown in Figure 1(b), i.e., Has (Peach tree, Flower) $\wedge$?R (Flower, Bee) $\wedge$?R (Flower, Pollination) $\wedge$ Do (Bee, Pollination). However, after unfolding and dividing, some structural information in the original question is missing, which embodies two aspects:

(a) There are no explicit relations between some entities. After unfolding, the original relation between the question and the answer (e.g., Purpose in Figure 1(a)) is no longer the relation of some atomic triplets. For example, in Figure 1(b), the relation of Flower, Pollination is unknown, and the relation of Flower, Bee is not Attract anymore because of Purpose's influence.

(b) There are no associations among atomic hypotheses. After dividing, we assume that atomic hypotheses are independent and each atomic hypothesis is inferred by its specific formulas. For example, after dividing, Eat (Bee, pollen) $\Rightarrow$ Do (Bee, Pollination) which supports Do (Bee, Pollination) is irrelevant with Has (Peach tree, Flower). Actually, there should be associations among these hypotheses, and it is these associations that make atomic hypotheses the original question.

To resolve these problems, this paper proposes decoupling KB-based inference from question answering by transforming a complex QA pair into a virtual high-level hypothesis on the KB.

(a) For the entity pairs which have no explicit relations, we create a virtual relation Rq to replace unknown relations of entity pairs, and represent Rq into distributed semantic space according to the linguistic expressions of the question. We expect different dimensions of Rq can capture relations of different entity pairs. For example, Rq (Flower, Pollination) and Rq (Flower, Bee) focus on different aspects of Rq.

(b) For the atomic hypotheses among which there are no associations, we create a virtual hypothesis Rq (H, T) to combine all possible atomic hypotheses, and adapt original KB-based inference methods to infer it. For example, atomic hypotheses in Figure 1(b) are combined into Rq ({Peach Tree, Flower}, {Bee, Pollination}) (in Figure 1(c)). Therefore, the virtual hypothesis is treated as a whole and can be supported by evidences obtained from any pair of entity $h \in H$ and entity $t \in T$. For example, in Figure 1(c), a path Flower $\longrightarrow {}^{\text{Has}}$Pollen $\longrightarrow {}^{\text{Feed}}$Bee on the KB can produce a formula Has (h, x) $\wedge$ Feed (x, t) $\Rightarrow$ Rq (H, T), which is an evidence for the virtual hypothesis. At last, we build a joint inference model to eliminate irrelevant or noisy evidence (including formulas and embeddings) which may be introduced by meaningless entity pairs, e.g., Peach Tree, Pollination irrelevant to the original question.

We conduct experiments on AI2 Science Dataset to examine whether our inference method can acquire extra long-distance knowledge and bring improvement for read-world QA task. Moreover, in order to explore more deeply the effect of inference and focus on the questions that definitely need inference, we propose a new dataset, named as InfQAD. This dataset totally contains more than 11,000 real-world examination questions in seven subjects with two languages (English and Chinese), where questions that can be answered only by simple retrieval have been filtered out. The experimental results on InfQAD show that logic inference and embedding-based method concentrate on different aspects of questions and they can complement each other.

In summary, the contributions of this paper are shown as follows:

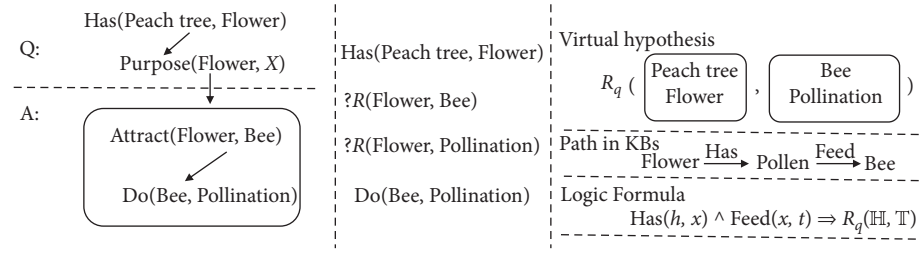

Figure 1: (a) A recursive structure obtained from semantic parsing. (b) Atomic hypotheses obtained by unfolding recursive structure. (c) A virtual hypothesis and its evidences.

We decouple KB-based inference from question answering by creating a virtual hypothesis and apply inference to answer complex questions, which not only utilizes long-distance semantic associations but also bridges the gap between natural language questions and hypotheses for inference.

