Knowledge-Based Systems 250 (2022) 108943

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

# Causality-aware Enhanced Model for Multi-hop Question Answering over Knowledge Graphs

Yuan Sui [a], Shanshan Feng [a,*], Huaxiang Zhang [a,b], Jian Cao [c], Liang Hu [d,e], Nengjun Zhu [f]

[a] School of Information Science and Engineering, Shandong Normal University, Jinan, China
[b] Shandong JiaoTong University, Jinan, China
[c] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China
[d] School of Electronics and Information Engineering, Tongji University, DeepBlue Academy of Sciences, Shanghai, China
[e] DeepBlue Academy of Sciences, Shanghai, China
[f] School of Computer Engineering and Science, Shanghai University, Shanghai, China

ARTICLE INFO

*Article history:*
Received 29 December 2021
Received in revised form 26 April 2022
Accepted 27 April 2022
Available online 5 May 2022

*Keywords:*
Knowledge graph-based question answering
Causal representation learning
Constraint pairwise-based clustering
Knowledge graph embedding
Confounding bias

ABSTRACT

To improve the performance of knowledge graph-based question answering system (KGQA), several approaches have been developed to construct a semantic parser based on entity linking, relation identification and logical/numerical structure identification. However, existing methods arrive at answers only by maximizing the data likelihood only on the sparse or imbalanced explicit relations, ignoring the potentially large number of latent relations. It makes KGQA suffer from a high level of spurious entity relations and missing link challenge. In this paper, we propose a causal filter (CF) model for KGQA (CF-KGQA), which performs causal interference on the relation representation space to reduce the spurious relation representation in a data-driven manner, *i.e.*, the goal of this work is to comprehensively discover disentangled latent factors to alleviate the spurious correlation problem in KGQA. The model comprises a causal pairwise aggregator ($A_P$) and a disentangled latent factor aggregator ($A_C$). The former filters out most spurious entity relations inconsistent to their dense groups' neighborhood, and generates a causal pairwise matrix among all the candidate relations. The latter learns the latent relation representation via an encoder–decoder on the causal pairwise matrix. It disconnects the latent factor and the causal confounder beneath the knowledge embedding space by causal intervention. To prove the effectiveness and efficiency of the proposed approach, we test CF-KGQA and other state-of-the-art methods on four public real-world datasets. The experiments indicate that our approach outperforms the recent methods and is also less sensitive to the spurious correlation problem, thus demonstrating the robustness of CF-KGQA.



## 1. Introduction

Due to the increasing size of data, real-world knowledge graphs often contain millions or even billions of facts, and their large volume and complexity have hindered the access of regular users [1]. Knowledge graph-based question answering (KGQA) is proposed to bridge this gap between users' demands and a complex knowledge graph. With KGQA, users' natural language questions can be automatically translated into structured queries such as SPARQL, and answers are returned as entities or predicates [2–4]. For example, given the question "Where was Walter Chrysler born?", KGQA identifies its corresponding fact, *i.e.*, (Walter Chrysler, place_of_live, Wamego_US). With applications ranging from search engine design to AI-based conversational agents, KGQA enables artificial intelligence systems to incorporate knowledge graphs as a key building block to answer human questions, as evidenced by a large number of exciting works in recent years [5–11].

Generally, KGQA can be decomposed into a semantic parsing problem [12], *i.e.*, translating a natural language question (NLQ) $q$ into an executable representation $f$. The correct $f$ should satisfy the following conditions: (a) it can accurately capture the meaning of $q$; and (b) the execution of $f$ on KG yields the correct answer to $q$ [13,14]. The solution of such a semantic parsing problem requires linking entities, detecting predicates and identifying logical/numerical operators. Fig. 1 illustrates a general protocol of the semantic parsing and KGQA process. The topic entity and the predicate involved in the question are first detected using entity linking and predicate detection, then the candidate entity with the highest confidence in the KG is returned as the answer to the question.

* Corresponding author.
*E-mail addresses:* yuansui08@gmail.com (Y. Sui), fswl6869@foxmail.com (S. Feng), huaxzhang@163.com (H. Zhang), cao-jian@sjtu.edu.cn (J. Cao), rainmilk@gmail.com (L. Hu), zhu_nj@shu.edu.cn (N. Zhu).

https://doi.org/10.1016/j.knosys.2022.108943
0950-7051/

Based on the key steps, we discuss the challenges in terms of the knowledge expression considering the KGQA's exclusive content and structure.

- **Predicate expression variety**. Predicate expressions often differ from predicate names in many ways in natural language questions. For instance, the predicate *person. nationality* can be expressed as "What is ... 's", "which country is ... from", *etc.*
- **Entity name ambiguity**. Ambiguous entity names and partial names causes difficulties in finding the correct entity, especially when the number of candidates is large. Moreover, many entities share the same names, and partial names could be used in end users' utterances as well.
- **Grounding difficulty**. Entity grounding in large search space may cause the candidate triplet search to be expensive and even invalid in terms of obtaining executable logic forms, as hundreds or even thousands of relations could exist between one entity and other entities in the knowledge graph.
- **Missing link**. Some answers may be incorrect due to a missing link between the entities (often occurs in a sparse knowledge graph). It blocks the generation of the extracted question sub-graph and interferes the predicting process. If the knowledge graph is too sparse, the QA tasks cannot be undertaken.

Regarding the above challenges, it would be hard for KGQA to produce reasonable answers for a wide range of questions. Most of these recent approaches [15–19] obtain executable representation $f$ by maximizing the statistical likelihood only on the sparse already-existing-facts in the knowledge graph, ignoring a large number of latent facts. This yields answers far from the truth for human understanding, especially for complex questions. In recent years, the latent-factor-based approaches are considered to be potential to alleviate the difficulty by discovering more latent relations for knowledge completion [20,21]. Unfortunately, when building the syntax trees, there are some expressions with the same semantics but different syntactic structure, and some conditional predicates are inconsistent with the entities. This will further produce a large number of correlations describing pseudo-facts, leading to more spurious information in the knowledge representation space.

In fact, the structure hidden in knowledge embedding space should be fully investigated for modeling the explicit and implicit relations of KGs. According to Fig. 1 (upper diagram), given an ideal knowledge graph, the reasoning link is "United States-House Committee-Judiciary department-Member-Jerry Nadler", and the answer is found through a 3-hop reasoning propagation. However, as shown in Fig. 1 (lower diagram), a real-life knowledge graph is always grounded with spurious relation issues such as missing links and entity name ambiguity. Due to the missing link from "House Committee" to "Judiciary department", the reasoning link has changed to "United States-Congressional Committee-Judiciary department-Member-Edward Thomas", which results in a wrong answer. However, the real-life evidence (or, commonsense) implies that there is a subtle correlation between Congressional Committee and House Committee, so the reasoning link "United States-Congressional Committee" should be removed.

Nevertheless, how to identify those truly meaningful latent relations is definitely not a trivial issue. Intuitively, the relations that conform the commonsense contain clear semantic meaning. In a good knowledge representation space, their embeddings tend to be formed as clusters that can be described by some certain super-ordinate concepts, while those irrational relations may deviate from these clusters. In practice, we may first explicitly indicate all the possible latent relations by syntax tree,

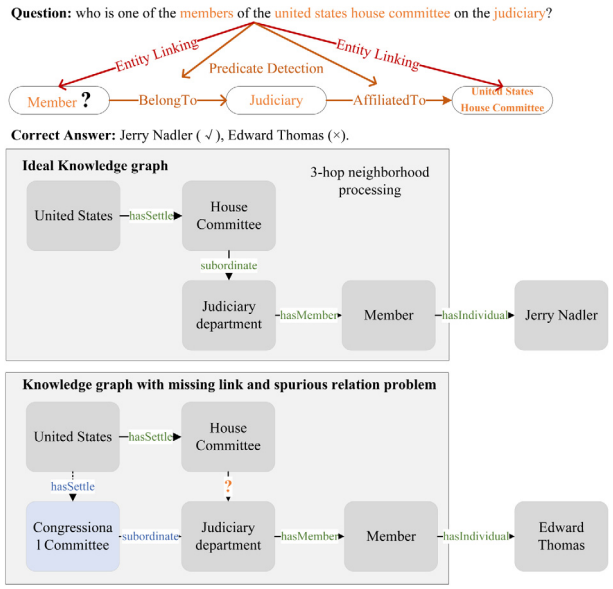


**Fig. 1.** Multi-hop KGQA examples and its key components. We first identify the topic entity using entity linking and then detect the predicate asked by the question using predicate detecting. According to the upper ideal knowledge graph, the reasoning link has 3-hop neighborhood processing. However, the real multi-hop KGQA still suffers from the problem of missing links and spurious relations, *i.e.*, United States has settle House Committee and Congressional Committee, due to the missing link from House Committee to Judiciary department, the reasoning link will be changed which results in the wrong answer.

and then cluster them to obtain the latent factors and filter out most of the irrational relations. The process is as shown at the bottom of Fig. 1. However, the relation clusters may still include some spurious but seems vital relations that distract the answer prediction process. Introducing external causal knowledge may somehow rectify part of these spurious relations, at the cost of tedious construction of domain-specific causal triplet [22] and low coverage on the KG. To achieve more generality, one still needs to go back into the relation representation space, to find a way to reliably identify those truly informative latent relations.

In this paper, we propose a causal interference-based module called causal filter (CF) for KGQA. It performs causal interference on the relation representation space to improve the quality of KG relations and QA accuracy, by squeezing out those spurious relations in a data-driven manner. Specifically, it is composed of causal pairwise aggregator ($A_P$) and disentangled latent factor aggregator ($A_C$). $A_P$ filters out most spurious relations that are inconsistent to their dense groups' neighborhood, and generates a causal pairwise matrix among all the candidate relations. $A_C$ learns the representation of the latent relations via an encoder–decoder on the causal pairwise matrix from $A_P$. It disconnects the latent factor and causal confounder beneath the knowledge embedding space by causal intervention.

Compared with the recent methods, CF-KGQA scales up the statistics likelihood to jointly model the statistics likelihood and latent causal correlations, and the aforementioned challenges can be well addressed by our method. First, the causal-enhanced CF make it easier to filter out the spurious relations brought by linguistic diversity. The variety of predicate expression and the ambiguity of entity name can be alleviated, even when the number of candidates is large. For the problem of grounding difficulty, we have redesigned the scoring function to quickly locate the most likely answers. It utilizes the causal information as judgments without having to explore all the possible forms of a complex question. Second, we adopt the embedding mechanism

of the head entity and predicate, and add the causal correlations derived from the CF. It envelops all the observed and unobserved links surrounding each node. Unlike other QA systems, *e.g.*, PullNet [23] and GraftNet [24], even if there is no path between the head entity and answer entity, our model is capable to answer the question as the KG contains enough information for determining the path.

The key contributions of our work are as follows:

- We develop a novel causal inference method CF using clustering methods. It filters out most spurious relations inconsistent to their dense groups' neighborhood, and disconnects the latent factor and causal confounder beneath the knowledge embedding space by causal intervention. The proposed model can well tackle the spurious entity relations and missing link challenges in KGQA.
- We design a contextual framework on the basis of causal relevance for triplet extraction. It learns to represent the entity and predicate of the triplet over embedding knowledge graphs and disentangled latent factors detected through the CF module.
- Our method functions in a data-driven manner rather than being process driven. The model can obtain finer grained latent relations via clustering algorithms and is capable of better utilizing the KG sources.
- We devise a new mechanism for the KGQA task on the basis of their causal relevance probabilities produced by CF, and experiments on several real-world benchmarks demonstrate the effectiveness and robustness of our method.

The remainder of this article is organized as follows. The related work is briefly summarized in Section 2. Then, the core related technique of CF is detailed in Section 3. In Section 4, we illustrate the framework of our CF-based KGQA model and how to integrate the causal filter module to KGQA task. In Section 5, we provide the experimental results and discuss related topics to support our proposed model. Finally, we conclude and provide the directions for future work in Section 6.

## 2. Related work

In this section, we discuss some foundation works in the field, describing studies on commonsense learning, latent factor based learning and causality in QA. Also, we detail some of the approaches that are most widely adopted in KGQA.

