Seeing the wood for the trees: a contrastive regularization method for the low-resource Knowledge Base Question Answering

Junping Liu[†], Shijie Mei[†],Xinrong Hu[†], Xun Yao[†], Jie Yang[‡], Yi Guo[◇]
[†]School of Computer Science and Artificial Intelligence, Wuhan Textile University
[‡]School of Computing and Information Technology, University of Wollongong
[◇]School of Computer, Data and Mathematical Sciences, Western Sydney University
{jpliu, hxr, yaoxun}@wtu.edu.cn, jackeymeis1@gmail.com
jiey@uow.edu.au,
y.guo@westernsydney.edu.au

Abstract

Given a context knowledge base (KB) and a corresponding question, the Knowledge Base Question Answering task aims to retrieve correct answer entities from this KB. Despite sophisticated retrieval algorithms, the impact of the low-resource (incomplete) KB is not fully exploited, where contributing components (*i.e.* key entities and/or relations) may be absent for question answering. To effectively address this problem, we propose a contrastive regularization based method, which is motivated by the `learn-by-analogy` capability from human readers. Specifically, the proposed work includes two major modules: the knowledge extension and sMoCo module. The former aims at exploiting the latent knowledge from the context KB and generating auxiliary information in the form of question-answer pairs. The later module utilizes those additional pairs and applies the contrastive regularization to learn informative representations, that making `hard` positive pairs attracted and `hard` negative pairs separated. Empirically, we achieved the state-of-the-art performance on the WebQuestionsSP dataset and the effectiveness of proposed modules is also evaluated.

1 Introduction

The task of Knowledge Base Question Answering (KBQA) refers to answering a question given a background knowledge base (KB). A large number of studies on KBQA can be cast into two mainstream categories: semantic parsing (SP) and information retrieval (IR) based. The former focuses on parsing questions into symbolic logic forms, such as query graph (Hu et al., 2018) and skeleton grammar (Sun et al., 2020), before identifying final answer(s) from the KB. The IR approach, on the other hand, aims to perform semantic matching between topic entities from questions and candidate answers within the KB (Xiong et al., 2019; Sun et al., 2019; Saxena et al., 2020; Yadati et al., 2021). This paper is on a novel method for information retrieval based KBQA.

Approaches to IR-based KBQA usually follow a three-step process, including question analysis, subgraph reasoning and answer matching. At first, the question analysis is to understand reasoning instructions behind questions for extracting topic entities and involved relations, *etc*. The second step is performed to retrieve relevant entities and relations (as subgraphs of the context KB) according to reasoning instructions, and further formulate candidate answers. The last step is to identify the best answer by estimating and ranking the semantic-relationship matching score between the given question and candidates.

However, existing IR-based matching pays insufficient attention to the `low resource` (incompleteness) nature of KB, where contributing components (*i.e.* neighboring entities, key relations and/or reasoning path) may be absent. The limit or lack of enough information poses challenges for the subsequent effective question answering. Some work have been proposed to utilize auxiliary information, such as extra question-related texts (Sun et al., 2019) and pre-trained KB embeddings (Saxena et al., 2020), which unfortunately could introduce noisy and misleading facts, not to mention the extra computational cost.

Notably, when human performs QA reasoning, one could infer other cases from one similar instance. Even with a new question, experienced readers could still make a guess using similar concepts or facts from their current knowledge, and potentially discriminate candidate answers from different aspects. For instance, given a question of "`What state did Al Gore represent?`", one could infer from his working experience (*i.e.* `The Tennessean (News)`) and/or the graduate university (*i.e.* `Vanderbilt University`), to further predict the correct answer of `Tennessee`.

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1085 - 1094

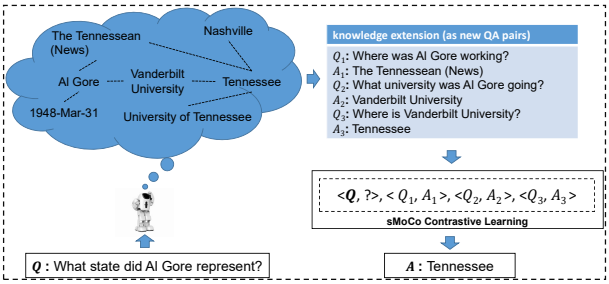

Figure 1: Overview of the proposed `learn-by-analogy` framework for KBQA, where latent knowledge is exploited and represented as the form of QA pairs. The sMoCo module is then applied to extract informative features for question answering.