We create a new dataset from real-world examination questions to specifically evaluate the performance of inference methods on those complicated questions which need inference to resolve. It contains seven subsets of different subjects and may promote study on domain specific inference.

We conduct an experiment on AI2 Science Questions Dataset to prove that inference can improve the performance of retrieval. After that, we compare several inference methods on InfQAD, and our method outperforms baselines.

2. KB-Based Inference

An inference task should contain a hypothesis and evidence, and inference is a process of collecting evidence to prove the hypothesis. For simple inference task on the KB, the hypothesis usually has the form of $r(h, t)$, and the evidence can be a path, loop, or subgraph. There are mainly two types of models: probabilistic logic inference and knowledge graph embedding.

2.1. Probabilistic Logic Inference. Probabilistic logic inference utilizes various logic formulas to perform probabilistic inference. A logic formula has a head and a body, respectively, corresponding to hypothesis and evidence, noted as Body $\Longrightarrow$ Head. For example, Has $(x, y) \wedge$ Feed $(y, z) \Longrightarrow$ Attract (x, z) is a formula which can infer Attract (Flower, Bee), and its head is Attract (x, z), and its body is an abstraction of evidence, Has $(x, y) \wedge$ Feed (y, z).

Logic formulas usually are mined automatically from KB with confidence or weights. For a specific head or class of heads, bodies of formula are frequent structures on the KB, and mining such frequent structures is an important component in several probabilistic logic models, e.g., Markov Logic Network (MLN) [4]. Random walk algorithm is proposed to perform sampling frequent structures on knowledge graph, and the frequent structures are conceptualized as formula bodies. After that, the counts of formulas are used to calculate formulas' confidence or learn formulas' weights. Algorithm 1 shows a general process of mining formulas on KB by random walk.

The algorithm takes a class of hypotheses $H(X, Y)$ as input and starts random walks from the head entity x in each ground hypothesis (Lines 1–3). For example, the algorithm finds a path from Flower to Bee for Attract (Flower, Bee), e.g., Flower $\longrightarrow^{\text{Has}}$ Pollen $\longrightarrow^{\text{Feed}}$ Bee, and then the path is conceptualized as the body of formula (Lines 4-5), i.e., Has $(x, y) \wedge$ Feed (y, z). After obtaining weighted formulas (Lines 6-7), probabilistic inference model employs the formulas as features and estimates the probability of the hypothesis as

$$p(h) = \frac{\phi\left(\sum_{f_i \in F} w_i x_i\right)}{\sum_{x_i \in X_i} \phi\left(\sum_{f_i \in F} w_i x_i\right)}, \tag{1}$$

where x_i is a value about f_i, e.g., truth value or count, and the denominator is a normalizing constant. φ is a nonlinear function; e.g., ϕ is an exponential function in Markov Logic Network.

2.2. Knowledge Graph Embedding. Knowledge graph embedding (KGE) model represents entities and relations as low-dimension numeric vectors or tensors and expects that arithmetical operations among embeddings can capture implicit relationships among elements. KGE can apply to inferring the hypothesis by defining a score function, noted as $F_{r(h,t)}$; e.g., TransE [5] defines its score function as

$$F_{r(h,t)} = -\left\| E_h + E_r - E_t \right\|, \tag{2}$$

where E_h, E_t, E_r are the embeddings of two entities and one relation, respectively. At training, the score of the triplet in KB is expected to be larger than triplets not in KB. After multiple rounds, embeddings are considered to contain implicit semantics and can perform inference.

3. Decoupling Inference from QA

The above KB-based inference methods all take a triplet $r(h, t)$ in the KB as a hypothesis for inference, so if a complex question can be transformed into a triplet, KB-based inference method can solve it. This section describes a method that can distill a high-level triplet from the question as the hypothesis, and we call it virtual hypothesis.

Input: KB, hypothesis $H(X, Y)$
Output: F, W
(1) Collect all instances $H(x, y)$
(2) For $H(x, y) \in H(X, Y)$
(3) Repeat Random walk from x
(4) If terminal $t = y$
(5) Conceptualize path x to y, as $B(X, Y)$, Put $B(X, Y) \Rightarrow H(X, Y)$ into F
(6) Prune $f \in F$ by supports or heuristic rules
(7) Learning weights W for F
(8) Output F and W

Algorithm 1: Mining formulas by random walk.