### 2.1. Commonsense learning

When humans communicate with each other, they often rely on broad implicit assumptions. Humans learn and use these kinds of assumptions in everyday life, making their language concise without lacking precision. However, machines by nature do not have such background knowledge. We find AI systems currently are good at expressing "what" (*e.g.*, classification, segmentation) and "where" (*e.g.*, detection, tracking), but are not adept at knowing "why" [25], *e.g.*, why open umbrellas when raining? which requires high-level, commonsense reasons, such as "umbrellas can prevent humans from getting wet". In real situations, it is common that machines make cognitive errors due to a lack of common sense.

Commonsense reasoning (learning) is a long-standing challenge for deep learning, and popular and effective commonsense learning approaches follow the strategy of applying background knowledge (corresponds to reasonable common sense) on specific tasks. These approaches can be grouped into two: (1) data-central methods assume that knowledge can be learned directly from large corpora in an unsupervised way. These methods use pre-trained language models as an unified knowledge encoder to solve the question answering task [26,27]. Although they continue to make groundbreaking advances in relation to various NLP tasks by training with larger corpora and using more advanced hardware support and tuning techniques, some works argue that they are good at memorizing facts and finding "false statistical cues" as a result of overfitting. However, this kind of language model is unable to reasoning the process of model updation, which limits its potential to solve NLP understanding tasks; and (2) model-central methods still rely on powerful pre-trained models to learn language semantics [28]. The difference is that they extend the underlying pre-trained models with structures which are designed to explicitly incorporate knowledge that is helpful for specific tasks. However, the model-central methods face the main challenge, which is how to extract relevant evidence and how to perform reasoning over the extracted evidence.

When commonsense knowledge outside the given text is needed to answer the question is called commonsense question answering. Therefore, the main focus of the commonsense learning in QA is how to incorporate commonsense knowledge and conduct reasoning. Recently, some commonsense question answering methods have been proposed with quite promising outcomes [29,30]. For example, [30] proposed an unsupervised framework on self-talk as a novel alternative to multiple-choice commonsense tasks and queries language models with a number of information seeking questions such as "What is the definition of ..." to discover additional background knowledge. The work in [29] utilized a novel system for selecting grounded multi-hop relational commonsense information from ConceptNet via a pointwise mutual information and term-frequency based scoring function and the extracted commonsense information is used to fill the gaps in reasoning between context hops, using a selectively gated attention mechanism which boosts the QA model's performance significantly. However, these works still limited to outer knowledge, and cannot identify the latent causal correlation beneath the data automatically.

### 2.2. Latent factor-based learning

Most previous KGQA methods are required to specify the factors that are considered responsible for the query's candidate on a particular class of questions. However, it may causes the algorithms being generalizable and inadequate, *i.e.*, the factors responsible for the choice of footwear are different from those responsible for selecting movies. The latent factor based model is generalized under the assumption that the factors responsible for a query's candidate do not have to be explicitly stated. A query $q$ can be analyzed by its affinity towards these latent factors, and the underlying questions can be identified by the latent factor corresponding to it.

Latent factor-based learning has been studied in combination with deep learning [20,21,31,32], and has been explored in several directions, including disentangled lexicon induction [22,33], explainability of neural NLP [34,35], text classifying [36] and text matching [21]. For example, [20] proposed a LEGO framework to utilize a query synthesizer and a latent space executer to execute the reasoning action in the latent embedding space to combat the missing information in the knowledge graph. In addition, [21] developed a method based on text matching to discover the latent information for matching both the topical content of confounding documents and the probability that each of these documents is treated. However, these methods are still generally unable to support KGQA tasks well because the latent factors learned from the syntax tree and the spatial–temporal co-occurrence are still noisy, and the influence of these latent factors hidden in the knowledge embedding space has not been fully investigated.

Inspired by the description above, we propose CF-KGQA to comprehensively discover the disentangled latent factors and causal confounders based on the data-driven clustering algorithm to alleviate the aforementioned drawbacks. It can better purify the latent factors and better investigate their influence.

### 2.3. Causality for question answering

Causality can link two phrases representing a cause and its effect which is viewed as a semantic relation that exists between different parts of a sentence because causality provides a new dimension of thinking NLP methods and it plays an major role in understanding the meaning of natural language text. Several usages of causal relations have been used in different modules of QA systems [37–39] and helping to find candidate answers to a question asked on QA systems. For example, [39] proposed the use of causality on corpus preparation which extracted sentences containing cue phrases from the text document and identified cause and effect parts from the sentence. For example, consider the sentence "The sun rises in the east" is the cause of "the earth rotates around its axis toward the east". The sentence helps to form a question and its answer. It has been stated that cause part is regarded as an expected answer to a question which can be automatically extracted from the effect part. Thus, the cause part "the earth rotates around its axis toward the east" serves as an answer to the question formulated from the effect part "Why does the sun rises in the east?"

However, this kind of causality usage for QA is still trapped in the stage of statistics correlation and suffers from the spurious correlation problem. In CF-KGQA, we inherit the core idea of there causality usage and propose a more general causality model based jointly on statistics and causal correlation to address the spurious correlation problem.

### 2.4. Knowledge graph based question answering

Question answering through knowledge graph has grown in popularity as a natural approach to search structured data sources, with considerable advancements being made over time. We note that the previous methods fall into three types: (i) text-matching based; (ii) semantic parsing based; and (iii) seq2seq.

The text-matching-based method is the most common method, involving multiple components in solving the different sub-tasks of KGQA, such as entity detection [5,6], entity linking [2–4], relation prediction [7,18,19,40] and evidence integration [15]. With the development of cross domain knowledge, several information retrieval style solutions have been studied to assist these sub-tasks, *e.g.*, [16] adopted a two-phase approach, candidate generation and candidate re-ranking to answer questions; [14] improved subgraph selection through a ranking method and leveraging the subject-relation dependency by a joint scoring CNN model. In addition to the retrieval style, some transition style methods have also been proposed. [10] developed an unrestricted-hop framework to relax the restriction (the number of hops in QA is generally restricted to two or three) using a transition-based search framework instead of a relation-chain-based search one.

Furthermore, semantic parsing-based methods focus on the structure of the semantic parser (relationships between entities and directions of their relations). In semantic parsing, a question is converted into query graphs and used to generate logical queries for further processing [8,9]. For example, [41] encoded such a complex query structure to be represented as an uniform vector sequence which effectively capture the interactions between individual semantic components and correct answers. Also, [42] proposed to encode the graph structure of the semantic parse using gated graph neural networks.

Additionally, the seq2seq based methods are proposed due to the existing semantic parsing approaches in KGQA which mainly focus on relations rather than the association among relations [1,11]. [43] proposed an attention seq2seq-based semantic parsing approach to improve the performance of KBQA by converting the identification problem of question types to a machine translation problem.

However, these 3-fold models either limited the common sense to human-annotated knowledge or, essentially, learn from statistical correlation rather than causality. So, as they are still trapped in the spurious correlation problem, they do not work well for selecting reasonable answers to individual queries. Therefore, we propose to perform causal interference on the relation representation space in a data-driven manner to address this gap. CF-KGQA uses two core components $A_P$ and $A_C$. The former component captures the causal pairwise matrix among the candidate relations by removing the most spurious relations that are inconsistent with common sense, and disconnect the latent factor and causal confounder beneath the knowledge embedding space for intervention. The latter component learns the latent relation representation via an encoder–decoder on the causal pairwise matrix. Compared with the recent methods, CF-KGQA scales up the statistics likelihood to jointly model the of statistics likelihood and latent causal correlations, and address the problem of spurious correlation.

## 3. The Causal Filter (CF)

This section details the core technical contributions of CF-KGQA and its implementation.

### 3.1. Problem formulation

We formulate the problem based on the causal structure and use the language of structural causal models (SCMs) for the basic statement. The formulation of the model is illustrated in Fig. 2. SCMs consider a set of observables (or variables) $X_1, X_2, \ldots, X_n$ with a directed acyclic graph (DAG) and assume that each observable is the result of an assignment: $X_i = f_i(PA_i, U_i)$, $(i = 1, \ldots, n)$ using a deterministic function $f_i$ depending on $X_i$'s parents in the graph (denoted by $PA_i$) and on an unexplained random variable $U_i$. The unexplained random variable (actually noisy variable) $U_i$ ensures that the overall object can represent a general conditional distribution $P(X_i|PA_i)$, and the set of noise $U_1, \ldots, U_n$ is assumed to be independent. Variables in a causal graph may be unobserved, making the causal inference particularly challenging.

In KGQA, the causal structure needs to be modified. As presented in this article, the goal of this work is to comprehensively discover disentangled latent factors to alleviate the spurious correlation and integrate these disentangled latent factors with the KGQA model. Specifically, as shown in Fig. 2(a), the influence from $Z$ on $W$ is denoted as $s(e_h, p_l, z_i)$ which scores the latent relations' influences on head entity/predicate representation. However, the unobserved confounder $C$ causes both the latent relation and the answer, and eventually creates a spurious correlation. Figure (b) and (c) illustrate the causal models by blocking the backdoor path from $C$ to $Z$, or by intervening on a small set of $Z$. The disentangling process is denoted as $s(e_h, p_l, g(z_i))$ which scores the causal influences, *i.e.*, deliberately forces $e_h$ and $p_l$ to incorporate every $z_i$ fairly, where $g(\cdot)$ denotes the disentangling process on latent relations.

The fundamental problem we intend to solve is modeling a function for disentangling the latent relations $g(z_i)$. Here, we formally define the casual confounder, disentangled latent factors and some basic definitions to model the entity-predicate causal interactions based on clustering algorithms.

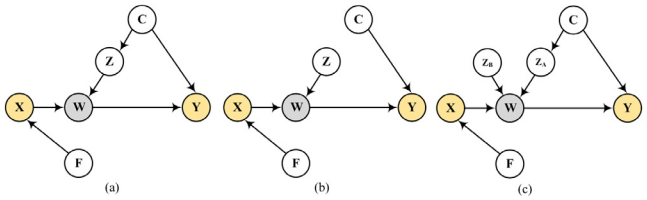


**Fig. 2.** Causal graph for KGQA. The yellow variables are observed. (a) $F$ is the variable that generates the head entity and predicate representation illustrated as $W$ by entity linking and predicate detecting. The unobserved confounder $C$ causes both the latent relation $Z$ and answer $Y$, which creates a spurious correlation between the question $X$ and answer $Y$. (b) An ideal intervention blocks the backdoor path from $C$ to $Z$, which produces causal models. (c) In practice, we cannot guarantee to intervene on all $Z$ variables. However, by properly intervening on even a small set of nuisance factors $Z_i$, the confounding bias is addressed.

**Definition 1.** Latent relations: To implement the theoretical and imaginative intervention, we define the latent relations $\mathcal{Z}$ among the KG as the basic form of the causality objects which can be calculated easily. Note that the acquisition of latent relations is based on the knowledge graph embedding (KGE) algorithms. For example, $r_{\text{latent}} = h - t$ for TransE.

**Definition 2.** Informative objects: The object in a dense cluster $\mathcal{N} = \{\mathcal{N}_1, \mathcal{N}_2 \ldots \mathcal{N}_K\}$ is defined as the most informative object if it has the highest neighborhood density in the dense group. The objects closer to the most informative object in the same dense group are called informative objects.

**Definition 3.** Causal confounder: some unobserved variables cause a spurious correlation from question to answer. It should be noted that not only objects in the KG space are confounders, some confounders cannot be observed, *e.g.*, color, attributes, and nuanced scene contexts. In this paper, we propose a novel implementation to disentangle the latent relations from the causal confounders are not necessarily restricted to concrete objects.

**Definition 4.** Disentangled latent factors: The disentangled latent factor $\mathcal{Z}' = [z'_1, \ldots, z'_N]$, disentangled cluster of the latent relations $\mathcal{Z} = [z_1, \ldots, z_M]$, in the $N \times d$ matrix for practical use, where $N$ is the category size determined by the output of the clustering algorithm, $M$ is the number of latent relations we utilize, $d$ is the feature dimension of the disentangled latent factors. $\mathcal{Z}'$ is composed of $g(z_1, \ldots, z_M)$, where $g(\cdot)$ refers to the CF process and $M$ is the size of the latent relations utilized.

