



An empirical study of pre-trained language models in simple knowledge graph question answering

Nan Hu¹ · Yike Wu¹ · Guilin Qi¹ · Dehai Min¹ · Jiaoyan Chen² · Jeff Z Pan³ · Zafar Ali¹

Received: 20 October 2022 / Revised: 18 February 2023 / Accepted: 12 March 2023 /

Published online: 17 May 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Large-scale pre-trained language models (PLMs) such as BERT have recently achieved great success and become a milestone in natural language processing (NLP). It is now the consensus of the NLP community to adopt PLMs as the backbone for downstream tasks. In recent works on knowledge graph question answering (KGQA), BERT or its variants have become necessary in their KGQA models. However, there is still a lack of comprehensive research and comparison of the performance of different PLMs in KGQA. To this end, we summarize two basic KGQA frameworks based on PLMs without additional neural network modules to compare the performance of nine PLMs in terms of accuracy and efficiency. In addition, we present three benchmarks for larger-scale KGs based on the popular Simple-Questions benchmark to investigate the scalability of PLMs. We carefully analyze the results of all PLMs-based KGQA basic frameworks on these benchmarks and two other popular datasets, WebQuestionSP and FreebaseQA, and find that knowledge distillation techniques and knowledge enhancement methods in PLMs are promising for KGQA. Furthermore, we test ChatGPT (<https://chat.openai.com/>), which has drawn a great deal of attention in the NLP community, demonstrating its impressive capabilities and limitations in zero-shot KGQA. We have released the code and benchmarks to promote the use of PLMs on KGQA (<https://github.com/aannonymouss/PLMs-in-Practical-KBQA>).

Keywords Knowledge graph question answering · Pretrained language models · Accuracy and efficiency · Scalability

This article belongs to the Topical Collection: *Special Issue on Knowledge-Graph-Enabled Methods and Applications for the Future Web*

Guest Editors: Xin Wang, Jeff Pan, Qingpeng Zhang, and Yuan-Fang Li.

✉ Guilin Qi
gqi@seu.edu.cn

Extended author information available on the last page of the article

1 Introduction

As a milestone work in the field of NLP, substantial work has shown that the pre-trained language models (PLMs) with self-supervised learning on large-scale corpora have learned the rich semantic knowledge that facilitates a variety of NLP downstream tasks [54]. Manning [1] explains that PLMs learn the meaning of words because meaning can be considered as a connection of linguistic forms, and PLMs have seen many connections of words to understand the meaning of words. For example, PLMs understand the word “Washington, D.C.” by these two sentences “Washington, D.C. is the capital city and federal district of the United States” and “Washington, D.C. is located on the east bank of the Potomac River”. As a result, PLMs significantly improve the performance of most NLP tasks through such learning.

Today, it has become a consensus to use PLMs, through fine-tuning or prompting, as the backbone of downstream tasks. With the proposed initial PLMs BERT [20] and GPT [49], the PLMs community began to flourish. Subsequently, various PLMs are proposed to address different needs and tasks. For example, some works propose knowledge distillation models such as DistilBERT [48] and TinyBERT [47] to reduce the number of parameters in PLMs and to increase the speed of training and inference. Some works present knowledge-enhanced models such as ERNIE [35] and KEPLER [36] to address knowledge-driven downstream tasks. However, there is a lack of a comprehensive study of the application of these PLMs to an important sub-task of NLP, namely knowledge graph question answering (KGQA).

KGQA aims to find answers to natural language questions from the knowledge graph (KG) which is typically store structured knowledge in the form of triples, denoted as (*subject, relation, object*). The study of various PLMs applied to KGQA is valuable for the following reasons.