Inspired from the aforementioned `learn-by-analogy` process, we propose a novel KBQA framework as illustrated in Fig. 1, which includes two major modules: knowledge extension and sMoCo. The former aims to expand knowledge from the existing KB to discover latent information. That is, hidden knowledge is exploited and manipulated in a form of question-answering pairs via making use of existing relations and entities (simulating human readers to use similar concepts/facts). The sMoCo adopts the contrastive learning mechanism, with the ultimate aim of capturing discriminative features from correct and misleading question-answering pairs (simulating human's inference skill). Furthermore, different from traditional contrast methods, the proposed sMoCo is particularly designed to utilize **hard** positive and negative pairs, which further improves the model generalizability. [1]

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to advance the contrastive learning-based algorithm to resolve the KBQA task; the proposed method is motivated by the `learn-by-analogy` capability from human readers;
- Simulating the human capability of utilizing similar facts, the proposed knowledge extension module explores the context KB to capture latent knowledge and further manipulates them as the task-related question-answer pairs;
- A novel contrast algorithm (sMoCo) is further introduced to simulate human inference capability; compared to existing methods, the proposed one mainly relies on `hard` positive and negative pairs during the training process. Related theoretical analysis is also provided for the discussion of its applicability and sensitivity;
- Experimental evaluations show that the proposed algorithm outperforms state-of-the-arts, in particular with the low-resource cases. We also conduct comprehensive ablation studies to characterize the proposed algorithm, by investigating the impact from the knowledge extension and hard positives/negatives.

[1] The source codes is publicly available at `https://github.com/JakeyMei/sMoCo`

2 Related work

2.1 KBQA

Knowledge Base Question Answering (KBQA) is one of the most popular and challenging research topics for machine reading comprehension (MRC). Aiming at determining correct answer(s) given the background knowledge base (KB) and one question, a large amount of research efforts have been put forward to supplement the KBQA task via either semantic parsing (SP) or information retrieval (IR) based strategies. The work from (Hu et al., 2018; Sun et al., 2020), for instance, belongs to the former, which introduced the state-transition and skeleton-based parsing approach to convert the target question into a semantic graph and structural tree, respectively, before querying answers.

Another line of studies aims to retrieve answers by following a more end-to-end training style, that is, to learn representations of the target question and candidate answers. KDReader from (Xiong et al., 2019) performed an attention-based fusion to combine the question and answer features. PullNet (Sun et al., 2019) employed the question-related content as a guideline to extract supporting components (entities and relations from KB). The work of (Saxena et al., 2020) utilized the pre-trained KB embeddings. In addition, Han *et al.* proposed a hypergraph-based reasoning strategy with dynamic relation and entity embeddings (Han et al., 2020). Similarly, RecHyperNet (Yadati et al., 2021) applied recursive hypergraphs to form groups (relations and entities with similar semantic) in the KB. With well-represented features, a matching step is usually followed to identify/rank the best candidate(s).

Despite some promising results from the afore-

mentioned feature learning and matching, less work has been put forward to explore underlying/hidden knowledge from the context KB. We argue that making full use of those existing but latent knowledge is beneficial to the subsequent QA task.

2.2 Contrastive learning

Contrastive learning (CL) has attracted a lot of attention in the recent several years, which utilizes input data itself as the additional supervision signal for training (Chen et al., 2020a; Grill et al., 2020; Chen and He, 2021; He et al., 2020; Chen et al., 2021; Zhu et al., 2021). Specifically, for an input sample (anchor) $\boldsymbol{x}_i$ and an encoder $f(\cdot)$, the overall contrastive loss is formulated as follows:

$$sim\left(f(\boldsymbol{x}_i), f(\boldsymbol{x}_i^+)\right) \gg sim\left(f(\boldsymbol{x}_i), f(\boldsymbol{x}_i^-)\right), \quad (1)$$

where $sim(\cdot,\cdot)$ is a user-defined similarity function, and $\boldsymbol{x}_i^+$ and $\boldsymbol{x}_i^-$ are contrastive pairs of positive and negative, respectively. The training purpose via the contrastive loss is to form informative features ($i.e. \boldsymbol{z}_i^{+/-} = f(\boldsymbol{x}_i)$) such that positives stay close to anchors and negatives are pushed away. Yet, existing CL methods fail to differentiate the pair significance, while the proposed method focuses on `hard` contrastive pairs.

3 Methodology

The illustration of our framework is shown in Fig 2. The KBQA task has been formulated as searching for optimal entities, given the question $\boldsymbol{q}$ and the external KB (*i.e.* a set of triples $(\boldsymbol{e}_h, \boldsymbol{r}, \boldsymbol{e}_t)$ representing head entity, semantic relationship and tail entity, respectively).

3.1 Knowledge extension module

Most existing KBQA work, unfortunately, focuses on the inference accuracy by offering sophisticated reasoning models. They have neglected the low resource nature of KB, from which key reasoning paths could be absent. Note that, experienced human readers could still utilize knowledge from limit but similar circumstances to infer correct answers. As such, this paper argues that it is beneficial to leverage hidden knowledge, from the context KB (even incomplete), for the subsequent QA process. That is the main aim for the proposed knowledge extension module.

Notably, knowledge extension serves the same role as data argumentation. Yet, this knowledge extension task is nontrivial. In the domain of image processing, argumentation can be carried out by using randomized operations, such as image rotation, cut-off, and scaling. However, in the context of KB, it is impractical to simply randomize knowledge, which could lead to meaningless or inaccurate information.