3.1. Virtual Hypothesis. We try to transform a pair of question and option into a high-level triplet $R_q(H, T)$ and we propose the first assumption here.

Assumption 1. The virtual hypothesis $R_q(H, T)$ for a pair of question and option is the combination of all possible triplets $r(h, t)$, where r is an implicit relation between an entity h in question and an entity t in option. Thus, $h \in H$, $t \in T$, and R_q is an integration of r.

We employ TransE model to explain the correction of Assumption 1. According to TransE, if $r(h, t)$ is true, $h + r \approx t$, and the distance of $r(h, t)$ in distributed space, $D_{r(h,t)} = \|E_h + E_r - E_t\|$, should be close to zero. We represent the sum of distances for all possible triplets as D_q, e.g., triplets in Figure 1(b), and $D_q = \sum_{i=1}^{n} \|E_{hi} + E_{ri} - E_{ti}\|$. According to Triangle Inequality, $\|\sum_{i=1}^{n} E_{hi} + \sum_{i=1}^{n} E_{ri} - \sum_{i=1}^{n} E_{ti}\| \le D_q$. However, r_i may have no clear definition; e.g., the relation between Flower and Pollination in Figure 1(c) is unknown. To handle this, we create a virtual relation type R_q and set $E_{R_q} = \sum_{i=1}^{n} E_{r_i}$, so the distance of $R_q(H, T)$ is $D_{R_q}(H, T) = \|EH + E_{R_q} - ET\|$, and $D_{R_q}(H, T) \le D_q$ which means the virtual hypothesis $R_q(H, T)$ is true if and only if all atomic hypotheses $r_i(h_i, t_i)$ are true. Thus, we believe the virtual hypothesis covers all semantics in atomic hypotheses, but there is temporarily no association between the virtual hypothesis and the original question. We propose the second assumption as follows.

Assumption 2. In distributed space, the embedding of virtual relation R_q is close to the embedding of the original question without entities.

According to Assumption 1, R_q is a combination of implicit relations between question entities and option entities, which should be what the question describes. For example, a simple question "Who is first emperor of Tang Dynasty" can be represented by a triplet FirstEmperor (Tang Dynasty, X), and then "Who is first emperor of" describes the semantics of the relation type FirstEmperor. Therefore, Assumption 2 is reasonable, and we get the concrete definition of a virtual hypothesis $R_q(H, T)$—H and T, respectively, are the entity sets in question and option, and R_q is the question without entities. At present, we have distilled a high-level triplet from a pair of question and option as the simple hypothesis for inference.

3.2. Logic Inference with Virtual Hypothesis. When we employ logic formulas to infer a normal triplet $r(h, t)$ on the KB, we select applicative logic formulas by the relation r; e.g., the formula Has $(x, y) \wedge$ Feed $(y, z) \Rightarrow$ Attract (x, z) which is obtained from the instance Has (Flower, Pollen) Feed (Pollen, Bee) $\wedge$ Attract (Flower, Bee) can be also used to infer Attract (Honeycomb, Bear). However, the virtual relation R_q is specific to a question, and no two R_q can share their formulas. To capture associations between R_q (H, T) and formulas, we propose the third assumption.

Assumption 3. If a formula f can be used to infer $R_q(H, T)$, the body of f should be close to R_q in distributed space.

Intuitively, some formulas have such a property; e.g., in Father $(x, y) \wedge$ Father $(y, z) \Rightarrow$ Grandfather (x, z), Father + Father should be close to Grandfather. More formally, we still employ TransE to explain the correction. For a formula f, $r_1(x_1, x_2) \wedge \ldots \wedge r_n(x_n, x_{n+1}) \Rightarrow r_1(x_1, x_{n+1})$, if the body of f is true, and $r_i(x_i, x_{i+1})$ is true. Thus, $E_{x_i} + E_{r_i} = E_{x_{i+1}}$, and then $E_{x_i} + \sum_{i=1}^{n} E_{r_i} = E_{r_{n+1}}$. We define the embedding of f as $E_{x_f} = \sum_{i=1}^{n} E_{r_i}$, so $E_{x_1} + E_{r_f} = E_{x_{n+1}}$. Since the head of f is also true, $E_{x_1} + E_{r_r} = E_{x_{n+1}}$, so we get $E_f = E_r$, which is exactly Assumption 3.