1. The general domain KGQA is difficult to use in practice because of the efficiency issue. The huge size of the general domain KG¹ leads to a large training and inference time for the KGQA system. Some KGQA works limit the search range in KG to reduce complexity [40–42], yet the training time for a well-performing KGQA model [40] still exceeds 2 weeks. Without this search limitation, the model [40] would even take a few months to train as noted in [41]. Several works have attempted to reduce complexity by optimizing KGQA approaches [43–45], but at the expense of performance. In the recent KGQA system, PLMs have become a regular component of the system due to their obvious gains in performance. However, the application of powerful but large PLMs further increases the difficulty of using KGQA systems in practice. **Therefore, it is necessary to explore the trade-off between performance and efficiency of PLMs on KGQA.**
2. KGQA is a knowledge-intensive task and involves several common NLP subtasks² such as mention detection, entity disambiguation and relation detection as shown in Figure 1. There are substantial works to tackle these subtasks using PLMs as the cornerstone and with success [28–32]. Moreover, several works utilize structured knowledge to enhance PLMs for more than just self-supervised training on the large-scale corpus, with significant improvements in mention detection [33] and relation detection [34–37] tasks. Nevertheless, there is lack of work on making a comprehensive comparison on each subtasks from the perspective of PLMs. **Therefore, exploring the use of PLMs for KGQA is also instructive for these subtasks.**

¹ Freebase [38] contains over 3 billion triples across over 100 domains, while Google Knowledge Graph [39] has amassed over 500 billion triples.

² These tasks are similar to named entity recognition, entity linking and relation extraction.

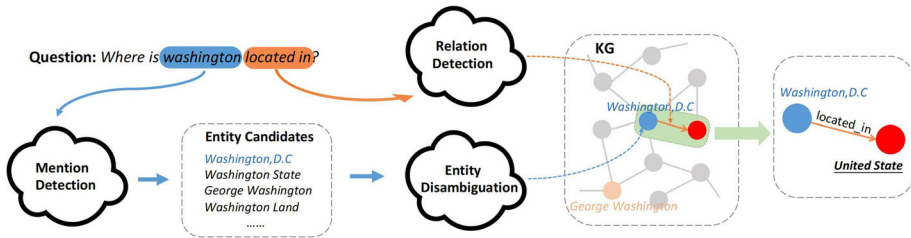


Figure 1 The general structure of KGQA

This work aims to evaluate the overall performance of various PLMs on KGQA comprehensively. We not only examine the **accuracy** and **efficiency** metrics of KGQA systems based on different PLMs, but also study their **scalability**³. Specifically, we designed four KGs of increasing size to explore the variation of these KGQA systems. **Three** classes of **nine** PLMs are used for evaluation, including the common PLMs BERT [20], ROBERTA [21], XLNET [50] and GPT2 [49], the lightweight PLMs ALBERT [22], DISTILBERT [48] and DISTILROBERTA [48], and the knowledge-enhanced PLMs LUKE [33] and KEPLER [36]. As common models serve as the backbone models for lightweight PLMs and knowledge-enhanced PLMs, we follow [54] to further classify them according to pre-trained task categories, namely, masked language modeling (i.e., BERT and ROBERTA), language modeling (i.e., GPT2), and permuted language modeling (i.e., XLNET). The investigation focused on the simple but common Simple KGQA task, which can be answered by a triple in KG. Moreover, we summarize two basic KGQA frameworks from previous works for the experiment. These two frameworks are vanilla without additional neural network modules except for PLMs and simple linear layers. It allows us to focus on comparing PLMs instead of various KGQA approaches with complex neural network modules. We also conduct experiments to compare the performance of these PLMs under fine-tuning with ChatGPT under zero-shot settings on the KGQA task.

In summary, our main contributions are as follows.