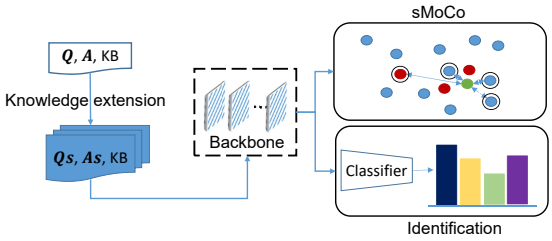

Figure 2: Overview of the proposed framework for KBQA which consists of four modules. First, the knowledge extension module is introduced to explore hidden knowledge from the context KB and produce new QA pairs. Second, a backbone module is applied to extract latent representations, after which the sMoCo module is employed, utilizing only hard contrastive pairs, to ensure that the similarity (of feature representations) between correct QA pairs is maximized by contrasting to that of inaccurate ones. At last, an overall identification module is utilized to identify the best-matching answer.

Intuitively, we propose to create new KB-relevant QA pair(s) from available triple(s) using an unsupervised manner. By doing so, we can alleviate the aforementioned challenge by not only effectively utilizing existing knowledge facts, but also providing additional samples for training the QA model. In this regard, we implement the knowledge-extension module using a *template*-oriented strategy[2].

Template
Triple: (<Psou River><fb:geography.river.origin> <Greater Caucasus>)
Desc.: <X> originates from <Y>
Extended question **Wh + [e_h] + Desc.[r] ?**: Where Psou River originates from?
Answer: Greater Caucasus

Figure 3: Example of producing the template-oriented QA pairs.

Given an known triple of $(\boldsymbol{e}_h, \boldsymbol{r}, \boldsymbol{e}_t)$, as the example illustrated in Figure 3. The relation of `fb:geography.river.origin` is associated with a factual description (`Desc.[`$\boldsymbol{r}$`]`), that can be interpreted as `[X] originates from`

[2]This strategy can also be viewed as the task of question generation. As such, some general generation techniques (Chen et al., 2020b; Bi et al., 2020) may help if complicated questions (with multiple hops) are preferred and we leave the investigation to future work.

`[Y]` (`[X]` and `[Y]` is the placeholder for $\boldsymbol{e}_h$ and $\boldsymbol{e}_t$ respectively). Then, we produce a template question "`[Wh]` + `[`$\boldsymbol{e}_h$`]` + `Desc.[`$\boldsymbol{r}$`]`?", where `Desc.[`$\boldsymbol{r}$`]` is the relation description. The `[Wh]` component belongs to one of the types (such as `who`, `where`, and `when`), depending on the tag of $\boldsymbol{e}_t$ (such as person, location, and time).

3.2 Backbone module

We adopt KDReader from (Xiong et al., 2019) as the backbone module. For a given question $\boldsymbol{q}$, KDReader firstly retrieves relevant entities from the context KB, according to topic entities $\boldsymbol{e}_q$ from $\boldsymbol{q}$, and further formulates entity embeddings $\mathcal{G}_{\boldsymbol{e}}$. Secondly, with the question embedding $Q_{\boldsymbol{q}}$, a classifier C is trained via optimizing the following loss function:

$$\mathcal{L}_{Matching} = SoftMax(C(\mathcal{G}_{\boldsymbol{e}}, Q_{\boldsymbol{q}})). \quad (2)$$

3.3 sMoCo module

Learning distinguishable features to match correct entities with the given question lies at the heart of KBQA, for which we accordingly propose a new contrastive learning algorithm in this module. Yet, majority contrast methods are characterized by a slow momentum encoder (E_p) and an independent queue ($\mathcal{Q}$) with thousands of negative samples. Despite its simplicity, the existing design poses the following problems: (i) the slow encoder E_p could produce `easy` positive samples (highly similar to anchors); similarly (ii) the lengthy queue $\mathcal{Q}$ may contain `easy` negative samples (highly different to anchors).

Notably, existing methods fail to differentiate `easy` and `hard` positives/negatives. We argue that the contribution of easy ones to the contrastive loss is trivial, while the significance from hard ones is the success toward the contrastive learning. As a result, we propose a screening-based Momentum Contrast method (termed **sMoCo**) via *screening* hard contrastive pairs.

To illustrate this idea, let $\mathbf{z}_i$ and $\mathcal{P}$ represent the i-th anchor feature and the set of corresponding positives. We propose to select only one `hard` positive from $\mathcal{P}$, say $\mathbf{z}_i^{h,+}$, with the lowest similarity to $\mathbf{z}_i$. On the other hand, we select `hard` negatives according to their similarity to $\mathbf{z}_i$ to form a subset $\mathcal{I}_-$ (with $\mathcal{I}_- \in \mathcal{Q}$). More precisely, $\forall \mathbf{z}_i^{h,-} \in \mathcal{I}_-$ and $\forall \mathbf{z}_n \in \mathcal{Q}$, we have

$$s(\mathbf{z}_i^{h,-}) > s(\mathbf{z}_i^-), \forall \mathbf{z}_i^- \in \mathcal{Q}, \ \mathbf{z}_i^- \notin \mathcal{I}_-,$$
$$\text{and} \sum_{\mathbf{z}_i^{h,-} \in \mathcal{I}_-} s(\mathbf{z}_i^{h,-}) = \lambda_{\mathcal{Q}} \sum_{\mathbf{z}_n \in \mathcal{Q}} s(\mathbf{z}_n), \quad (3)$$