We search paths from any entity in H to any entity in T and transform them into the body of formulas f. Then, we represent f as $E_f = \sum_{i=1}^{n} E_{r_i}$ and calculate the similarity between f and the virtual relation R_q as

$$\mathrm{Sim}\left(f, R_q\right) = \left\|E_f - E_{R_q}\right\|. \tag{3}$$

Finally, we employ the similarity Sim (f, R_q) to replace the count of f's instances between H and T, so equation (1) changes to

$$P_l(h) = \frac{1}{Z}\phi\left(\sum_{f_i \in F} w_i * -\left\|E_h - E_{f_i}\right\|\right), \tag{4}$$

where Z is the normalizing constant.

3.3. KB Embedding with Virtual Hypothesis. To adapt KB embedding model, i.e., TransE, to the virtual hypothesis, we give the fourth assumption.

Assumption 4. In distributed space, the entity set H in question is close to the entity set T in option under the translation of virtual relation R_q.

Assumption 4 can be deduced from Assumption 2. In Assumption 2, E_{R_q} is the embedding of the original question without entities, so $E_{R_q} + E_H$ is exactly the embedding of the question E_q. T is the set of entities in option, so E_T is close to the embedding of the option E_o. Thus, $\|E_H + E_{R_q} - E_T\| \approx \|E_q - E_o\|$, and its right part means question is close to option in distributed space, which is a truth when the option is the correct answer. Therefore, TransE with virtual hypotheses is a kind of text inference.

3.4. Joint Objective Formalization. To utilize different types of evidence from both logic formulas and KB embedding, we build a joint objective G which combines P_l and P_h as

$$G(h) = \alpha P_l(h) + \beta P_l(n), \tag{5}$$

where α and β are hyper-parameters, and $\alpha + \beta = 1$. To simultaneously learn word embeddings and KB embeddings, we minimize a margin-based ranking criterion over the training set as

$$\beth = \sum_{h' \in O_w} \left[G(h') + \gamma - G(h)\right]_+, \tag{6}$$

where γ is a margin, O_w is the wrong option set, and h' is a hypothesis formed by the question and a wrong option. The optimization is carried out by stochastic gradient descent with the additional L2 regularization on parameters.

4. Experiments

To explore whether our method would acquire long-distance knowledge and bring an improvement for read-world QA task, we combine our methods with a retrieval-based method and conduct an experiment on AI2 Science Question Dataset1. After that, to further explore the effect of inference and focus on questions which need inference, we create an Inference Question Answering Dataset (InfQAD), in which questions cannot be answered by search or retrieval. After that, we compare several types of inference on InfQAD.

4.1. Evaluation on AI2 Science Questions

4.1.1. Dataset. AI2 Science Question Dataset contains 5343 4-way multiple-choice science questions without diagrams at the elementary and middle school levels, and they are divided into Train, Dev, and Test. Table 1 shows the statistics of this dataset. We employ three types of resources as backgrounds, including Wikipedia, Freebase [8], and ConceptNet [9].

4.1.2. Setting. We implement a retrieval method based on Lucene which is an open-source information retrieval software library, and we employ the method to build reverse index on the whole Wikipedia dump. We concatenate a question with an option as the query for retrieval and calculate the average of Top-3 scores. We rank options by the average scores, and the highest one is the final answer. All questions and options are preprocessed by CoreNLP [10]. For our logic-based method, we use simple maximal matching algorithm to extract entities from the question and options, respectively. When collecting ground formulas for hypotheses, we employ a typical random walk algorithm to run on both Freebase and ConceptNet and limit the maximal length of formula to 4. For our embedding-based method, we represent a question by the sum of embeddings of its words which were pretrained by GloVe [11] with 100 dimensions. We combine results of the retrieval-based method and our methods by two steps:

(1) For each solver, we normalize scores across the answer options for a given question

(2) We send normalized scores into a classifier which can output correct/incorrect with confidence, and the correct option with the maximal confidence is treated as the final answer

4.1.3. Results and Analysis. We show the accuracy of methods in Table 2, where +Emb and +Logic represent adding embedding and logic formulas, respectively. We can obtain the following observations:

(1) Combining two types of inference methods with retrieval can improve performances, which proves that decoupling inference by virtual hypothesis is effective and KB-based inference can utilize a mass of extra long-distance knowledge to improve the performance of the retrieval method.