1. To the best of our knowledge, this is the first attempt to comprehensively study the overall performance of various PLMs in KGQA tasks. For this purpose, we summarize two basic KGQA frameworks from popular simple KGQA approaches to exclude the interference of complex neural network modules. We have implemented 18 KGQA systems based on these two basic KGQA frameworks using a total of nine PLMs. Further, we propose three KGQA benchmarks based on the popular SimpleQuestions benchmark. These four benchmarks have a linearly increasing KG scale.
2. We conduct comprehensive experiments to evaluate the overall results of all implemented KGQA systems on all benchmarks. We make detailed analyses regarding overall accuracy, efficiency and scalability from the perspective of different PLMs and KGQA frameworks. In addition, we further analyze the overall performance of the sub-modules of the KGQA systems to investigate the impact of the different PLMs and KGQA frameworks on these subtasks. We also compare the performance of these PLMs under fine-tuning with ChatGPT under zero-shot settings on three KGQA datasets.
3. We find that knowledge-distilled lightweight PLMs and knowledge-enhanced PLMs are promising for use in KGQA. This leads us to delve into this direction in the future to explore practical KGQA systems. Besides, we observe that ChatGPT has an excellent

³ Scalability is the measure of a system's ability to increase or decrease in performance and cost in response to changes in system processing demands. In our work, we explore the variation in accuracy performance and time cost with increasing KG size.

performance in KGQA tasks, while there are still some limitations. Section 7 summarizes all the important findings. Our KGQA frameworks based on PLMs provide new strong baselines of simple KGQA. We have released code and benchmarks as publicly accessible resources to help the future development of the KGQA community.

The rest of the article is structured as follows. In Section 2, we introduce related works on simple KGQA and PLM Applications On KGQA. In Section 3, we present the preliminary knowledge of this work. In Section 4, we summarise the existing simple KGQA methods and describe the two summarised KGQA basic frameworks in detail. In Section 5, we introduce four benchmarks and evaluation metrics. In Section 6, we describe the results and analysis of the experiment. Finally, Section 7 concludes this work and introduces future works.

2 Related works

2.1 Simple knowledge graph question answering

Knowledge graph question answering (KGQA) aims to find answers to natural language questions from the knowledge graph (KG). Simple KGQA means that a natural language question can be answered by a triplet in KG. The two mainstream branches of the current KGQA methods are information retrieval (IR) and semantic parsing (SP) [51–53]. The former attempts to retrieve answers directly from a subgraph centred on the topic entity and then models answer features for ranking. The latter tries to train a semantic parser to transform the question into intermediate logical forms and then execute it against KG. In simple KGQA, the IR method employs various neural networks to score the similarity between the question and each candidate fact in the subgraph and then find the best match. It follows the process of *retrieving* question-specific subgraph and then *ranking* the facts in it. Bordes et al. [24] used a memory network to encode questions and facts into the same vector space and score their similarity. Dai et al. [15] proposed a two-step conditional probability estimation problem and adopted a BiGRU network as an encoder. Yu et al. [7] designed two independent hierarchical residual BiLSTMs to represent questions and relations with different granularities. Yin et al. [11] used two independent models, a character-level CNN and a word-level CNN with attentive max-pooling. Lukovnikov et al. [9] proposed an end-to-end word/character-level encoding network for ranking subject-relation pairs and retrieving relevant facts. In simple KGQA, the SP method is simplified to a *classification* model because only a relation or a predicate needs to be generated. Ture and Jovic [6] employed a two-layer BiGRU model as a classifier. Petrochuk and Zettlemoyer [5] used a BiLSTM to classify relations and achieve state-of-the-art performance. Mohammed et al. [2] only adopted simple neural networks (i.e. LSTMs and GRUs) or non-neural network models (i.e. CRFs). In Section 4, we name IR method and SP method as the *retrieval and ranking-based* method and the *classification-based* method to show the differences more clearly.