### 3.2. Model architecture

The architecture of CF consists of four components: Latent Relation Representer $R_L$, Causal Constraint Pairwise Aggregator $A_P$, Disentangled Latent Factor Aggregator $A_C$ and Grounding Fact Scorer $S_G$:

- Latent Relation Representer $R_L$: obtains the latent relations representation based on knowledge graph embedding algorithms according to Definition 1.
- Causal Pairwise Aggregator $A_P$: generates a causal pairwise matrix among all the candidate relations, *i.e.*, Causal Link (CF), Statistics Link (SL), Non-Link (Non-link), which can remove the most spurious confounder beneath the knowledge embedding space for intervention using both the explicit and implicit correlation.
- Disentangled Latent Factor Aggregator $A_C$: learns the latent relation factor representation via the encoder–decoder on the causal pairwise matrix from $A_P$.

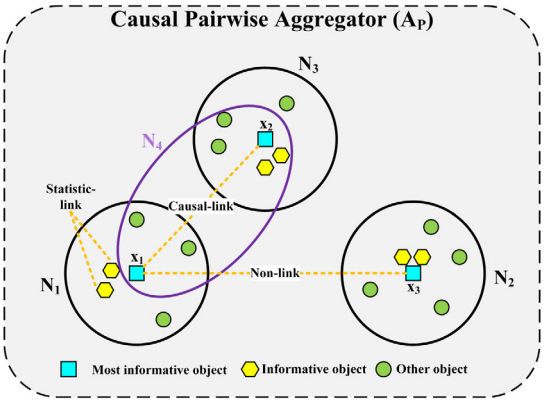


**Fig. 3.** Illustration of the different links. $x_1, x_2, x_3$ are the most informative objects of dense group $N_1, N_2, N_3$, which are the candidates for the causal-link and non-link constraints. The informative objects shown by the yellow hexagons are candidates for the statistic-link constraints in the dense group. For $x_1, x_2, x_3$, we compute the distinctness and concentration distance for each object pair and rearrange them in ascending order of their score. The pair of objects with smaller scores which can form a new dense group $N_4$ is denoted as a causal-link, while the other pairs are denoted as a non-link. In this case, $x_1, x_2$ denotes a causal-link, while $x_1, x_3$ denotes a non-link.

- Grounding Fact Scorer $S_G$: learns to score whether a fact should be candidate from the causal level and statistics level which is detailed in Section 3.3.3, Eq. (9).

This architecture is implemented through a mixed neural network model and different components have deviated functions. Moreover, our model is type of a propagation model, *i.e.*, the output of $A_P$ will be the input of $A_C$.

### 3.3. Clustering based causal filter model

In this section, we present the details of each components in CF.

#### 3.3.1. Causal constraint pairwise aggregator $A_P$

We design the causal constraint pairwise aggregator based on an unsupervised clustering algorithm for two purposes. First, due to the lack of a ground truth of the latent factors and a cluster of the latent relations embedded in the knowledge embedding space, we build a function to classify the unlabeled factors to purify them to a certain extent. Second, we identify the pairwise constraints for the next stage, *i.e.*, whether there is causal link between two entities or only a statistic link among the knowledge embedding space.

To do this, $A_P$ follows a two-stage process. First, it adopts the K-DBSCAN [44] as the initial clustering and generates dense groups from the initial clusters based on distinctness and concentration to purify them to a certain extent. Second, it selects causal-link and statistic-link constraints based on the local density estimation. The key idea of our $A_P$ architecture is illustrated in Fig. 3.

*Generating dense groups from initial clusters.* We develop a mechanism to clarify the outliers of the initial cluster and obtain the dense clusters based on the concepts of distinctness and concentration distances. $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_K$ is the set of the initial clusters generated from an unlabeled latent relation representations by K-DBSCAN and $n_1, n_2, \ldots, n_K$ are the number of objects in these clusters. Given an object $x_o \in \mathcal{G}_c$ for $1 \leq c \leq K$, its neighboring objects are denoted as a set of objects in the initial cluster excluding itself. We assume that $\mathcal{G}'_c$ is a set of $n'_c$ objects neighboring $x_o$.

The distinctness distance of $x_o$ to $\mathcal{G}_c'$ is defined as the average of distances from $x_o$ to all objects in $\mathcal{G}_c'$, given as:

$$Dd_{x_o} = \frac{1}{n_c'} \sum_{x_i \in \mathcal{G}_c'} d\left(x_i, x_o\right) \tag{1}$$

Similarly, the concentration distance of $x_o$ to $\mathcal{G}_c'$ is denoted as the average of the mutual distances among the objects in $\mathcal{G}_c'$ as:

$$Cd_{x_o} = \frac{1}{n_c'\left(n_c' - 1\right)} \sum_{x_i, x_j \in \mathcal{G}_c', i \neq j} d\left(x_i, x_j\right), \tag{2}$$

Based on $Sd_{x_o}$ and $Cd_{x_o}$ defined, the distinctness of $x_o$ from $\mathcal{G}_c$ can be measured by the ratio of the distinctness and the concentration distance, which is named the group outlier criterion (GOC):

$$GOC_{x_o} = \frac{Dd_{x_o}}{Cd_{x_o}} = \frac{1}{n_C' - 1} \frac{\sum_{x_i \in \mathcal{G}_c'} d\left(x_i, x_o\right)}{\sum_{x_i, x_j \in \mathcal{G}_c', i \neq j} d\left(x_i, x_j\right)} \tag{3}$$

For an initial cluster $\mathcal{G}_c$, $GOC_{x_o}$ is scored by measuring how far away object $x_o$ is from the dense group of objects in $\mathcal{G}_c'$. All objects in $\mathcal{G}_c$ have the *GOC* scores computed. Then, we arrange them in ascending order of the *GOC* scores. The objects with smaller GOC scores in $\mathcal{G}_c$ are denoted as dense groups. According to the distribution of the *GOC* scores in $\mathcal{G}_c$, a threshold is set to select objects into a dense group. For all initial clusters, this process yields a set of dense groups of objects as candidates for pairwise constraints.

*Selecting causal-link and statistic-link constraints.* For a set of dense groups, we derive the causal-link and statistic-link pairwise constraints from the dense groups using a local density estimation method based on semi-supervised clustering. It is assumed that $\mathcal{N}_1, \mathcal{N}_2 \ldots \mathcal{N}_K$ represents the set of dense groups of objects that remain after removing the separate objects from the initial clusters. These density groups serve to derive the pairwise link constraints: the causal-link, the statistic-link, and the non-link pairwise constraints. In the semi-supervised clustering result, on the one hand, the objects classified with the statistic-link should be selected from the same dense group, while on the other hand, the objects classified with the causal-link or non-link should be selected from two separate dense groups and classified into different clusters.

$\mathcal{N}_c = \left\{x_1, x_2, \ldots, x_{L_c}\right\}$ is a dense group with $L_c$ objects in a high density neighborhood $\mathcal{N}_c$. We calculate the mutual distances between the $L_c$ objects using Euclidean distance to generate a non-negative and symmetric matrix $D = \left(d_{ij}\right)$. Then we compute the local density for each object $x_i$ in the neighborhood $\mathcal{N}_c$ as:

$$LD\left(x_i\right) = \frac{1}{|\mathcal{N}_c|} \sum_{j \in N_c} d_{ij} \tag{4}$$

where $x_i$ is the $i$th object in $\mathcal{N}_c$ and $LD\left(x_i, \mathcal{N}_c\right)$ is the local density of object $x_i$ relative to other objects in $\mathcal{N}_c$. In this way, we calculate the local densities for all objects in $\mathcal{N}_c$. A large value of $LD\left(x_i\right)$ indicates the high local density of the object $x_i$. Based on the objects' local densities, we can rank them in $\mathcal{N}_c$ and according to the assumption that objects with a large *LD* are more informative, we identify the most informative object from each dense group.

For a dense group of objects, we calculate the local densities of all the objects and rank them according to their densities. The first object with the highest local density is regraded as the most informative object which is used to generate causal-link and non-link pairwise constraints. In addition, according to the ranking list of informative objects, the second and third objects with the highest local densities are selected as the first pair of statistic link pairwise constraints. Similarly, the next pair of informative objects is derived as the second statistic-link pairwise constraint. The process of dense group selection is continuous until the required number of pairwise statistical-link constraints are obtained. We use the same process to obtain the set of statistic-link pairwise constraints from all dense groups.

Then, we formulate the causal-link and non-link pairwise constraints set. For each dense group, we obtain the most informative object, which is the candidate for the causal-link and non-link pairwise constraint. Given the $K$ most informative objects from $K$ dense groups, we compute the *GOC* for each pair object according to Eq. (3) and arrange the objects in ascending order of their *GOC*. The pair of objects with smaller *GOC* scores is denoted as causal-link so we assume there is a latent causal relation between this pair, while the other pairs are denoted as non-link.

For $K$ objects from different dense groups, any pair of the causal-link and non-link objects should occur in different clusters, *i.e.*, the candidate objects should be obtained from two different dense groups. Furthermore, any pair of statistic-link objects should be obtained from the same dense group. We define a matrix describing the pairwise constraints CL and SL as:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{bmatrix} \tag{5}$$

As $x_i$ and $x_k$ are assigned to be causal-link pair, $a_i k = 1$. If $x_i$ and $x_k$ satisfy the statistic-link constraints, $a_i k = -1$. Other entities in

**Algorithm 1** Causal Constraint Pairwise Aggregator $A_P$: unsupervised deep embedded clustering

**Require:** A high-dimensional unlabeled latent relation embeddings $X = x_1, x_2, \cdots, x_n$.
**Ensure:** Call DBSCAN algorithm to produce initial cluster assignments $y_{i_{i=1}}^n$;
**Step 1** → Discover dense groups
**for** iter1 = 1 to N **do**
  **for** iter2 = 1 to M **do**
    Compute $GOC_{x_o}$ in initial cluster $y_{i_{i=1}}^n$ by Eq. (3).
    **if** $GOC_{x_o} > threshold1$ **then**
      $y_{i_{i=1}}^n \leftarrow y_{i_{i=1}}^n - x_o$
    **end if**
    $\mathcal{N}_c \leftarrow y_{i_{i=1}}^n$
  **end for**
**end for**
**Step 2** → Select Statistic-link, Causal-link and Non-link pairwise constraints
**for** iter = 1 to K **do**
  Compute pre-distance matrix $D$ of $L_c$ objects in dense group $\mathcal{N}_c$ with Euclidean distance function.
  Compute local density $LD(x_i)$ for each object $x_i$ in the neighborhood $\mathcal{N}_c$ by Eq. (4).
  Sort the indices of $LD(x_i)$ with descending order.
  **for** i = 1 to $L_c$ - 1 **do**
    **if** $i \neq 1$ **then**
      Add the pair of $(x_i, x_{i+1})$ into the Statistic-link set (SL).
    **else**
      Compute $GOC_{x_o}$ for each pair objects again and re-arrange the objects in ascending order.
      **if** $GOC_{x_o} < threshold2$ **then**
        Add the pair of $(x_i, x_{i+1})$ into the Causal-link set (CL).
      **else**
        Add the pair of $(x_i, x_{i+1})$ into the Non-link set.
      **end if**
    **end if**
  **end for**
**end for**
Return the pairwise constraints: Causal-link (CL), Statistic-link (SL), and Non-link.

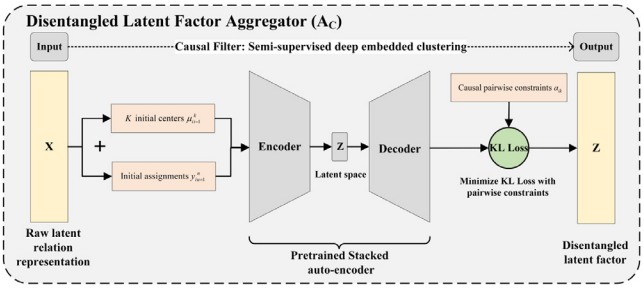


**Fig. 4.** The architecture of $A_C$. We use the encoder layers of a pre-trained SAE to initialize the DNN structure. The causal pairwise constraints obtained from $A_P$ are added to the embedding layer latent space representation to directly learn the feature representation where the next block denotes the soft assignment of each data point and is used to minimize the KL divergence loss.

this matrix are all set to zero. The pairwise constraint specifies if two data examples are causally linked (CL) or statistically linked (SL). The overall process of the $A_P$ is listed in algorithm 1.