where $s(\cdot)$ is the normalized inner product (for simplicity say $s(\mathbf{z}_i^{h,-}) = \langle \mathbf{z}_i^{h,-}, \mathbf{z}_i \rangle$), and $\lambda_{\mathcal{Q}}$ is a screening term. As such, the proposed contrastive loss for $\mathbf{z}_i$ is formulated as follows:

$$\mathcal{L}_{sMoCo} = -\log \frac{e^{\langle \mathbf{z}_i^{h,+}, \mathbf{z}_i \rangle}}{e^{\langle \mathbf{z}_i^{h,+}, \mathbf{z}_i \rangle} + \sum_{j \in \mathcal{I}_-} e^{\langle \mathbf{z}_j, \mathbf{z}_i \rangle}} \quad (4)$$

In addition to utilizing the proposed hard positive $\mathbf{z}_i^{h,+}$ and hard negatives $\mathcal{I}_-$, we further consider a liner combination of updating $\mathbf{z}_n$ from the negative queue $\mathcal{Q}$ via:

$$\mathbf{z}_n = \lambda_U \mathbf{z}_i + (1 - \lambda_U) \mathbf{z}_i^{h,+}, \quad (5)$$

where $\lambda_U \in [0, 1]$ is a hyperparameter.

Comparison with existing contrastive methods. The mainstream approaches for implementing contrastive learning include SimCLR (Chen et al., 2020a), MoCoV1/2 (He et al., 2020), MoCoFT (Zhu et al., 2021), MoCoV3 (Chen et al., 2021), SimSiam (Chen and He, 2021). Although our approach shares similar idea of utilizing positives and/or negatives as the above, our algorithm is different in the following: (i) the majority existing work neglects the difference of easy and hard contrastive pairs; for instance, MoCoV1/2/3 and SimSiam only considers one positive, while SimCLR selects the easiest one (with the highest similarity with anchors) with the presence of multiple positives; meanwhile, they take all available negatives into account. By contrast, ours is particularly interested in hard positives and negatives; (ii) MoCoFT applies feature-level transformation (FT) to generate hard positives/negatives, which in fact has no direct impact for/from encoders; another drawback then lies in the sensitivity of those FT hyperparameters, that further reduces its generalization capability; Similarly, SimSiam is also significantly impacted by the hyperparameters to produce positives, as no negative exists; and (iii) for updating the negative queue, the proposed update can be cast as a linear combination of SimCLR and MoCos, where other methods (such as MoCoFT) require one entire queue for each single anchor, that is computationally expensive. The aforementioned difference is also summarized in Table 1.

Table 1: Comparison between sMoCo and existing methods, where $|b|$, $|\mathcal{Q}|$, 1-E and 1-H represents the batch size, queue size, one easy and one hard positive, respectively.

	$\mathcal{P}$	$\mathcal{Q}$	Update $\mathcal{Q}$
SimCLR	1-E	$\|b\|$-2	$\mathbf{z}_n = \mathbf{z}_i$
MoCoV1/2	1	$\|\mathcal{Q}\|$	$\mathbf{z}_n = \mathbf{z}_i^+$
MoCoFT	1-H	$\|\mathcal{Q}\|$	$\mathbf{z}_n = \lambda \mathbf{z}_i + (1-\lambda)\mathcal{Q}$
MoCoV3	1	$\|b\|$-2	$\mathbf{z}_n = \mathbf{z}_i$
SimSiam	1	$\times$	$\times$
sMoCo	1-H	$\mathcal{I}_{n,-}$	$\mathbf{z}_n = \lambda_U \mathbf{z}_i + (1-\lambda_U)\mathbf{z}_i^{h,+}$

3.4 Analysis

We present the sensitivity analysis hereafter to justify the choice of the screening parameter (or $\lambda_{\mathcal{Q}}$ from Eq. (3)). Note that the traditional contrastive loss, for the i-th anchor, is defined as follows

$$\mathcal{L}_c = -\log \frac{e^{\langle \mathbf{z}_i^{h,+}, \mathbf{z}_i \rangle}}{e^{\langle \mathbf{z}_i^{h,+}, \mathbf{z}_i \rangle} + \sum_j e^{\langle \mathbf{z}_j^-, \mathbf{z}_i \rangle}}. \quad (6)$$

Denoting $s_i^+ = e^{\langle \mathbf{z}_i^{h,+}, \mathbf{z}_i \rangle}$ and $s_i^- = \sum_j e^{\langle \mathbf{z}_j^-, \mathbf{z}_i \rangle}$, Eq. (6) can be written more concisely as

$$\mathcal{L}_c = -\log \frac{s_i^+}{s_i^+ + s_i^-} = \log(1 + \frac{s_i^-}{s_i^+}). \quad (7)$$