2.2 PLM-based methods for KGQA

Pretrained language models have been widely served for various downstream tasks, including KGQA, due to the powerful representation capabilities learned from large-scale text corpora. For the IR method, PLMs provide a unified way to model unstructured text and structured KG information in a unified semantic space, which facilitates question-specific subgraph reasoning. Zhang et al. [59] trained a PLM-based path retriever to retrieve

hop-by-hop question-related relations. At each step, the retriever ranked the top-k relations based on the question and the relations selected in the previous step. Hu et al. [58] introduced PLM to help align questions and paths in a step-wise reasoning manner from explicit text semantic matching and implicit KG structure matching. Luo et al. [14] proposed a BERT-based model to preserve the original question-fact interaction information and reduce the semantic gap. For the SP method, PLMs significantly improve the understanding of questions, especially complex ones. Lukovnikov et al. [3] made the first attempt to use PLMs as classifiers to predict relations, with a significant performance improvement over shallow neural networks. In addition, Lukovnikov et al. demonstrated the greater advantage of PLMs on limited training data. Some works [60, 61] used PLMs to directly generate executable programs based on a given question and other relevant KG information. Substantial improvement in model performance demonstrates the effectiveness of such usages of PLMs. However, few KGQA works have taken into account the efficiency of PLMs. This is crucial for KGQA, which is inherently difficult to apply in practice.

3 Preliminaries

In this section, we introduce the definition of simple KGQA task (Section 3.1) and large-scale pre-trained language models (PLMs) (Section 3.2).

3.1 Task definition

This work focuses on evaluating PLMs on simple knowledge graph question answering, where the natural language question can be answered by a triple in KG. Simple questions are frequently queried in search engines and question-answering robots. The 100 most frequently asked questions on Google search engines in 2021 are simple questions⁴, and most of them can be answered by the KGQA system.

For ease of understanding, we define some notations used in this paper. Formally, a knowledge graph (KG) is typically a collection of subject-relation-object triples, denoted by $\mathbb{G} = \{(s, r, o) | s, o \in \mathbb{E}, r \in \mathbb{R}\}$, where (s, r, o) denotes that relation r exists between subject s and object o , \mathbb{E} and \mathbb{R} denote the entity set and relation set, respectively. Given the available KG \mathbb{G} , KGQA aims to answer natural language questions $Q = \{w_1, w_2, \dots, w_n\}$ in the format of a sequence of words with the answers $\mathcal{A}_q \subset \mathbb{E}$. For simple KGQA task, the answers directly connect to the topic entity and a KGQA system is trained using a dataset $D = \{Q, \langle s, r \rangle\}$, where $\langle s, r \rangle$ refers to a subject-relation pair.

In inference stage, given a natural language question Q “*The film Forrest Gump is directed by who?*” as shown in Figure 1, the KGQA system can answer this question by the answer *Robert Zemeckis* which is retrieved by the subject-relation pair $\langle \text{Forrest Gump}, \text{directed_by} \rangle$ in the KG \mathbb{G} .

3.2 Large-scale pretrained language models

Neural network-based language models represent everything through vectors of real numbers. They can learn better representations on a large corpus by back-propagating from the language model prediction task to the representation of words. Early work on language

⁴ Available online: <https://keywordtool.io/blog/most-asked-questions/> (accessed on 12 April 2022)

models trained shallow networks to capture the semantic meaning of words, such as Word2Vec [25] and GloVe [26]. However, they suffer from the drawback of not being able to represent polysemantic words in different contexts.