### 3.3.2. Disentangled latent factor aggregator $A_C$

We design the disentangled latent factor aggregator $A_C$ using a semi-supervised clustering algorithm. It aims to use the constraints from $A_P$ to improve the learning ability and rearrange the causal clusters. We propose a pairwise constraint based semi-supervised clustering algorithm inspired by SDEC [45]. Specifically, $A_C$ makes use of the CL and SL constraints in the feature learning process such that the data samples with the causal-link should be enforced closer to each other according to their causal relevance. While samples with a statistics link are forced far away from each other in the learned feature space where the final cluster assignment is conducted. The idea of $A_C$ is illustrated in Fig. 4.

$A_C$ utilizes the same students's t-distribution similar to SDEC [45] to measure the similarity between embedded point $z_i$ and center $\mu_j$ as:

$$m_{ij} = \frac{\left(1 + \left\|\omega_i - \mu_j\right\|^2\right)^{-1}}{\sum_{j'}\left(1 + \left\|\omega_i - \mu_{j'}\right\|^2\right)^{-1}}, \tag{6}$$

where $\omega_i = f_\theta(x_i) \in \omega$ corresponds to $x_i \in X$ after embedding, $\mu_j$ is the center of the $j$th cluster in the embedded space. $m_{ij}$ is the probability of assigning data point $i$ to cluster $j$, and $m_i = [m_{i1}, m_{i2}, \ldots, m_{ik}]^T$ is a soft assignment of data point $i$. In each step, $A_C$ matches the soft assignment $M$ to an auxiliary target distribution $N$ as:

$$n_{ij} = \frac{m_{ij}^2/f_j}{\sum_{j'} m_{ij}^2/f_{j'}}, \tag{7}$$

where $f_i = \sum_i m_{ij}$. In $A_C$, KL divergence between the soft assignment $M$ and the target distribution $N$ is minimized to refine the nonlinear transformation $f_\theta$, *i.e.*, the encoder layers of SAE [46]. To utilize the causal constraints obtained from the $A_P$ to lead the direction of clustering and embedding, we define the objective of $A_C$ as:

$$\begin{aligned} L &= KL(N \parallel M) + \lambda \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}\left\|\omega_i - \omega_k\right\|^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{K} n_{ij}\log\frac{n_{ij}}{m_{ij}} + \lambda\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}\left\|\omega_i - \omega_k\right\|^2, \end{aligned} \tag{8}$$

where $KL$ refers to KL divergence, $n$ is the number of data points and $\lambda$ is a trade-off parameter which is defined by the user. When $\lambda = 0$, the causal constraints influence degenerates to zero. Actually, minimizing Eq. (8) can minimize the costs of the violated constraints, thus being able to simultaneously learn feature representations and perform clustering assignments to favor the causal-link constraints. The overall idea of the $A_C$ is listed in algorithm 2 and is consistent with the schematic diagram of our model shown in Fig. 4.

**Algorithm 2** Disentangled latent factor aggregator $A_C$: semi-supervised deep embedded clustering

**Require:** : Dataset $X$; coefficient *lambda*; number of clusters $K$; pairwise constraints matrix $A$; stopping threshold *tol*%.
**Ensure:** : Cluster assignments $y_{i\,i=1}^{n}$; cluster centers $\mu_{i\,i=1}^{k}$; deep mapping $f_\theta$

**Step 1** → Initialization with SAE ▷ SAE: Stacked auto-encoder refers to [46].
Pretrain SAE and obtain $K$ initial centers $\mu_{i\,i=1}^{k}$ and cluster assignments $y_{i\,i=1}^{n}$ from $A_P$ in the latent relation space.
**Step 2** → Clustering with pairwise constraints
**for** *iter* $\in 0, 1, \ldots,$ *MAXITER* **do**
  Choose a batch of samples $S \in X$.
  **if** *iter*%$T == 0$ **then**
    $\omega_i \leftarrow f_\theta(x_i), \forall x_i \in X$.
    Compute all $m_{ij}$ values according to Eq. (6).
    Compute all $n_{ij}$ values according to Eq. (7).
    Save old assignments: $y_{oldi} \leftarrow y_i$.
    Update label assignments: $y_i \leftarrow argmax_j m_{ij}$.
    **if** $(sum_{i=1}^{n} y_{oldi} \neq y_i)/n <$ *tol*% **then**
      Stop training.
    **end if**
  **end if**
  Update $\theta$ and $\mu_{i\,i=1}^{k}$ via Eq. (8).
**end for**

### 3.3.3. Grounding fact scorer $S_G$

Each sample fact can be measured by a score function $s$ in terms of the statistics level and causal level. According to the aforementioned definitions, the joint score function $s$ can be decomposed into four scores:

$$s = \beta_1 s(e_h, \hat{e_h}) + \beta_2 s(p_l, \hat{p_l}) + \beta_3 s(e_h, p_l, \hat{e_t}) + \beta_4 s(e_h, p_l, g(z_i)) + \delta \tag{9}$$

where $s$ scores a fact which matches the learned entities and predicates the most in both causal-level and statistical-level; $s(e_h, \hat{e_h})$ scores the distance between entity representation $e_h$ and ground truth; $s(p_l, \hat{p_l})$ scores the distance between predicate representation $p_l$ and ground truth; $s(e_h, p_l, \hat{e_t})$ scores the similarity between tail entity, obtained from head entity $e_h$ and predicate $p_l$, and ground truth tail entity $\hat{e_t}$, and $s(e_h, p_l, g(z_i))$ scores the causal influences, *i.e.*, deliberately forces $e_h$ to incorporate every $z_i$ fairly, where $g(\cdot)$ denotes the disentangling process on latent relations. $\beta_1, \beta_2, \beta_3, \beta_4$ are the scale parameters for weighing these scores, and $\delta$ is used for practical optimization.

## 4. CF-based KGQA model

In this section, we detail the whole framework of our proposed model as illustrated in Fig. 5 and supplement the core implementation of how to integrate the causal filter module with the KGQA task. The knowledge graph embedding (KGE) algorithm is the primary method utilized to represent each entity and predicate. Then, we develop a novel causal inference based module that we denote as the causal filter (CF) based on the combination of $A_P$ and $A_C$ to disentangle the latent factors. In addition, we train a head entity learning model and a predicate learning model for forecasting. Lastly, a carefully tailored distance metric (adding causal scores) is used to search for the candidate entities' closest entities in the KG considering causal inference.

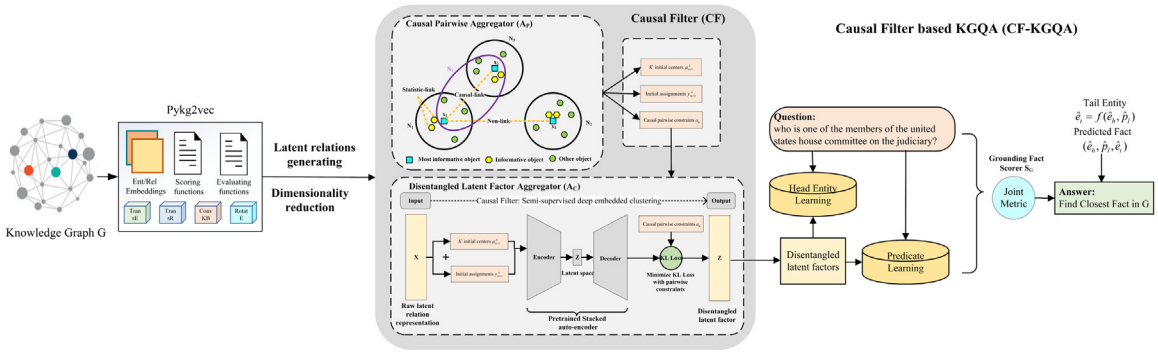


**Fig. 5.** Overview of the CF-KGQA: The primary process is to obtain the representation of each entity and predicate by KGE algorithms. Then, we develop the core block CF module to disentangle the latent factors. After this, we train the head entity learning model and predicate learning model for forecasting. Finally, searching the candidate entities on embedding spaces based on a carefully crafted distance metric (with causal score adding), the predicted entities' closest entities in KG which also correspond to the causal inference are identified as the answers.

### 4.1. Knowledge graph embedding

CF-KGQA employs the knowledge embedding representations as the external resources to guide the language model for training. We utilize an existing KG embedding algorithm TransE [47] to learn the entities collections $\mathbb{E}$ and use pykg2vec [48] as implementation.

For each triplet in KG, we denote its embedding representations as $(e_h, r_l, e_t)$, and the knowledge representation learning algorithm initializes the values randomly or based on trained word embedding models. Then, a function $f(\cdot)$ is used to define the relation $r_1$ of a triplet in triplet by $e_h, e_t$, *i.e.*, TransE [47] defines the relation as $e_t \approx e_h + r_l, r_l \approx e_t - e_h$. Finally, the KGE algorithms minimize the overall distance $e_t$ between $f(e_h, r_l)$ and for all facts and train the method on both positive and negative samples (facts and synthetic facts that do not exist in KG).

### 4.2. Predicate and head entity learning model

To represent a question, we derive a point in the entity embedding space as its head entity representation $\hat{e}_h$, and a point in the predicate embedding space as its predicate representation $\hat{p}_\ell$. A conventional solution is to learn the mapping dependent on semantic parsing and physically made dictionaries, or basically consider each kind of predicate as a label class to convert it into a classification issue. However, since the areas of the end clients' inquiries are frequently unbounded, the spurious correlation problem might occur, *i.e.*, the new inquiry's predicate may not be the same as every one in the training data. As a consequence, the traditional solutions are not able to handle this scenario. To address the gap, we apply the global relation information on KGQA and utilize the disentangled latent factors obtained from CF as compensation. Fig. 6 illustrates the architecture of the proposed predicate and head entity learning model.

Given a question with length $L_q$, we first map its $L_q$ tokens into a sequence of word embedding vectors $w_j$, for $j = 1, 2, \ldots, L_q$. Then we employ a bidirectional LSTM to learn a forward hidden state sequence $(\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_{Lq}})$ and a backward hidden state sequence $(\overleftarrow{h_1}, \overleftarrow{h_2}, \ldots, \overleftarrow{h_{Lq}})$. The specific calculation process of the LSTM hidden layer output representation could be referenced in [12]. We concatenate the forward and backward hidden state vectors and obtain $h_j = [\overrightarrow{h_j} : \overleftarrow{h_j}]$.

The attention weight of the $j$th token $a_j$ is calculated using the following formulas:

$$a_j = \frac{\exp(q(w_j, h_j))}{\sum_{Lq} \exp(q(w_i, h_i))}, \tag{10}$$
$$q(w_j, h_j) = \tanh(W_a{}^{\mathrm{T}}[w_j : h_j] + \theta)$$

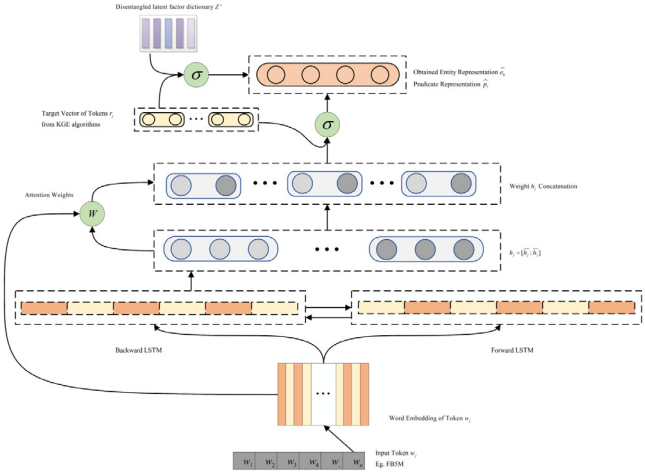


**Fig. 6.** The model for predicate and head entity learning: An attention-based bidirectional LSTM module for predicate and head entity learning with the assistance of disentangled latent factors from the CF model.