(2) The promotion on middle school dataset is more obvious than that on elementary dataset, which implies middle school examination is more difficult than elementary examination, and difficult questions need inference more.

(3) Only adding embedding into retrieval leads to performance reduction. We think the reason is that wrong answers from the unsuitable solver may affect others. Logic inference tends to refuse to answer with low confidence, while embedding method gives answers in any case, which may distract retrieval from giving the correct answer.

4.2. Inference QA Dataset Construction. Retrieval-based method achieves a good performance on AI2 dataset, but the experiment above shows that retrieval may affect the further exploration of inference. Therefore, we propose constructing a new Inference Question Answering Dataset, named as InfQAD, which only contains complicated questions that need inference. InfQAD contains 11,393 examination questions in seven subjects with two languages (five subjects

Table 1: Statistics of AI2 Science questions.

Dataset	Train	Dev	Test	Total
Elementary	574	143	717	1636
Middle	1583	485	1639	3707

Table 2: Accuracy of methods on AI2 Science questions.

Methods	Elementary (%)	Middle (%)
Retrieval	45.33	40.02
+Emb	43.65	39.29
+Logic	45.33	40.81
+Emb + Logic	45.46	41.12

in English and two subjects in Chinese), and in the dataset, questions that only need retrieval have been filtered out. Table 3 shows the statistics of InfQAD. We construct InfQAD by two major steps: question collection and question filtration.

For five subjects in English, we download questions from the CK12 website2. There is a downloadable quiz in almost every topic, and the quiz usually contains ten questions. We only keep 4-way multiple-choice questions without diagrams as AI2 does. For two subjects in Chinese, we collect about 200 senior high school entrance examination papers, which also only keeps 4-way multiple-choice questions without diagrams.

To filter out questions which can be answered by retrieval, we treat Lucene as a standard retrieval method and employ it to score each pair of question and option. We sort questions according to the difference between the score of its correct answer and the maximal score of its incorrect option in descending order. We remove top questions and make the accuracy of Lucene on the rest of questions 25% which equals the accuracy of random choice. We believe Lucene fails for the rest of the questions, and they can be viewed as questions that need inference to resolve, approximatively.

4.3. Evaluation on Inference Questions

4.3.1. Methods Compared. We compare different kinds of methods on InfQAD, including probabilistic logic inference (in Table 4(b)), embedding-based inference (in Table 4(c)), and ensemble inference (in Table 4(d)). Probabilistic logic inference includes the following:

(1) Traditional MLN [4], which treats all hypotheses as the same relation, and questions share all weighted formulas

(2) Cluster-based MLN, which first clusters questions by the similarities between questions and then trains an MLN Model for each cluster of questions

(3) Our method described in Section 3.2, noted as VHLogic

Embedding-based inference contains two approaches which both estimate the similarity between questions and options but employ different methods of representing text, i.e., SUM [12] and GRU [13]. Ensemble inference is combining VHLogic with two methods of embedding-based inference as in Section 3.4. We also add the results of random choice (Random) and the retrieval-based method (Retrieval) into the result in Table 4(a) for comparison.

Table 3: Statistics of InfQAD.

Datasets	Train	Dev	Test	Total
English				
Biology	242	48	193	483
Chemistry	749	149	599	1497
Earth	737	221	516	1474
Life	421	126	295	842
Physical	293	87	205	585
Chinese				
Biology	1919	300	1618	3837
History	1338	300	1037	2675

Table 4: Accuracy of methods on InfQAD Chinese.

	Datasets	Chinese	
	Methods	Biology (%)	History (%)
4(a)	Random	25.00	25.00
	Retrieval	24.43	27.17
4(b)	MLN	32.14	29.60
	MLN (cluster)	28.18	28.45
	VHLogic	35.04	35.10
4(c)	SUM	44.44	42.62
	GRU	40.30	36.84
4(d)	VHLogic + SUM	44.31	42.81
	VHLogic + GRU	40.60	36.07

4.3.2. Setting. We implement MLN as described in [14]. We implement SUM method and employ a GRU tool in Java3. We still use pretrained word vectors by GloVe [11] with 100 dimensions for English questions and train word vectors with 100 dimensions on Baidu Baike for Chinese questions. In this experiment, we only employ ConceptNet as the KB for both English and Chinese questions.