We then quantify the screened proportion as $\epsilon \in [0, 1)$ for negative samples, and note that

$$\epsilon = 1 - \lambda_{\mathcal{Q}}.$$

By removing a given proportion of negative samples, it is equivalent to reduce s_i^- to $\tilde{s}_i^-$ by a suitable value of ϵ such that $\tilde{s}_i^- = (1-\epsilon)s_i^-$. Hence the perturbed loss function $\tilde{\mathcal{L}}_c$ becomes

$$\tilde{\mathcal{L}}_c(\epsilon | r_i) = \log\left(1 + (1-\epsilon) r_i\right), \quad (8)$$

where $r_i = \frac{s_i^-}{s_i^+}$. The perturbed loss is caused by ϵ and $\tilde{\mathcal{L}}_c(0|r_i)$ recovers the original loss in Eq. (7). Therefore the sensitivity of the function in Eq. (8) at $(\epsilon = 0)$ determines the change to the loss by introducing screening. The first order Taylor expansion shows that:

$$\tilde{\mathcal{L}}_c(\epsilon | r_i) = \tilde{\mathcal{L}}_c(0|r_i) + \frac{\partial \tilde{\mathcal{L}}_c}{\partial \epsilon}|_{\epsilon=0}[r_i] \cdot \epsilon + O(|\epsilon|^2),$$

where $O(|\epsilon|^2)$ is the negligible higher order term when ϵ is approximate zero. Note that $\tilde{\mathcal{L}}_c(0|r_i) = \mathcal{L}_c$, *i.e.* the original loss. Plugging in the partial derivatives evaluated at $\epsilon = 0$ with r_i ($\frac{\partial \tilde{\mathcal{L}}_c}{\partial \epsilon}|_{\epsilon=0}[r_i]$), *i.e.*

$$\frac{-r_i}{r_i + 1},$$

we obtain

$$\tilde{\mathcal{L}}_c(\epsilon | r_i) \approx \mathcal{L}_c - \frac{r_i \epsilon}{r_i + 1}. \quad (9)$$

Remark 1. *The perturbation (or the difference in loss aroused by screening) is approximately $\frac{-r_i\epsilon}{r_i+1}$, which is an apparent reduction to the original loss $\mathcal{L}_c$ as all variables in Eq. (9) are positive.*

Furthermore, it is worth noting that normalized similarities are between $[-1, 1]$, that is, $\langle \mathbf{z}_i^{h,+}, \mathbf{z}_i \rangle \to 1$, and $\langle \mathbf{z}_j^-, \mathbf{z}_i \rangle \to -1$ at the convergence of the optimization, and hence

$$r_i \to e^{-2}|\mathcal{Q}| > \frac{|\mathcal{Q}|}{9},$$

where $|\mathcal{Q}|$ is the size of the negative queue or number of negative samples. In many contrast methods (He et al., 2020; Zhu et al., 2021; Chen et al., 2021), $\mathcal{Q}$ consists of thousands of negatives, which leads to $r_i \gg 1$.

At last, combining $\frac{\partial \tilde{\mathcal{L}}_c}{\partial \epsilon}|_{\epsilon=0}[r_i]$ with the value of r_i, we obtain

$$\frac{\partial \tilde{L}_c}{\partial \epsilon}|_{\epsilon=0}[r_i] \approx -1,$$

and the perturbation is approximately $-\epsilon$ at convergence.

Remark 2. *The perturbation indicates that that if we screen out a small proportion of negative samples, say $1 - \lambda_{\mathcal{Q}}$ (or ϵ), the function loss is reduced by approximately that much. In other words, the sample screening connects to the loss reduction directly. Although we discussed only ideal situation where the contrast reaches maximum, similar result exists in expectation sense as well because a large value of $|\mathcal{Q}|$ outweighs less contrast in negative and positive similarities.*

Notably, the above perturbation analysis is performed at 0 requiring ϵ to be small or even close to zero (say $\epsilon \in [0.01, 0.1]$), and hence satisfies this condition.The loss function change caused by screening will eventually transfer to gradients to the model parameters. Our results indicates the scaling factor in front of the final gradients. Nonetheless, when r_i is too large, it saturates the loss function to the "plateau" stage where gradients are small and hence the aforementioned analysis results hold.

3.5 Overall identification

With the matching and contrastive loss from the Backbone and sMoCo modules, we propose the following *joint* loss for the overall identification:

$$\mathcal{L}_{Identification} = \mathcal{L}_{Matching} + \lambda_{sMoCo}\mathcal{L}_{sMoCo}, \tag{10}$$

where λ_{sMoCo} is a penalty term[3].

4 Experiment

4.1 Setup

The WebQuestionsSP datasets is employed (Yih et al., 2016) with a total of 4737 questions, that are answerable through the Freebase KB. This entire KB consists of 601,145 distinct entities, 568 different relations, 1,261,849 unique triples. Followed by the work from (Xiong et al., 2019; Saxena et al., 2020), the low-resource KB settings have been constructed by down-sampling a percentage of facts in the background KB (we randomly retain a triple with probability of 0.1, 0.3, and 0.5). Accordingly, the resultant datasets are referred as KB0.1, KB0.3, KB0.5 and KBFull (original), respectively, and Table 2 shows the statistics of those four datasets.