We apply the attention weight $a_j$ to $h_j$ and concatenate it with the word embedding $w_j$, resulting a hidden state $s_j = [w_j; a_j h_j]$. A fully connected layer is then applied to $s_j$, and its result $n_j \in \mathbb{R}^{d\times 1}$ is denoted as the intermediate variables. Moreover, we utilize the disentangled latent factors obtained from CF as compensation for spurious correlations, *i.e.*, we calculate the distance $d_j$ between intermediate variables $n_j$ and disentangled latent factor $z_j \in \mathcal{Z}'$ and concatenate them together, resulting another hidden state $v_j = [n_j; d_j]$. The predicted predicate representation is computed as the mean of concatenation of $s_j$ and $v_j$, that is,

$$\hat{p}_l = \frac{1}{L}\sum_{j=1}^{L}[s_j; v_j]^{\mathrm{T}} \tag{11}$$

All the weight matrices, weight vector $W_a$ and bias term $\theta$ are calculated according to the training data.

Additionally, for head entity learning, given a question, instead of inferring the head entity directly, we recover its representation in the KG embedding space. Similar to the computation of $\hat{p}_\ell$, the entity representation $\hat{e}_h$ is computed using the same neural network as shown in Fig. 6. The entity representation $\hat{e}_h$ and predicate representation $\hat{p}_\ell$ are eventually obtained.

### 4.3. Jointly searching on embedding spaces

This section is the progressive version of Section 3.3.3. For each sample question we have its predicate, head entity representation and disentangled latent factors' vector centers. Our goal is to find a fact in KG that most closely matches these learned representations and candidates. We inherit the joint distance metric defined by [12], however, we add the causal score as an additive evaluation. Mathematically, the proposed joint distance metric is defined as:

$$\begin{aligned}\underset{(h,\ell,t)\in C}{\text{minimize}} \quad \text{Eq. (9)}:&\beta_1 \left\| e_h - \hat{e}_h \right\|_2 + \beta_2 \left\| p_\ell - \hat{p}_\ell \right\|_2 \\ &+ \beta_3 \left\| f\left(e_h, p_\ell\right) - \hat{e}_t \right\|_2 + \beta_4 \frac{\left\| p_\ell - z' \right\|_2}{\left\| p_\ell - z' \right\|_2 + \left\| e_h - z' \right\|_2} \\ &- \alpha_1 \operatorname{sim}\left[n(h), \mathrm{HED}_{\mathrm{entity}}\right] - \alpha_2 \operatorname{sim}\left[n(\ell), \mathrm{HED}_{\mathrm{non}}\right]\end{aligned} \tag{12}$$

where $\hat{e}_t = f\left(\hat{e}_h, \hat{p}_\ell\right)$. Function $n(\cdot)$ returns the name of the entity or predicate. $HED_{entity}$ and $HED_{non}$ denote the tokens that are classified as the entity name and non entity name by a simple binary classification model according to TextCNN [49]. Function $sim[\cdot]$ measures the similarity of two strings. $\beta_1, \beta_2, \beta_3, \beta_4$ are the predefined weights to balance the contribution of each term. $\alpha_1, \alpha_2$ are the predefined weights for practical optimization. In this paper, we utilize the same $\beta_n$ and $\alpha_n$ according to [12] for experiments comparing.

The first three terms in Eq. (12) measure the distance between a fact $(e_h, p_l, e_t)$ and ground truth in the KG embedding spaces. We use $f(e_h, p_\ell)$ rather than $e_t$ to represent the embedding vector of the tail entity, because there might be several facts that share the same head entity and predicate but different tail entities. Thus, a single tail entity without the pair of head entity and predicate might not be capable of answering the question. We tend to select a fact which most corresponds to the causal correlation. In particular, we assume that a causal fact has to be closer to the disentangled latent factor embedding vectors. In this manner, we add the causal score to the distance metric $\frac{\|p_\ell - z'\|_2}{\|p_\ell - z'\|_2 + \|e_h - z'\|_2}$. Meanwhile, the function balance the degree of the causal effect on predicates and entities. The fact $(e_h^*, p_l^*, e_t^*)$ that minimizes the objective function is returned.

## 5. Experiment

In this section, we demonstrate the effectiveness and efficiency of CF-KGQA in the following three ways: (1) CF-KGQA has better performance than comparative methods with various evaluation metrics on four different datasets; (2) CF-KGQA is less sensitive to the missing link problem among the knowledge graph; (3) CF-KGQA is more robust and generalized than previous models on real case studies.

### 5.1. Datasets

In the experiment, there are two public knowledge graph subsets and four real-world question answering benchmarks. All the data are publicly available. Their statistics information are as shown in Table 1.

- **FB2M** and **FB5M**: Since Freebase is primarily collected and trimmed by community members, it is regarded as a reliable knowledge graph. In this paper, we employ two large subsets of Freebase, *i.e.*, FB2M and FB5M. The number of predicates and the entities are listed in Table 1 and the repeated facts have been removed. Freebase's application programming interface (API) is not available anymore. For this reason, we use a mapping tool from [12] to obtain the entities and their relations.
- **SimpleQuestions**: More than ten thousand simple questions are presented in SimpleQuestions along with corresponding facts. All these facts belong to FB2M and all questions are formulated by English speakers based on their context and the facts. It has been used for many recent KGQA methods.
- **Webqsp**: It involves complete semantic parses in SPARQL query for 4737 questions, and partial parses for 1073 questions where no valid parse can be found or where the questions themselves require a descriptive answer. We use the same train/dev/test splits as GraftNet [24].
- **MetaQA**: The dataset contains more than 400k movie-related questions in a multi-hop KGQA context. It contains 1-hop, 2-hop, and 3-hop questions. In our experiments, we considered full KG to be defined by the MetaQA KG-Full and half KG to be defined by the MetaQA KG-50.
- **OpenBookQA**: The dataset involves 5957 multiple-choice elementary level science questions, with 4957 for training, 500 for validation and 500 for testing, which assess the understanding of a small "book" containing 1326 core science facts and the application of these facts to novel situations.

Different datasets for KGQA have varying levels of open-domain capacity requirement. Webqsp, for instance, can be used to observe most of the gold test relations during training, thus some prior works on this task adopted the closed domain assumption as in general relations extraction research, whereas for data sets like SimpleQuestions, it becomes more important to support large relation sets and unseen relations. Therefore, we utilize these four benchmarks to evaluate our proposed model from different aspects. Additionally, SimpleQuestions has been rarely used since 2019, and as a result, our model utilizes Webqsp and MetaQA as alternative datasets to compare with the most recent models.

### 5.2. Comparative methods

As described in the previous section, KGQA has grown in popularity as a natural approach to search structured data sources, with considerable advancements over time. To evaluate the effectiveness of CF-KGQA, we include several state-of-the-art KGQA algorithms (sorted by the timeline below).

- **AMPCNN** [50]: It adopts a character-level convolutional neural network for matching the questions and predicates.
- **GruQA** [51]: It employs a character-level gated recurrent unit neural network to embed questions and predicates or entities into the same space.
- **StrongBaseslineQA** [52]: It converts the problem of predicate prediction into a classification problem and adopts multiple neural networks as solutions.
- **GraftNet** [24]: It uses a variant version of the graph convolution network (GCN) to execute multi-hop reasoning on heterogeneous graphs.
- **PullNet** [23]: It trains the graph retrieval module by utilizing the shortest path as supervision and complete multi-hop reasoning on the retrieved sub-graph with GraftNet.
- **KEQA** [12]: It utilizes a knowledge embedding algorithm to jointly extract the head entity, predicate from the question and tail entity representation in the KG embedding spaces. Our work is based on this framework and has been extended.
- **KEQA_noEmbed**: It is the version of [12] without knowledge embedding. It randomly generates the predicate and entity embedding representations.
- **EmbedKGQA** [53]: It conducts multi-hop reasoning by matching the pre-trained entity embeddings with question embedding obtained from RoBERTa.

**Table 1**
Statistics of knowledge graph based question answering datasets.

| | FB2M | FB5M | SimpleQuestions | OpenBookQA | Webqsp |
|---|---|---|---|---|---|
| Training | 14,174,246 | 17,872,174 | 75,910 | 4015 | 2848 |
| Validation | N.A. | N.A. | N.A. | 1302 | 250 |
| Test | N.A. | N.A. | N.A. | 640 | 1639 |
| Predicates | 6701 | 7523 | 1837 | 153 | N.A. |
| Entities | 1,963,130 | 3,988,105 | 131,681 | 10,973 | N.A. |
| Vocabulary size | 1,963,130 | 1,213,205 | 61,336 | 12,839 | N.A. |

*Note:* The statistics description of FB2M, FB5M, and SimpleQuestions is according to [12].

- **T5-11B** [54]: It introduces a unified framework that converts all text-based language problems into a text-to-text format.
- **CBR-KBQA** [55]: It employs a neuro-symbolic case-based reasoning (CBR) approach for KGQA with non-parametric memory and parametric models.
- **NSM** [56]: It utilizes a teacher–student method to conduct the multi-hop KGQA task, where the student network tries to find the correct answer for the query, and the teacher network tries to perform intermediate supervision to improve the reasoning skill of the student network.

### 5.3. Experiment settings

To compare and evaluate the performance of the KGQA methods, we use the same training, validation, and test splits as in the original paper and follow the same settings. Either FB2M and FB5M is employed as the external knowledge resource, and a knowledge embedding algorithm such as TransE or TransR is applied for knowledge representation based on Pykg2vec [48]. We use the same dimension of the knowledge embedding representations according to [12]. The initial pre-trained word embedding is based on GloVe [57] and we employ the fuzzy wuzzy algorithm [58] to measure the similarity of two strings.

The setting of the Statistical analysis and hyper parameters are as follows. It should be noted that the choices of hyper parameters $\beta_{1-4}$ and $\alpha_{1-2}$ has a very limited impact on the performance of the whole model.

- **Statistical Analysis:** The analyzed data are collected from three separate experiments and express as means $\pm$ standard deviation. We use one-sample t test to determine the level of significance, and $P$ values $< 0.05$ are considered to be significant.
- **Hyper parameters Setting:** As aforementioned in Section 4.3, $\beta_{1-4}$ and $\alpha_{1-2}$ are some hyper parameters for contribution balancing and practical optimization. In this paper, we utilize the same $\beta_1, \beta_2, \beta_3$ and $\alpha_n$ according to [12], and $\beta_4$ is set to 0.43. The dimension of the KG embedding representations d is set to be 250, and the hidden size of the model is set to be 300. The max index of the latent relations is set to be 1000. We use the grid search to determine the best hyper parameters.

### 5.4. Evaluation and discussion

We evaluate the recent state-of-the-art KGQA algorithms listed in Section 5.2 with our proposed model on SimpleQuestion and Webqsp. The results are detailed in Tables 2 and 3.

On the basis of the results in Table 2, we observe that: first, the proposed framework CF-KGQA outperforms all the baselines and obtains a 6.02% improvement compared to accuracy when SimpleQuestions is released on AMPCNN [50]. Moreover, CF-KGQA achieves 0.5% higher accuracy compared to KEQA [12]. This demonstrates that the latent relations causally intervene the head entity and the predicate. It helps narrow the gap between prediction and the ground truth, and help to select reasonable facts in a causal manner. Second, the performance of CF-KGQA decreases by 0.23% when it is applied to FB5M. This can be attributed to the same reason as [12], that all the ground truth facts belong to FB2M, and FB5M has 26.1% more facts than FB2M. Third, CF-KGQA achieves 0.48% higher accuracy compared to CF-KGQA without the embedding algorithm. This demonstrates that the separate task KGE is necessary for CF-KGQA.

**Table 2**
The performance of different methods on SimpleQuestions.

| | FB2M (Accuracy) | FB5M |
|---|---|---|
| AMPCNN [50] | 68.3% | 67.2% |
| GruQA [51] | 71.2% | N.A. |
| StrongBaselineQA [52] | 73.2% | N.A. |
| KEQA_noEmbedded | 73.1% | 72.6% |
| KEQA [12] | 73.8% | 73.6% |
| CF-KGQA_noEmbedded | 73.4% | 72.9% |
| **Ours (CF-KGQA)** | **74.4%** | **74.1%** |

*Note*: CF-KGQA_noEmbedded refer to the version of CF-KGQA without knowledge embedding. It randomly generates embedding representations.