4.3.3. Results. We show the accuracy of methods in Tables 4 and 5, and we can obtain the following observations:

(1) Comparing VHLogic with other logic inference methods in Tables 4(b) and 5(b), VHLogic has the best performance on almost all subsets, which indicates decoupling logic inference is effective, and the distributed similarity between hypotheses and formulas improves the performance of inference.

(2) Comparing logic inference methods in Table 4(b) with SUM and GRU in Table 4(c), there is no obvious evidence that some kind of method could achieve better performance than another kind. It implies that different types of inference are better at questions in some subjects and may complement each other. The experimental results are also applicable to methods in Table 5(b) compared with Table 5(c).

Table 5: Accuracy of methods on InfQAD English.

	Datasets	English				
	Methods	Biology (%)	Chemistry (%)	Earth (%)	Life (%)	Physical (%)
5(a)	Random	25.00	25.00	25.00	25.00	25.00
	Retrieval	25.91	23.04	24.03	27.46	23.90
5(b)	MLN	27.46	23.87	27.71	24.07	31.71
	MLN (cluster)	**32.64**	23.54	27.32	28.14	26.83
	VHLogic	31.61	25.21	**34.11**	**31.19**	34.15
5(c)	SUM	29.53	26.38	33.53	24.75	37.07
	GRU	29.53	**32.22**	32.55	30.17	**39.02**
5(d)	VHLogic + SUM	34.20	27.88	**35.85**	30.85	37.56
	VHLogic + GRU	**34.72**	**33.39**	34.10	**31.53**	**39.51**

(3) Comparing ensemble inference methods in Table 4(d) with single methods in Tables 4(b) and 4(c), ensemble methods outperform single methods on almost all subsets, which proves that different types of inference concentrate on different aspects of questions and they can complement each other. The experimental results are also applicable to methods in Table 5(d) compared with Tables 5(b) and 5(c).

4.3.4. Data Analysis. To analyze various causes of breakdowns, we sample 100 questions answered incorrectly by VHLogic and roughly classify them into several categories (shown in Figure 2):

(1) *Complex Relation.* This category is that there is a relationship among more than two entities in the question, and the relationship is the key to answer the question. This category is the largest category, and 26% of questions belong to it.

(2) *Missing Entity from KB.* This category is that there is a key entity missing from the KB, which leads to key formulas not being found. This category contains 22% of questions.

(3) *No Entity in Answer.* This category is also about entities and contains 5% of questions. Answers of these questions contain no entity but numbers, modifications, or other elements.

(4) *No Formula.* Sometimes, there are still no paths or useful formulas between entities, though entities in the question and the answer are both linked in KB. 13% of questions belong to this category.

(5) *Irrelevant Formulas.* Irrelevant formulas are noise in inference process, which may disturb useful formulas, and this category has 14% of questions.

(6) *Key Modifier.* Comparative degree, superlative degree, and other modifiers may be the key of answering the question, but these modifiers cannot be captured by VHLogic. There are 8% of questions belonging to this category.

There are other categories: 7% of questions belong to Math category which needs mathematical operations. 4% of questions need Global Information; e.g., the answer is All of the above, and there are 1% of questions with Wrong Answer.

5. Related Work

Our work is related to two types of work: question answering and KB-based inference. In recent years, various QA tasks and datasets have been emerging in an endless stream. WebQuestions [7], BAbI [15], and SimpleQuestions [16] mainly evaluate question understanding and assume the correct parsing results must be able to get the correct answer. MCTest [17] contains questions with 4 answer choices per question like ours, but each question and its answer in MCTest come from a given story. The Children Book Test [18] and CNN/Daily Mail dataset [19] view cloze test as a kind of QA task, while SQuAD [20] also employs word or phrase in original text as the answer. AI2 Science Dataset [1] is the most related to our InfQAD, but InfQAD only contains the questions which need inference. Aristo [21] is a QA system for science questions which combines 5 solvers including IR, MLN [22], and other inference methods. Aristo extracts inference rules from texts by patterns, while our method mines formulas from KB.