Table 2: Summary of four adopted datasets, where **#*e***, **#*r*** and **#triples** is the averaged number of entities, relations, and triples per question, respectively.

Dataset	#*e*	#*r*	#triples
KB0.1	152.5	51.3	184.8
KB0.3	182.3	54.8	567.1
KB0.5	183.5	56.1	837.3
KBFull	191.2	56.7	1484.2

To make a fair comparison, the hyperparameter setting of the backbone module is adopted explicitly from (Xiong et al., 2019), such as implementing the 300-d GloVe embeddings for question entities, maximal number of neighboring entity as 50, maximal question length as 10. The matching loss has been implemented using binary cross-entropy loss with 0.1 smoothing factor. In addition, the mini-batch size is 16, the Adam optimizer with a learning rate setting of 0.001, and the number of training epoch is set as 100. For the implemented knowledge extension module, factual descriptions about entity and relation are available publicly[4]. Meanwhile, we generate three additional QA pairs for one single QA input. For the sMoCo, the momentum rate is 0.99, $\tau = 0.07$, $\lambda_{sMoCo} = 0.2$, $\lambda_{\mathcal{Q}} = 0.95$, $\lambda_U = 0.4$, and the negative queue capacity is set as 10000. The proposed model is trained on a machine with four Tesla V100 GPUs. The Hits@1 score is used to measure the performance.

4.2 Main Results

We compare the proposed with different methods, including the baseline model (*i.e.* KDReader (Xiong et al., 2019)) and state-of-the-arts (*i.e.* PullNet (Sun et al., 2019), 2HR-DR (Han et al., 2020), EmbedKGQA (Saxena et al., 2020), and RecHyperNet (Yadati et al., 2021)). The comparison results over 10 trails are shown in Table 3.

Table 3: Results in Hits@1 obtained by sMoCo and existing methods re-implemented for four test sets. The number within the bracket indicates the original result reported by the paper.

Algorithm	KB0.1	KB0.3	KB0.5	KBFull
KDReader	33.5(33.6)	42.6(42.6)	52.8(52.7)	67.4(67.2)
PullNet	33.7	42.8	52.1(51.9)	68.0(68.1)
2HR-DR	33.5	42.5	52.0(52.2)	66.9(67.0)
EmbedKGQA	34.3	41.5	53.2(53.2)	67.0(66.6)
RecHyperNet	34.5	43.1	53.6(53.7)	68.4(68.4)
sMoCo	36.1	44.2	54.1	69.2

sMoCo shows superior performance compared to the state-of-the-arts via achieving a considerable margin. For instance, competing with the strongest baseline RecHyperNet, the proposed method outperforms by 2.55%, 0.93%, 1.17% with respect to KB0.3 KB0.5 and KBFull. It is also worth noting that sMoCo obtains the highest performance boost with KB0.1 (7.76% and 4.63% compared to the baseline and RecHyperNet), which demonstrates its strong capability of handling the extreme low-resource KB.

In addition, to further evaluate the improvement from sMoCo, we also compare ours with cutting-edge contrastive learning techniques, including SimCLR (Chen et al., 2020a), MoCoV1 (He et al., 2020), MocoFT (Zhu et al., 2021), MoCoV3 (Chen et al., 2021), and SimSiam (Chen and He, 2021). Again, the backbone module from KDReader is

[3]There are another two training strategies, including *pre-train* and *alternate*. The former is to update the model first using $\mathcal{L}_{sMoCo}$ before fine-tuning with $\mathcal{L}_{Matching}$, while the latter is to train the model with $\mathcal{L}_{Matching}$ for $(N_t - 1)$ iterations and switch to $\mathcal{L}_{sMoCo}$ once, for every N_t iterations. We leave these as the future work.

[4]Description about entities and relations can be found https://developers.google.com/freebase and https://freepal.appspot.com/, respectively

implemented for all contrastive methods for a fair comparison. In addition, hyperparameters for employed contrastive methods are set to similar as sMoCo: the momentum rate is 0.99, $\tau = 0.07$, $|\mathcal{Q}|$ = 10000, while the prediction head is implemented as a 2-layer MLP with a 512-hidden dimension.

Table 4: Average Hits@1 results (10 trials) obtained by sMoCo and other contrastive learning methods.

Algorithm	**KB0.1**	**KB0.3**	**KB0.5**	**KBFull**
SimCLR	34.5	42.9	52.9	67.3
MoCoV1	34.5	43.2	52.8	67.8
MocoFT	35.4	43.6	53.3	67.9
MoCoV3	35.0	43.4	53.3	68.1
SimSiam	34.5	43.2	52.9	67.5
sMoCo	36.1	44.2	54.1	69.2

Table 4 clearly demonstrates the superiority of applying contrastive based methods, as averaged results (over 10 trials) show all methods achieve better performance compared to the baseline model (*i.e.*, KDReader). In particular, contrastive methods perform specifically well with low-resource KBQA, as they lead to from 2.98% to 7.76% averaged improvement for KB0.1 (higher than other three cases). This result indicates the advantage of using contrastive learning for extremely low-resource KBQA tasks. Additionally, the proposed sMoCo outperforms existing methods, by achieving a notable accuracy (50.9 for overall averaged Hits@1) compared to that of SimCLR(49.4), MoCoV1(49.6), MocoFT(50.1), MoCoV3(50.0), and SimSiam(49.5), respectively.