By jointly predicting the question's predicate and head entity with the help of causal filter, the CF-KGQA achieves an accuracy of 74.3%, which is 0.502% higher than the baseline model KEQA on the SimpleQuestions dataset. Table 3 shows that CF-KGQA still achieves the state-of-the-art on the Webqsp dataset. However, on the OpenBookQA dataset, our CF-KGQA obtains 79.6% precision, which is little worse than [60] (80.2%). This suggests that our framework might be further improved using a more sophisticated model, *e.g.*, more layers of the proposed CF module; and deeper neural networks. Fortunately, CF-KGQA still outperforms the state-of-the-art on these three datasets, which confirms the effectiveness of our proposed framework in disentangling the latent factors to remove the confounding bias among the embedding search spaces.

#### 5.4.1. Parameter analysis

We now investigate the contribution of each term in the objective function of CF-KGQA. As shown in Eq. (12), there are six terms in our objective functions. To study the contribution of every single term, we validate the performance of CF-KGQA on SimpleQuestions, Webqsp and OpenBookQA as shown in Table 4.

On the basis of Table 4, there are two significant observations. First, the predicted predicate representation $\hat{p_\ell}$ has the most significant impact in our framework. Only with metric $\left\| p_\ell - \hat{p_\ell} \right\|_2$, does the framework achieve an accuracy of 72.8% on the SimpleQuestions dataset. Second, the disentangled latent factor complements $\hat{p_\ell}$ and $\hat{e_h}$ in the joint learning. The accuracy increases from 73.8% to 74.3% as causal evaluation $\frac{\|p_\ell - z'\|_2}{\|p_\ell - z'\|_2 + \|e_h - z'\|_2}$ is used which demonstrates the effectiveness of our method.

#### 5.4.2. Ablation study

Furthermore, we test whether the number of latent relations influences the accuracy of the framework. We validate the performance of CF-KGQA *w.r.t.* three groups of combinations of situations. To study the influence of the number of latent relations,

**Table 3**
The performance of the different methods on Webqsp and OpenBook.

| Model | OpenBookQA | | | | Webqsp | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1-score | Accuracy | Precision | Recall | F1-score | Accuracy |
| STAGG [59] | 68.5% | 67.5% | 67.9% | 67.3% | 65.9% | 63.3% | 65.6% | 63.9% |
| GraftNet [24] | N.A. | N.A. | N.A. | 66.4% | N.A. | N.A. | N.A. | 66.4% |
| PullNet [23] | N.A. | N.A. | N.A. | 68.1% | N.A. | N.A. | N.A. | 68.1% |
| EmbedKGQA [53] | N.A. | N.A. | N.A. | 70.8% | N.A. | N.A. | N.A. | 66.6% |
| T5-11B [54] | 68.8% | 67.6% | 68.2% | 67.5% | 62.1% | 62.6% | 62.3% | 61.5% |
| KEQA [12] | 71.8% | 73.6% | 72.7% | 71.2% | 71.8% | 73.6% | 72.7% | 69.3% |
| CBR-KBQA [55] | 73.1% | 75.1% | 74.0% | 72.9% | 73.1% | 75.1% | 74.1% | 72.0% |
| QA-GNN [60] | **80.2%** | 75.7% | 77.9% | 77.6% | 73.9% | 75.6% | 74.7% | 72.5% |
| NSM [56] | N.A. | N.A. | N.A. | 74.5% | N.A. | N.A. | N.A. | 73.9% |
| **Ours (CF-KGQA)** | 79.6% | **77.8%** | **78.7%** | **78.5%** | **75.6%** | **74.3%** | **74.9%** | **74.3%** |

**Table 4**
The performance of CF-KGQA with different objective functions on three datasets.

| | SimpleQuestions | Webqsp | OpenBookQA |
|---|---|---|---|
| $\left\|p_\ell - \hat{p}_\ell\right\|_2$ | 72.8% | 68.3% | 73.2% |
| $\left\|e_h - \hat{e}_h\right\|_2$ | 73.4% | 70.6% | 75.4% |
| $\left\|f\left(e_h, p_\ell\right) - \hat{e}_t\right\|_2$ | 73.4% | 71.9% | 75.4% |
| $\mathrm{sim}\left[n(h), \mathrm{HED}_{\mathrm{ent}}\right]$ | 73.5% | 72.3% | 75.4% |
| $\mathrm{sim}\left[n(\ell), \mathrm{HED}_{\mathrm{non}}\right]$ | 73.8% | 72.5% | 76.5% |
| $\frac{\left\|p_\ell - z'\right\|_2}{\left\|p_\ell - z'\right\|_2 + \left\|e_h - z'\right\|_2}$ | 74.4% | 74.3% | 78.5% |

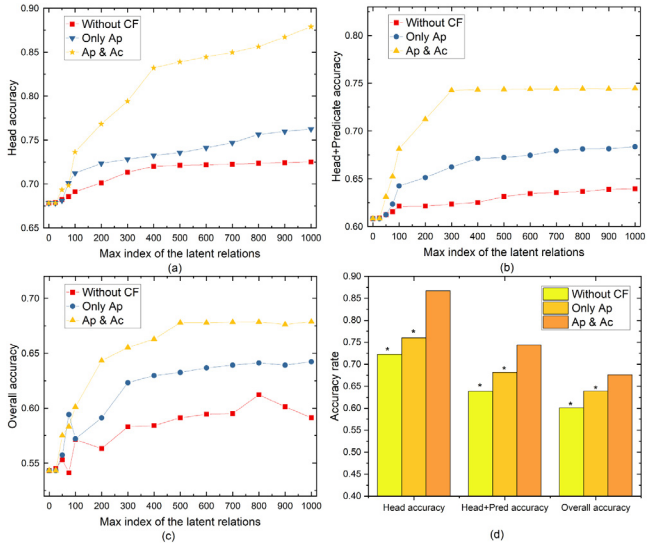


**Fig. 7.** Accuracy curves of CF-KGQA on SimpleQuestions. (a) Comparison of the accuracy of detecting the head entity in three situations, e.g., without CF, $A_P$ model only, integrated $A_P$ and $A_C$ model; (b) Comparison of the accuracy of detecting the head entity and predicate in three situations; (c) Comparison of the overall accuracy of the whole model. CF is short for causal Filter and the X-ray corresponds to how many latent relations we utilize, e.g., if the max index of the latent relations is 100, the total number of relations is 9900; (d) the statistics analysis of the three settings. $*$ refers to at the 0.05 level, the level of statistics significance is considered to be significant.

in the first group, *i.e.*, without CF, we only use the raw latent relations generated from the knowledge graph without the causal filter for purification; in the second group, *i.e.*, only $A_P$, we only use the causal pairwise aggregator $A_P$ and neglect the pairwise constraint influences; in the third group $A_P$&$A_C$, we use the integrated causal filter to disentangle the latent factors. The experiment results are illustrated in Fig. 7.

As shown in Fig. 7, we have three observations: (1) Our proposed CF module can well compensate the lack of strength of the latent factors and further explore the influences of latent factors

**Table 5**
Ablation over CF-KGQA model size on three datasets. #E = the dimension of the KG embedding representations; #H = hidden size.

| #E | #H | SimpleQuestions | Webqsp | OpenBookQA |
|---|---|---|---|---|
| 50 | 150 | 61.8% | 62.3% | 64.2% |
| 100 | 150 | 65.5% | 65.7% | 67.4% |
| 150 | 150 | 71.4% | 69.5% | 74.4% |
| 200 | 300 | 73.5% | 72.9% | 76.4% |
| **250** | **300** | **74.4%** | **74.3%** | **78.5%** |
| 300 | 300 | 74.3% | 74.5% | 78.5% |

**Table 6**
Different KG embedding algorithms on MetaQA.

| KG embedding | MetaQA KG-Full | MetaQA KG-50 |
|---|---|---|
| CF-KGQA_TransH | 97.4% | 83.4% |
| CF-KGQA_TransR | 97.2% | 83.3% |
| CF-KGQA_TransA | 97.6% | 83.5% |
| **CF-KGQA_TransE** | **97.7%** | **83.6%** |

*Note*: We only report the performance of 1-hop question results here.

hidden in knowledge space. Without CF module, the accuracy of head and predicate accuracy is limited from 0.60 to 0.62. (2) The more latent relations we utilize, CF-KGQA is more accurate and the trend gradually slows down. This is because our model is data-driven based which means the more high-quality data is fed in, the more accurate model can be constructed. (3) The results of using CF are significantly different from the other two settings which demonstrates the significance of the improvements.

Also, we test the sensitivity of the model size hyper parameter as shown in Table 5. It demonstrates that the model relies on some hyper parameters of the model size, especially when the hidden size and KG embedding dimension is lower than 150 and 300. It should be noted that the values of the model size are determined by the utilized dataset.

### 5.4.3. Robustness analysis

To further validate the robustness of CF-KGQA, we reshuffle the MetaQA KG-Full and MetaQA KG-50 randomly. The performance of CF-KGQA with different KG embedding algorithms on MetaQA is shown in Table 6.

According to Table 6, we observe that CF-KGQA could still achieve an accuracy of 83.6% using TransE method. And the difference between KG embedding algorithms has little impact on our proposed method. Furthermore, we test the capability of CF-KGQA. We remove all triples from the KG which can be used to answer all questions in the validation set of the MetaQA 1-hop dataset. For instance, given a question "Which artist created the city horizon? [Hans Hofmann]", we intentionally remove the triple (Hans_Hofmann, artist_create, city_horizon) from the KG. We also remove all the paraphrases of the same question in the validation set since we are only interested in evaluating the KG completion property of our model instead of linguistic

**Table 7**
Case study of CF-KGQA, comparing predictions by KEQA and our model (CF-KGQA). Our model causally corrects some triplet changes.

| Question (Originally taken from SimpleQuestion Dev) | Science fact | KEQA (baseline) prediction | Our prediction |
|---|---|---|---|
| Where was Walter Chrysler born? | Triplet: (people, person, place of birth) Answer: Wamego, Kansas, United States | Triplet: (people, person, place of live) Answer: Ellis, Kansas × (misunderstand the live and birth) | Triplet: (people, person, place of birth) Answer: Wamego, US √ |
| Which artist created the city horizon? | Triplet: (visual art, artwork, artist) Answer: Hans Hofmann | Triplet: (visual art, artwork, artist) Answer: Hans Hofmann √ | Triplet: (visual art, artwork, artist) Answer: Hans Hofmann √ |
| In which European nation was the Fit of Passion filmed? | Triplet: (film, film, country) Answer: Russia | Triplet: (film, film, country) Answer: Los Angeles × (lacking the hint European nation) | Triplet: (film, film, country) Answer: Belgium fairly× (Belgium is a European nation but not the correct answer) |
| Who is one of the members of the United States House Committee on the judiciary? | Triplet: (us congress, house committee, current members) Answer: Jerry Nadler (D) Since January 3, 2019 | Triplet: (US, house-committee, members) Answer: James Buchanan × | Triplet: (us, house committee, current members) Answer: Jerry Nadler √ |

generalization. In such a setting, it is expected that models relying only on sub-graph retrieval achieve 0 hits1. However, our model still has a 69.3% hits1. This demonstrates that our model is able to capture the KG completion property of complex embedding and answer questions that would have been impossible otherwise.

#### 5.4.4. Case study

To further validate the robustness and generalizability of CF-KGQA, we perform a case study on SimpleQuestions as shown in Table 7, which demonstrates that our CF-KGQA outperforms its baseline model KEQA in four cases; specifically, CF-KGQA can capture the latent/vague hint embedded in the questions, *e.g.*, the third question has a preliminary constraint that the nation should be in Europe, however both KEQA and our proposed model all give the wrong answers, but our model chooses a European nation as the candidate. Moreover, for the fourth question, the preliminary constraint is the tense which will determine whether the current members or previous members is the correct answer. CF-KGQA captures this subtle difference and returns the correct answer. These cases illustrate the superiority of removing causality confounding bias in our CF-KGQA.

### 5.5. Knowledge graph with missing-link evaluation

Some state-of-the-art KGQA models like PullNet [23] and GraftNet [24] require a path to be presented in the KG between the head entity and the answer entity for question answering. For example, in PullNet, the answers are restricted to contained within one of the entities from the extracted question subgraph, limiting the number of questions the model can answer in the setting. The CF-KGQA model, in contrast, inherits the idea of EmbeddedKGQA [53] and uses KGE algorithms instead of a localized sub-graph to answer the question. It adopts the head entity embedding and predicate embedding, which implicitly envelops all observed and unobserved links surrounding each node. Unlike other QA systems, even if there is no path between the head entity and answer entity, the CF-KGQA in other words should be able to answer the question if the KG contains enough information to determine where the path will lead (addressing the missing links of the KG).