Since the introduction of Transformer [17], it has become feasible to train deep neural models for NLP tasks. Transformer is a more complex model than the simple neural networks previously explored by humans for word sequences. One of its main ideas is the attention mechanism, by which the representation of one location is computed as a weighted combination of representations from other locations. With Transformer as the architecture, various PLMs trained on large-scale corpora such as BERT [20] and RoBERTa [21] have been proposed with the goal of language model learning. Large-scale PLMs with hundreds of millions of parameters can learn polysemantic words as well as factual knowledge from contextual semantics. Furthermore, numerous works have proposed the use of structured knowledge to enhance PLMs, such as KEPLER [36], and others have used distillation techniques to reduce the number of PLM parameters, such as DistilBERT [48]. Most of them are based on improvements of the primary PLMs, i. e., BERT and RoBERTa. One such large-scale pre-trained language model can be deployed for many specific NLP tasks, requiring only a small number of further instructions. A standard approach is to fine-tune the model with a small amount of additional supervised learning. By fine-tuning large-scale PLMs, the rich linguistic knowledge of PLMs shows great performance on downstream NLP tasks. Recently, ChatGPT PLM, released by OpenAI, has gained huge attention from the NLP community and many other fields. ChatGPT is fine-tuned from the GPT-3.5 series models through reinforcement learning from human feedback [65]. Several works [62–64] have shown that ChatGPT demonstrates powerful capabilities on a lot of NLP tasks, but testing in knowledge-intensive downstream tasks is lacking. This work aims to explore the practicability of various PLMs on knowledge-intensive downstream tasks, i.e., knowledge graph question answering, to help researchers select the appropriate PLMs according to their needs. Unfortunately, ChatGPT currently only supports limited access ways and times, limiting our testing. We will leave more work about ChatGPT for the future.

4 Two basic KGQA frameworks

4.1 Summary of the framework

To analyze the practicality of PLMs applied to KGQA. We summarise several simple KGQA approaches and propose two basic KGQA frameworks for evaluation⁵, a Classification-based KGQA framework (KGQACL) and a Retrieval and Ranking-based KGQA framework (KGQARR). Previous works [2–6] belonging to KGQACL designed various deep neural networks to encode the question and then map the question vector to the KG relational dictionary. Previous works [7–16] belonging to KGQARR first retrieved adjacent relations (one-hop) of linked entities and then designed new network architecture or introduced contextual information to rank these relations. Some works also propose approaches such as utilize relation detection models to reorder entities [7] or adopt a joint training strategy [10, 13] to improve performance but make the KGQA framework more complex. Our frameworks do not

⁵ There is an existing KGQA approach based on KG embedding, which introduces knowledge representation learning, is proposed by [23] and is not included in our frameworks. This work focuses on comparing various PLMs, so the discussion of the effect of different KG embedding methods is reserved for future work.

consider these approaches as improving the performance of KGQA is not the purpose of this work.

Both basic frameworks consist of four modules, including (1) **Mention Detection**, (2) **Entity Disambiguation**, (3) **Relation Detection** and (4) **Answer Query**. The main difference between these two basic frameworks is the relation detection module. For KGQARR, this module is intended to rank candidate relations (i.e. information retrieval). For KGQACL, this module aims to map question intent to KG relations (i.e. semantic parsing). Mention detection and entity disambiguation are also regarded as two steps of the entity linking task. Existing studies on KGQA typically treat entity linking as an individual task to be handled in advance [53]. While KGQACL usually treats entity linking and relation detection as separate modules, KGQARR considers the whole process as a pipeline, with relation detection coming after entity linking.

4.1.1 Mention detection

Given a natural language question, the model will first find the mention representing the entity's name in that question. Previous works usually treated mention detection as a named entity recognition task and employed various models such as RNN, CNN and their variants [2, 5, 6, 12] or BERT [14] to solve it. Other work regards it as span detection task [3] or adopts CNN-LSTM as an encoder-decoder to generate entities directly [4].

4.1.2 Entity disambiguation

The detected mentions will be used to collect candidate entities, and these candidates will then be ranked. Several works employ n-gram heuristics approaches [2, 3, 6] to collect entities efficiently, and then different methods such as character similarity [2–4], TF-IDF scores [6] are employed for entity disambiguation. [5] disambiguate candidate entities by such a simple method as the score of connected relations. Our frameworks have adopted this simple method.