4.3 Ablation study

Experiments are conducted to validate contributions from proposed modules, mainly the knowledge extension (KE) and sMoCo module. To highlight the low-resource nature of KG, we are particularly interested in the performance on datasets of KB0.1 and KB0.3. Again, all the results are reported as an averaged Hits@1 over 10 trials.

On knowledge extension module. To begin with, we consider the impact from the proposed knowledge extension (or data argumentation) by treating additional QA pairs as training samples, while no contrastive learning (sMoCo) is applied.

Results in terms of different numbers of additional pairs ($|Q|_a$) are then summarized in Table 5. Compared to the baseline (KDReader), the advantage of KE is observed from the performance

Table 5: Performance against different numbers ($|Q|_a$) of additional QA pairs (without sMoCo).

$\|Q\|_a$	**KB0.1**	**KB0.3**
baseline	33.5	42.6
1	33.9	42.9
3	33.7	43.1

improvement. With $|Q|_a = 3$, for instance, the model produces 38.4 for Hits@1 on average. Not surprisingly, the model performance is enhanced by the proposed KE (via providing additional training samples). In the following, we fix $|Q|_a = 3$ for KE and analyze the contrastive module.

Notably, compared to traditional method (such as MoCoV1), sMoCo is different from three aspects: 1) adopting hard positive and 2) negative samples for estimating the contrastive loss, and 3) queue updating with hard positives (anchors). We accordingly perform ablation study on individual aspect to manifest their efficacy and the results are summarized in Table 6. In particular, for comparison purposes we take "KE" and "MoCoV1" to represent the result from the knowledge extension module and normal contrastive learning, as the baseline; "+hard positive" differs from MoCoV1 by using one hard positive; "+hard negative" considers a subset of hard negatives while maintaining a 95% of total sum, in addition to "+hard positive"; "+queue updating" further applies the proposed combination strategy to update negatives iteratively.

Table 6: Ablation study on hard positive, hard negative, and queue updating from the proposed sMoCo.

Model Variants	**KB0.1**	**KB0.3**
KE	33.7	43.1
MoCoV1	34.5	43.2
+hard positive	34.8(↑ 0.3)	43.4(↑ 0.2)
+hard negative	35.7(↑ 0.9)	43.8(↑ 0.4)
+queue updating	36.1(↑ 0.4)	44.2(↑ 0.4)

Results from Table 6 show contributions from individual aspect to the final performance, which evidently states their effectiveness. At first, not surprisingly, all four contrastive variants improve the overall performance compared to KE, which again demonstrate the superiority of forming discriminate features to separate positives and negatives. Additionally, we observe the step of "+hard negative" brings the biggest performance boost, followed by

"+queue updating". The result highlights the significance of maintaining and utilizing hard negatives for contrastive learning, instead of employing a large negative queue as existing methods.

On hard positives. We then consider the individual impact from hard positives by fixing the rest setting of sMoCo, such as screening hard negatives ($\lambda_{\mathcal{Q}} = 0.95$) and updating $\mathcal{Q}$ with $\lambda_U = 0.4$. Precisely, four utilization, including "1-pos" (employing only one positive like MoCoV1/2), "3-pos" (with three positives), "1-easy" (picking up the most similar/easy positive like SimCLR), and "1-hard" (the proposed), are considered, while their comparison is illustrated in Fig. 4.

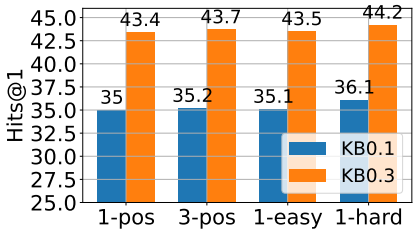

Figure 4: Comparison among four different ways of utilizing positives in sMoCo.

Notably, applying more positives (*i.e.*, "3-pos") performs competitively than cases with only one positive (either "1-pos" or "1-easy"). One reason could be the variety of positives helps in avoiding the model collapsing (that is, positives becomes very similar to anchors). Additionally, focusing on the hardest one achieves the best result, as shown in "1-hard", as it forces the model to pay attention to the positive far away from the anchor.

On hard negatives. Next we fix to use "1-hard" and evaluate the impact from negatives (by testing different $\lambda_{\mathcal{Q}}$, in line with the analysis presented in Section 3.4). Note that with a large $\lambda_{\mathcal{Q}}$, more negatives are included in the loss calculation. In particular, with $\lambda_{\mathcal{Q}} = 100\%$, sMoCo applies all negatives (same as MoCoV1/2/FT).