Table 8 shows that our model reaches the state-of-the-art for 1-hop and 3-hop questions and performance is close to the state-of-the art for 2-hop questions on MetaQA KG-Full setting. CF-KGQA achieves the SOTA performance in the 1-hop situation because there is a direct connection between the answer entity and the head entity. The model can learn the corresponding relation embedding from the question easily and selects the correct answer easily. On the other hand, the performance on 2-hop and 3-hop questions indicates that CF-KGQA can deduce the correct relation from the node's neighboring edges as KG embeddings are able to capture the composition of relations. Pullnet and GraftNet are also effective since the answer entity is generally found in the question sub-graph. Moreover, our model performs better than EmbedKGQA which is due to the usage of disentangled latent factors. Our causal inference-based module reduces the spurious relation representation and improves the effectiveness of the multi-hop QA task.

**Table 8**
The performance of different methods on MetaQA.

| Model | MetaQA KG-Full | | | MetaQA KG-50 | | |
|---|---|---|---|---|---|---|
| | 1-hop | 2-hop | 3-hop | 1-hop | 2-hop | 3-hop |
| STAGG [59] | 95.9% | 91.4% | 83.1% | N.A. | N.A. | N.A. |
| GraftNet [24] | 97.0% | 94.8% | 77.7% | 64.0% | 52.6% | 59.2% |
| PullNet [23] | 97.0% | 94.8% | 91.4% | 65.1% | 52.1% | 59.7% |
| EmbedKGQA [53] | 97.5% | **98.8%** | 94.8% | **83.9%** | 91.8% | 70.3% |
| **Ours (CF-KGQA)** | **97.7%** | 98.3% | **95.1%** | 83.6% | **92.5%** | **73.8%** |

*Note*: We have considered both full KG (MetaQA KG-Full) and 50% KG (MetaQA KG-50) settings. All baseline results were taken from [53].

Moreover, we find that there is a significant reduction in the accuracy of all baselines compared to the full KG setting on the incomplete KG setting (MetaQA KG-50, which only contains 50% of the whole KG), while EmbedKGQA and CF-KGQA still perform well. This is because MetaQA KGs only involve 135k triples for a total number of 43k entities, which is fairly sparse. So, as 50% of the triples are removed, the graph becomes very sparse, causing many head entity nodes of questions to have much longer paths to their answer nodes, *i.e.*, more than three. As a consequence, models which require the construction of a question-specific sub-graph do not involve the answer entity in their generated sub-graphs, and leads to poor performance. However, EmbedKGQA and our CF-KGQA, utilizing the link prediction properties from the KG embeddings instead of limiting themselves to a sub-graph, are able to infer the relation on missing links. With CF utilized, our CF-KGQA overcomes the spurious correlation problem and performs better than EmbedKGQA on 2-hop and 3-hop.

### 5.6. Multi-hop question answering evaluation

Single-hop questions often have relatively simple structures and can be answered using information contained in one sub-graph. The transformation from the single-hop questions to multi-hop questions means the model need to manage to achieve close

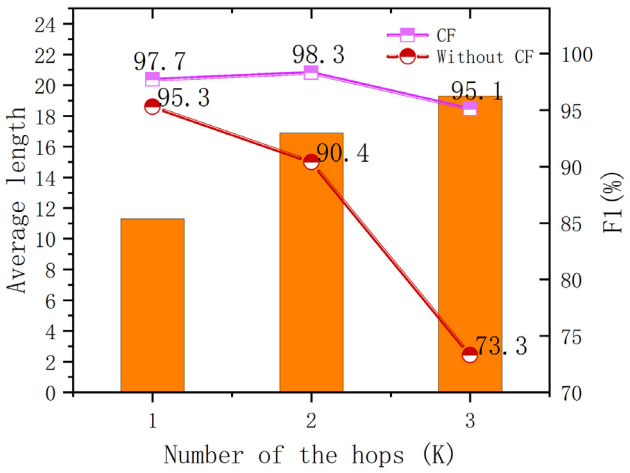


**Fig. 8.** Analysis of multi-hop question answering on MetaQA. We illustrate F1 score of CF-KGQA on MetaQA w.r.t. number of the hops (K).

to human performance. When generalizing such methods into the multi-hop area, we have to deal with two major challenges:

- Complex questions, particularly those with multiple question words, challenge the model's ability to understand the query correctly.
- Long context, which contains relevant information and noise, requires the model to identify noise, establish inference among multiple relevant sub-graphs, and process long sequence effectively.

**Why Causal Filtering?** Figure Fig. 8 shows that CF ensures the results remain stable across multi-hop questions, and can well trackle the two aforementioned challenges. Embedding-based model is well-known for being able to detect long-range dependencies. The multi-hop reasoning by matching the pre-trained entity embeddings with the queries embedding in the KG embedding spaces largely reduces the operation time for sequentially executed operations, and makes it easier to learn long-range dependencies because each token is connected to each other token. However, this method struggle in dealing with noise. Because of the large number of connections in the multi-hop questions, which are not all sequential, it will be thrown off by irrelevant data strongly. Thus, when the number of hop increases, the performance of the model without CF declines significantly. In general, the model without CF may not be optimal to multi-hop reasoning tasks because these tasks usually involve hopping over portions of data that are noisy and irrelevant-including their attention is detrimental to the model. However, with CF, the model can eliminates the effect of noise on the attention mechanism. It filters out the irrelevant confounding information according to the subtle correlation between the un-sequential multi-hop relations. With the CF module, the model can still achieves a high performance when $K = 3$ with only a limited decline.

**Impact of Number of Hops (K).** We evaluate the impact of number of hops in the query on MetaQA. As shown in Figure Fig. 8, the CF module reduces the F1 score by 19.4% compared to the original method by the increase of $K$. The impact of the hops not only involves the inference paths but also the number of tokens from each queries. We illustrates the average length of the each hop questions. It can be found that the average length of the 3-hop questions tokens is 18.9. With the increase of the numbers of the inferential chain, it will also have some impact on F1.

## 6. Conclusion and future work

Due to the increasing size of data, real world knowledge graphs often contain millions or billions of facts. KGQA suffers from the spurious correlation brought by entity name ambiguity, predicate expression variety, grounding difficulty and missing links. However, traditional solutions construct the relation embeddings based on a syntax tree and spatial–temporal co-occurrence by only maximizing the data likelihood. The learned relation representation contains a high level of spurious correlation. In this paper, we propose a causal filter based on the combination of a causal pairwise aggregator and disentangled latent factor aggregator to alleviate the challenges of the KGQA. Experiments show that our model (CF-KGQA) performs much better than the state-of-the-art methods on four public real-world datasets. In addition, CF-KGQA can effectively reduce the spurious correlations, demonstrating the robustness of our method.

It should be noted that the causal filter is a general causal inference-based model which can be applied to scale up the statistical likelihood to combine with the causal correlation. However, our CF-KGQA only handle the KGQA from the knowledge graph alone rather than using both the KG and text information, which is called the fusion model. Some fusion methods utilizing the combination of a KG and entity-linked text may be more appropriate when an incomplete KB is available with a large text corpus. Also, our model performs the causal filter and knowledge graph embedding based on pre-trained methods, and we intend to study how to jointly conduct the KG embedding, causal filtering, and question answering in an end-to-end manner in the future. Moreover, we plan to further extend the causal utility from the intervention level to the counterfactual level.

## CRediT authorship contribution statement

**Yuan Sui:** Methodology, Resource and writing, Including methodology development, Data collection, Experiment design, Paper writing. **Shanshan Feng:** Makes an important contribution on the data analysis, Algorithm design. **Huaxiang Zhang:** Mainly responsible for writing, including review and editing. **Jian Cao:** Mainly responsible for writing, including review and editing. **Liang Hu:** Mainly for model construction and data analysis. **Nengjun Zhu:** Mainly for software and coding.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported in part by Taishan Scholar Project of Shandong of China, the National Natural Science Foundation of China under Grant U1836216, the Major Fundamental Research Project of Shandong of China under Grant ZR2019ZD03, and the Shanghai Youth Science and Technology Talents Sailing Program under Grant 22YF1413700.

## References

[1] Y. Wang, R. Zhang, C. Xu, Y. Mao, The APVA-TURBO approach to question answering in knowledge base, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1998–2009, URL https://aclanthology.org/C18-1170.

[2] M. Yu, W. Yin, K.S. Hasan, C. dos Santos, B. Xiang, B. Zhou, Improved neural relation detection for knowledge base question answering, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 571–581, http://dx.doi.org/10.18653/v1/P17-1053, URL https://aclanthology.org/P17-1053.

[3] Y. Hao, Y. Zhang, K. Liu, S. He, Z. Liu, H. Wu, J. Zhao, An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 221–231, http://dx.doi.org/10.18653/v1/P17-1021, URL https://aclanthology.org/P17-1021.

[4] F. Ture, O. Jojic, No need to pay attention: Simple recurrent neural networks work!, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2866–2872, http://dx.doi.org/10.18653/v1/D17-1307, URL https://aclanthology.org/D17-1307.

[5] H. Bast, E. Haussmann, More accurate question answering on freebase, in: J. Bailey, A. Moffat, C.C. Aggarwal, M. de Rijke, R. Kumar, V. Murdock, T.K. Sellis, J.X. Yu (Eds.), Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015, ACM, 2015, pp. 1431–1440, http://dx.doi.org/10.1145/2806416.2806472.

[6] S. Jain, Question answering over knowledge base using factual memory networks, in: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 109–115, http://dx.doi.org/10.18653/v1/N16-2016, URL https://aclanthology.org/N16-2016.

[7] W. Yin, M. Yu, B. Xiang, B. Zhou, H. Schütze, Simple question answering by attentive convolutional neural network, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 1746–1756, URL https://aclanthology.org/C16-1164.

[8] W.-t. Yih, M.-W. Chang, X. He, J. Gao, Semantic parsing via staged query graph generation: Question answering with knowledge base, 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), Association for Computational Linguistics, Beijing, China, 2015, pp. 1321–1331, http://dx.doi.org/10.3115/v1/P15-1128, URL https://aclanthology.org/P15-1128.

[9] S. Hu, L. Zou, X. Zhang, A state-transition framework to answer complex questions over knowledge base, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2098–2108, http://dx.doi.org/10.18653/v1/D18-1234, URL https://aclanthology.org/D18-1234.

[10] Z.-Y. Chen, C.-H. Chang, Y.-P. Chen, J. Nayak, L.-W. Ku, UHop: An unrestricted-hop relation extraction framework for knowledge-based question answering, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 345–356, http://dx.doi.org/10.18653/v1/N19-1031, URL https://aclanthology.org/N19-1031.

[11] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, X. Zhu, Commonsense knowledge aware conversation generation with graph attention, in: J. Lang (Ed.), Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, ijcai.org, 2018, pp. 4623–4629, http://dx.doi.org/10.24963/ijcai.2018/643.

[12] X. Huang, J. Zhang, D. Li, P. Li, Knowledge graph embedding based question answering, in: J.S. Culpepper, A. Moffat, P.N. Bennett, K. Lerman (Eds.), Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019, ACM, 2019, pp. 105–113, http://dx.doi.org/10.1145/3289600.3290956.

[13] Y. Hao, H. Liu, S. He, K. Liu, J. Zhao, Pattern-revising enhanced simple question answering over knowledge bases, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3272–3282, URL https://aclanthology.org/C18-1277.

[14] W. Zhao, T. Chung, A. Goyal, A. Metallinou, Simple question answering with subgraph ranking and joint-scoring, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 324–334, http://dx.doi.org/10.18653/v1/N19-1029, URL https://aclanthology.org/N19-1029.

[15] S. Mohammed, P. Shi, J. Lin, Strong baselines for simple question answering over knowledge graphs with and without neural networks, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 291–296, http://dx.doi.org/10.18653/v1/N18-2047, URL https://aclanthology.org/N18-2047.