On the other hand, knowledge base (KB) organizes information as a graph. Graph learning has been widely used in many other fields such as image classification [23, 24]. KB-based inference mainly has two types of approaches: probabilistic logic inference and knowledge graph embedding. Besides MLN [1] mentioned in the previous sections, Inductive Logic Programming (ILP) [3], PSL [13], and PRA [12] all belong to probabilistic logic inference models. These models obtain logic formulas from knowledge graph and perform probabilistic inference, but they cannot handle virtual hypothesis as VHLogic does. TransE, RESCAL [6], TransH [25], and TransR [26] are all embedding-based methods, and relative to our method, they employ different similarity functions to calculate scores of hypotheses. There are also several methods to represent formula by embeddings, including PTransE [27], RNN [28], and ProPPR + MF [29–31], while these methods only represent formulas, but do not simultaneously represent texts as our method does.

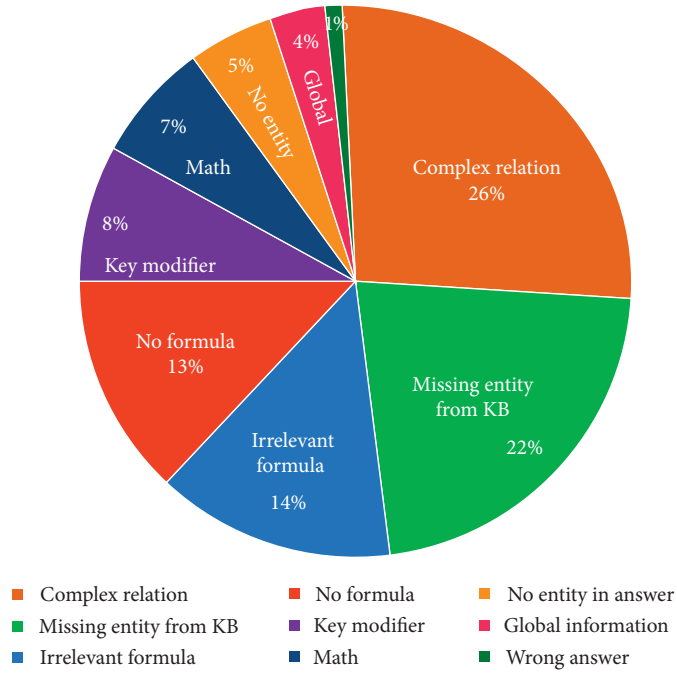

Figure 2: Statistics of breakdowns.

6. Conclusion

We propose a method to decouple KB-based inference from real-world QA by creating a high-level triplet on the KB named as virtual hypothesis, and adjust logic-based and embedding-based method to inferring it. The experimental results prove that our method is effective and is a promising method to apply inference to QA. In addition, we propose a specialized question answering dataset only for inference, named as InfQAD. We compare various inference methods on InfQAD and find that different types of inference are skillful in different subjects and combing them will improve the performance. At last, we analyze various causes of breakdowns, which can be helpful for the future study on domain specific inference. In the future, there are two aspects of our work that need deeper exploration. We try to find a better way to represent virtual hypotheses and try to reconstruct textual knowledge base to better capture long-distance evidence as formulas.

Data Availability

The dataset concerns commercial confidentiality, so it is not suitable for publishing.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] M. Richardson and P. Domingos, "Markov logic networks," *Machine Learning*, vol. 62, no. 1-2, pp. 107–136, 2006.

[2] P. Jansen, N. Balasubramanian, M. Surdeanu, and P. Clark, "What's in an explanation? characterizing knowledge and inference requirements for elementary science exams," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2956–2965, The COLING 2016 Organizing Committee, Osaka, Japan, December 2016.

[3] S. Muggleton and L. de Raedt, "Inductive logic programming: theory and methods," *The Journal of Logic Programming*, vol. 19-20, pp. 629–679, 1994.

[4] N. Lao, T. Mitchell, and W. W. Cohen, "Random walk inference and learning in a large scale knowledge base," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 529–539, Association for Computational Linguistics, Edinburgh, UK, July 2011.

[5] B. Antoine, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pp. 2787–2795, Lake Tahoe, NV, USA, December 2013.

[6] M. Nickel, V. Tresp, and H.-P. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 809–816, Bellevue, WA, USA, June 2011.