Table 7: Performance comparison in terms of hard negatives as a function of $\lambda_{\mathcal{Q}}$.

$\lambda_{\mathcal{Q}}$	100%	99%	97%	95%	90%
KB0.1	34.5	34.7	35.2	36.1	35.5
KB0.3	43.5	43.6	43.6	44.2	43.8

Table 7 shows the improvement by utilizing hard negatives, from which the model with the full queue ($\lambda_{\mathcal{Q}} = 100\%$) achieves the worst result. Additionally, the best performance is observed with $\lambda_{\mathcal{Q}} = 95\%$, approximately 2000 negatives. This findings clearly suggest that it is unnecessary to have a large number of negatives for a performance gain, not to mention its computational cost with a huge queue.

On updating negatives. At last, by fixing "1-hard" and $\lambda_{\mathcal{Q}} = 95\%$, we investigate the impact of updating negatives via changing λ_U. As discussed before, will a small λ_U (=0), sMoCo adopts positives directly to replace previous negatives like MoCoV1/2; on the other hand, sMoCo behaves similar to SimCLR or MoCoV3 if $\lambda_U = 1$.

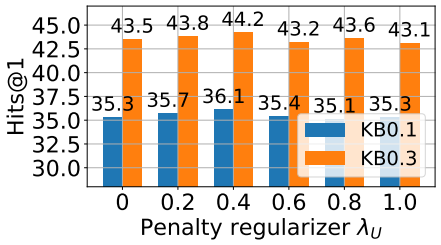

Figure 5: Model accuracy obtained from sMoCo as a function of λ_U via updating $\mathcal{Q}$.

Fig. 5 shows the model accuracy based on different settings of λ_U. Results indicate that the proposed combination of positives and anchors (with $\lambda_U = 0.4$) could bring in the performance boost (than that of $\lambda_U = 0$ or 1). Note that for KBQA, there exists QA pairs requiring same entities and relations. Using previous positives or anchors to replace previous negatives could mislead the model to separate them into different feature spaces. By contrast, the proposed updating maintains a good balance of forming hard but informative negatives.

5 Conclusion

We present a novel KBQA model that particularly tackles the low-resource (incompleteness) nature of the context knowledge bases (KBs). The proposed model is characterized by a knowledge extension and a sMoCo module, that is motivated by the learn-by-analogy capability of human readers. Precisely, the former extends existing knowledge via producing additional question-answer pairs, which are further utilized by the sMoCo module. The latter appropriately learns informative representations that grouping hard positives and pushing away hard negatives. Empirically, in comparison to existing approaches, the proposed algorithm produces the state-of-the-art performance on the WebQues-

tionsSP benchmark, in particular with the extreme incomplete KBs. In future work, we will extend the proposed idea to explore more contrastive behavior of utilizing hard positives and negatives. More importantly, sMoCo is agnostic to the downstream tasks, *i.e.*, we could incorporate it into other applications.

Acknowledgments

This work was partially supported by the Australian Research Council Discovery Project (DP210101426).

References

Sheng Bi, Xiya Cheng, Yuan-Fang Li, Yongzhen Wang, and Guilin Qi. 2020. Knowledge-enriched, Type-constrained and Grammar-guided question generation over Knowledge Bases. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2776–2786, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Xinlei Chen and Kaiming He. 2021. Exploring simple Siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758.

Xinlei Chen, Saining Xie, and Kaiming He. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9640–9649.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2020b. Toward subgraph guided Knowledge Graph question generation with Graph Neural Networks. *arXiv preprint arXiv:2004.06015*.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. 2020. Bootstrap Your Own Latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc.

Jiale Han, Bo Cheng, and Xu Wang. 2020. Two-phase hypergraph based reasoning with dynamic relations for multi-hop KBQA. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3615–3621. International Joint Conferences on Artificial Intelligence Organization.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738.

Sen Hu, Lei Zou, and Xinbo Zhang. 2018. A state-transition framework to answer complex questions over Knowledge Base. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2098–2108, Brussels, Belgium. Association for Computational Linguistics.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. 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*, pages 4498–4507, Online. Association for Computational Linguistics.

Haitian Sun, Tania Bedrax-Weiss, and William Cohen. 2019. 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)*, pages 2380–2390, Hong Kong, China. Association for Computational Linguistics.

Yawei Sun, Lingling Zhang, Gong Cheng, and Yuzhong Qu. 2020. Sparqa: Skeleton-based semantic parsing for complex questions over Knowledge Bases. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8952–8959.

Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Improving question answering over incomplete KBs with knowledge-aware reader. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4258–4264, Florence, Italy. Association for Computational Linguistics.

Naganand Yadati, Dayanidhi R S, Vaishnavi S, Indira K M, and Srinidhi G. 2021. Knowledge Base Question Answering through recursive hypergraphs. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 448–454, Online. Association for Computational Linguistics.

Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. 2016. 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), pages 201–206, Berlin, Germany. Association for Computational Linguistics.

Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, and Chang Wen Chen. 2021. Improving contrastive learning by visualizing feature transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10306–10315.