[16] V. Gupta, M. Chinnakotla, M. Shrivastava, Retrieve and re-rank: A simple and effective IR approach to simple question answering over knowledge graphs, in: Proceedings of the First Workshop on Fact Extraction and VERIfication (FEVER), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 22–27, http://dx.doi.org/10.18653/v1/W18-5504, URL https://aclanthology.org/W18-5504.

[17] Y. Qu, J. Liu, L. Kang, Q. Shi, D. Ye, Question answering over freebase via attentive RNN with similarity matrix based CNN, 2018, ArXiv preprint, abs/1804.03317 URL https://arxiv.org/abs/1804.03317.

[18] Z. Jia, A. Abujabal, R.S. Roy, J. Strötgen, G. Weikum, TEQUILA: temporal question answering over knowledge bases, in: A. Cuzzocrea, J. Allan, N.W. Paton, D. Srivastava, R. Agrawal, A.Z. Broder, M.J. Zaki, K.S. Candan, A. Labrinidis, A. Schuster, H. Wang (Eds.), Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, ACM, 2018, pp. 1807–1810, http://dx.doi.org/10.1145/3269206.3269247.

[19] M. Dubey, D. Banerjee, D. Chaudhuri, J. Lehmann, EARL: joint entity and relation linking for question answering over knowledge graphs, in: D. Vrandecic, K. Bontcheva, M.C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L. Kaffee, E. Simperl (Eds.), The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part I, in: Lecture Notes in Computer Science, vol. 11136, Springer, 2018, pp. 108–126, http://dx.doi.org/10.1007/978-3-030-00671-6_7.

[20] H. Ren, H. Dai, B. Dai, X. Chen, M. Yasunaga, H. Sun, D. Schuurmans, J. Leskovec, D. Zhou, LEGO: latent execution-guided reasoning for multi-hop question answering on knowledge graphs, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 8959–8970, URL http://proceedings.mlr.press/v139/ren21a.html.

[21] M.E. Roberts, B.M. Stewart, R.A. Nielsen, Adjusting for confounding with text matching, Amer. J. Polit. Sci. 64 (4) (2020) 887–903.

[22] K. Keith, D. Jensen, B. O'Connor, Text and causal inference: A review of using text to remove confounding from causal estimates, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5332–5344, http://dx.doi.org/10.18653/v1/2020.acl-main.474, URL https://aclanthology.org/2020.acl-main.474.

[23] H. Sun, T. Bedrax-Weiss, W. Cohen, PullNet: OPen domain question answering with iterative retrieval on knowledge bases and text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2380–2390, http://dx.doi.org/10.18653/v1/D19-1242, URL https://aclanthology.org/D19-1242.

[24] H. Sun, B. Dhingra, M. Zaheer, K. Mazaitis, R. Salakhutdinov, W. Cohen, Open domain question answering using early fusion of knowledge bases and text, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 4231–4242, http://dx.doi.org/10.18653/v1/D18-1455, URL https://aclanthology.org/D18-1455.

[25] T. Wang, J. Huang, H. Zhang, Q. Sun, Visual commonsense R-CNN, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, IEEE, 2020, pp. 10757–10767, http://dx.doi.org/10.1109/CVPR42600.2020.01077, URL https://doi.org/10.1109/CVPR42600.2020.01077.

[26] A. Talmor, J. Herzig, N. Lourie, J. Berant, CommonsenseQA: A question answering challenge targeting commonsense knowledge, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4149–4158, http://dx.doi.org/10.18653/v1/N19-1421, URL https://aclanthology.org/N19-1421.

[27] W. Zhong, D. Tang, N. Duan, M. Zhou, J. Wang, J. Yin, Improving question answering by commonsense-based pre-training, in: CCF International Conference on Natural Language Processing and Chinese Computing, Springer, 2019, pp. 16–28.

[28] A. Mitra, P. Banerjee, K.K. Pal, S. Mishra, C. Baral, Exploring ways to incorporate additional knowledge to improve natural language commonsense question answering, 2019, ArXiv Preprint abs/1909.08855 URL https://arxiv.org/abs/1909.08855.
[29] L. Bauer, Y. Wang, M. Bansal, Commonsense for generative multi-hop question answering tasks, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 4220–4230, http://dx.doi.org/10.18653/v1/D18-1454, URL https://aclanthology.org/D18-1454.
[30] V. Shwartz, P. West, R. Le Bras, C. Bhagavatula, Y. Choi, Unsupervised commonsense question answering with self-talk, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4615–4629, http://dx.doi.org/10.18653/v1/2020.emnlp-main.373, URL https://aclanthology.org/2020.emnlp-main.373.
[31] N. Egami, C.J. Fong, J. Grimmer, M.E. Roberts, B.M. Stewart, How to make causal inferences using texts, 2018, ArXiv Preprint abs/1802.02163 URL https://arxiv.org/abs/1802.02163.
[32] V. Veitch, D. Sridhar, D.M. Blei, Adapting text embeddings for causal inference, in: R.P. Adams, V. Gogate (Eds.), Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, Virtual Online, August 3-6, 2020, in: Proceedings of Machine Learning Research, vol. 124, AUAI Press, 2020, pp. 919–928, URL http://proceedings.mlr.press/v124/veitch20a.html.
[33] R. Pryzant, K. Shen, D. Jurafsky, S. Wagner, Deconfounded lexicon induction for interpretable social science, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1615–1625, http://dx.doi.org/10.18653/v1/N18-1146, URL https://aclanthology.org/N18-1146.
[34] A. Feder, N. Oved, U. Shalit, R. Reichart, CausaLM: Causal model explanation through counterfactual language models, Comput. Linguist. 47 (2) (2021) 333–386, http://dx.doi.org/10.1162/coli_a_00404.
[35] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, S.M. Shieber, Causal mediation analysis for interpreting neural NLP: The case of gender bias, 2020, ArXiv Preprint abs/2004.12265 URL https://arxiv.org/abs/2004.12265.
[36] Z. Wood-Doughty, I. Shpitser, M. Dredze, Challenges of using text classifiers for causal inference, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 4586–4598, http://dx.doi.org/10.18653/v1/D18-1488, URL https://aclanthology.org/D18-1488.
[37] J. Oh, K. Torisawa, C. Kruengkrai, R. Iida, J. Kloetzer, Multi-column convolutional neural networks with causality-attention for why-question answering, in: M. de Rijke, M. Shokouhi, A. Tomkins, M. Zhang (Eds.), Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017, ACM, 2017, pp. 415–424, http://dx.doi.org/10.1145/3018661.3018737.
[38] C. Pechsiri, R. Piriyakul, Developing a why–how question answering system on community web boards with a causality graph including procedural knowledge, Inf. Process. Agric. 3 (1) (2016) 36–53.
[39] R. Ishida, K. Torisawa, J. Oh, R. Iida, C. Kruengkrai, J. Kloetzer, Semi-distantly supervised neural model for generating compact answers to open-domain why questions, in: S.A. McIlraith, K.Q. Weinberger (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, AAAI Press, 2018, pp. 5803–5811, URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17199.
[40] D. Lukovnikov, A. Fischer, J. Lehmann, S. Auer, Neural network-based question answering over knowledge graphs on word and character level, in: R. Barrett, R. Cummings, E. Agichtein, E. Gabrilovich (Eds.), Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017, ACM, 2017, pp. 1211–1220, http://dx.doi.org/10.1145/3038912.3052675.
[41] K. Luo, F. Lin, X. Luo, K. Zhu, Knowledge base question answering via encoding of complex query graphs, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2185–2194, http://dx.doi.org/10.18653/v1/D18-1242, URL https://aclanthology.org/D18-1242.
[42] D. Sorokin, I. Gurevych, Modeling semantics with gated graph neural networks for knowledge base question answering, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3306–3317, URL https://aclanthology.org/C18-1280.
[43] L. Wu, P. Wu, X. Zhang, A seq2seq-based approach to question answering over knowledge bases, in: X. Wang, F.A. Lisi, G. Xiao, E. Botoeva (Eds.), Semantic Technology, Springer Singapore, Singapore, 2020, pp. 170–181.
[44] N. Gholizadeh, H. Saadatfar, N. Hanafi, K-DBSCAN: an improved DBSCAN algorithm for big data, J. Supercomput. 77 (6) (2021) 6214–6235, http://dx.doi.org/10.1007/s11227-020-03524-3.
[45] Y. Ren, K. Hu, X. Dai, L. Pan, S.C.H. Hoi, Z. Xu, Semi-supervised deep embedded clustering, Neurocomputing 325 (2019) 121–130, http://dx.doi.org/10.1016/j.neucom.2018.10.016.
[46] P. Vincent, H. Larochelle, Y. Bengio, P. Manzagol, Extracting and composing robust features with denoising autoencoders, in: W.W. Cohen, A. McCallum, S.T. Roweis (Eds.), Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008, in: ACM International Conference Proceeding Series, vol. 307, ACM, 2008, pp. 1096–1103, http://dx.doi.org/10.1145/1390156.1390294.
[47] A. Bordes, N. Usunier, A. García-Durán, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: C.J.C. Burges, L. Bottou, Z. Ghahramani, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a Meeting Held December 5-8, 2013, Lake Tahoe, Nevada, United States, 2013, pp. 2787–2795, URL https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html.
[48] S.Y. Yu, S.R. Chhetri, A. Canedo, P. Goyal, M.A.A. Faruque, Pykg2vec: A python library for knowledge graph embedding, 2019, arXiv:1906.04239.
[49] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751, http://dx.doi.org/10.3115/v1/D14-1181, URL https://aclanthology.org/D14-1181.
[50] W. Yin, M. Yu, B. Xiang, B. Zhou, H. Schütze, Simple question answering by attentive convolutional neural network, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 1746–1756, URL https://aclanthology.org/C16-1164.
[51] D. Lukovnikov, A. Fischer, J. Lehmann, S. Auer, Neural network-based question answering over knowledge graphs on word and character level, in: R. Barrett, R. Cummings, E. Agichtein, E. Gabrilovich (Eds.), Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017, ACM, 2017, pp. 1211–1220, http://dx.doi.org/10.1145/3038912.3052675.
[52] S. Mohammed, P. Shi, J. Lin, Strong baselines for simple question answering over knowledge graphs with and without neural networks, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 291–296, http://dx.doi.org/10.18653/v1/N18-2047, URL https://aclanthology.org/N18-2047.
[53] A. Saxena, A. Tripathi, P. Talukdar, Improving multi-hop question answering over knowledge graphs using knowledge base embeddings, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4498–4507, http://dx.doi.org/10.18653/v1/2020.acl-main.412, URL https://aclanthology.org/2020.acl-main.412.
[54] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67, URL http://jmlr.org/papers/v21/20-074.html.
[55] R. Das, M. Zaheer, D. Thai, A. Godbole, E. Perez, J. Lee, L. Tan, L. Polymenakos, A. McCallum, Case-based reasoning for natural language queries over knowledge bases, 2021, ArXiv Preprint abs/2104.08762 URL https://arxiv.org/abs/2104.08762.
[56] G. He, Y. Lan, J. Jiang, W.X. Zhao, J. Wen, Improving multi-hop knowledge base question answering by learning intermediate supervision signals, in: L. Lewin-Eytan, D. Carmel, E. Yom-Tov, E. Agichtein, E. Gabrilovich (Eds.), WSDM '21, the Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021, ACM, 2021, pp. 553–561, http://dx.doi.org/10.1145/3437963.3441753.
[57] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543, http://dx.doi.org/10.3115/v1/D14-1162, URL https://aclanthology.org/D14-1162.
[58] G.A. Rao, G. Srinivas, K. VenkataRao, P. Prasad Reddy, A partial ratio and ratio based fuzzy-wuzzy procedure for characteristic mining of mathematical formulas from documents, IJSC—ICTACT J. Soft Comput. 8 (4) (2018) 1728–1732.

[59] W.-t. Yih, M. Richardson, C. Meek, M.-W. Chang, J. Suh, The value of semantic parse labeling for knowledge base question answering, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 201–206, http://dx.doi.org/10.18653/v1/P16-2033, URL https://aclanthology.org/P16-2033.

[60] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, J. Leskovec, QA-GNN: Reasoning with language models and knowledge graphs for question answering, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 535–546, http://dx.doi.org/10.18653/v1/2021.naacl-main.45, URL https://aclanthology.org/2021.naacl-main.45.