[7] J. Berant, A. Chou, F. Roy, and P. Liang, "Semantic parsing on freebase from question-answer pairs," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1533–1544, Association for Computational Linguistics, Seattle, WA, USA, October 2013.

[8] K. D. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pp. 1247–1250, ACM, Vancouver, Canada, June 2008.

[9] H. Liu and P. Singh, "ConceptNet—a practical commonsense reasoning tool-kit," *BT Technology Journal*, vol. 22, no. 4, pp. 211–226, 2004.

[10] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of*

the Association for Computational Linguistics: System Demonstrations, pp. 55–60, Association for Computational Linguistics, Baltimore, MD, USA, June 2014.

[11] J. Pennington, R. Socher, and C. Manning, "GloVe: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Association for Computational Linguistics, Doha, Qatar, October 2014.

[12] L. Ni, T. Mitchell, and W. William Cohen, "Random walk inference and learning in a large scale knowledge base," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 529–539, Association for Computational Linguistics, Edinburgh, UK, July 2011.

[13] M. Brocheler, L. Mihalkova, and L. Getoor, "Probabilistic similarity logic," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, CA, USA, August 2012.

[14] Z. Wei, J. Zhao, and K. Liu, "Mining inference formulas by goal-directed random walks," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1379–1388, Association for Computational Linguistics, Austin, TX, USA, November 2016.

[15] Z. Wei, J. Zhao, K. Liu, Z. Qi, Z. Sun, and G. Tian, "Large-scale knowledge base completion: inferring via grounding network sampling over selected instances," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 1331–1340, ACM, Melbourne Australia, October 2015.

[16] B. Antoine, N. Usunier, S. Chopra, and J. Weston, "Large-scale simple question answering with memory networks," 2015, http://arxiv.org/abs/1506.02075.

[17] M. Richardson, C. J. C. Burges, and E. Renshaw, "MCTEST: a challenge dataset for the open-domain machine comprehension of text," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, WA, USA, October 2013.

[18] F. Hill, A. Bordes, S. Chopra, and J. Weston, "The goldilocks principle: reading children's books with explicit memory representations," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, May 2015.

[19] K. M. Hermann, T. Kocisky, E. Grefenstette et al., "Teaching machines to read and comprehend," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pp. 1693–1701, Montreal, Canada, December 2015.

[20] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Association for Computational Linguistics, Austin, TX, USA, November 2016.

[21] P. Clark, O. Etzioni, T. Khot et al., "Combining retrieval, statistics, and inference to answer elementary science questions," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 2580–2586, Phoenix, AZ, USA, February 2016.

[22] T. Khot, N. Balasubramanian, E. Gribkoff, A. Sabharwal, P. Clark, and O. Etzioni, "Exploring markov logic networks for question answering," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 685–694, Association for Computational Linguistics, Lisbon, Portugal, September 2015.

[23] F. Luo, L. Zhang, B. Du, and L. Zhang, "Dimensionality reduction with enhanced hybrid-graph discriminant learning for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, pp. 5336–5353, 2020.

[24] F. Luo, L. Zhang, X. Zhou, T. Guo, Y. Cheng, and T. Yin, "Sparse-adaptive hypergraph discriminant analysis for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 6, pp. 1082–1086, 2020.

[25] W. Y. Wang and W. W. Cohen, "Learning first-order logic embeddings via matrix factorization," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*, New York, NY, USA, July 2016.

[26] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2181–2187, Austin, TX, USA, January 2015.

[27] Y. Lin, Z. Liu, H. Luan, M. Sun, S. Rao, and S. Liu, "Modeling relation paths for representation learning of knowledge bases," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 705–714, Association for Computational Linguistics, Lisbon, Portugal, September 2015.

[28] A. Neelakantan, B. Roth, and A. McCallum, "Compositional vector space models for knowledge base completion," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 156–166, Association for Computational Linguistics, Beijing, China, July 2015.

[29] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pp. 926–934, Sydney, Australia, December 2019.

[30] K. Guu, J. Miller, and P. Liang, "Traversing knowledge graphs in vector space," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 318–327, Association for Computational Linguistics, Lisbon, Portugal, September 2015.

[31] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 1112–1119, Citeseer, Québec City, Canada, July 2014.