4.1.3 Relation detection

This module aims to obtain the correct relation in KG corresponding to the question. We summarise the two mainstream approaches, viz., KGQACL and KGQARR in this work. The former is based on the idea of classification and maps the questions directly into the KG relational dictionary as being independent of the previous modules. Previous works use various models like RNN [6], LSTM [2, 4, 5] and BERT [3] to encode sequences of questions, which are then classified into KG relation categories. The latter can be regarded as a similarity matching task, which will use linked entities to retrieve a set of candidate relations and then select the one with the highest similarity to the question. Various models [7, 8, 14, 15, 24], attention mechanism [11, 12] and external features such as context [16] and type [10] are designed to enhance the performance.

4.1.4 Answer query

Candidate entities and candidate relations with scores will be combined into pairs to query in KG to get the answers. The combination with the highest total score of the weighted sum of entities and relations is considered the correct pair [2–6]. Some works [9, 10, 12, 13] jointly train entity disambiguation and relation detection to select the pair with the highest model

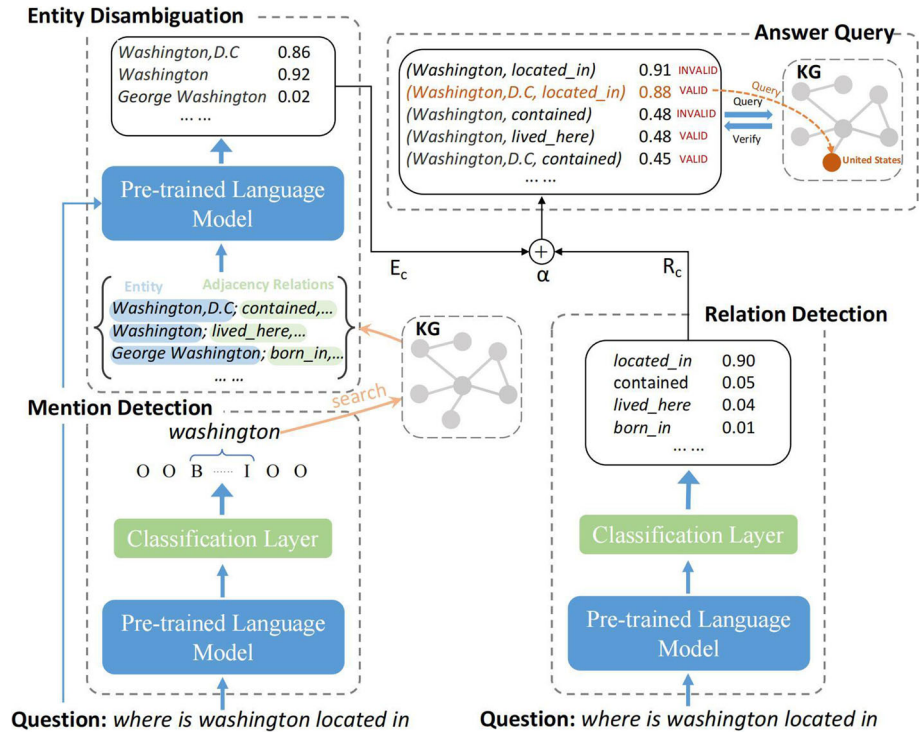


Figure 2 The basic classification-based KGQA framework

score. However, this approach cannot be implemented in our PLMs-based frameworks due to the limitations of the GPU⁶. So we do not consider this approach, which is difficult to ground because of its excessive hardware requirements.

Note that the two basic KGQA frameworks we summarised are vanilla and contain only PLMs and simple linear layers. This makes sense as it allows us to focus on comparing PLMs. Except for answer query, other modules are implemented based on PLMs. In addition, the modules are identical in both frameworks, excluding the relation detection. Next, we will detail these two PLMs-based KGQA frameworks.

4.2 The basic classification-based KGQA framework

The basic Classification-based KGQA framework (KGQACL) is shown in Figure 2. It consists of four modules described in Section 4.1, namely Mention Detection, Entity Disambiguation, Relation Detection and Answer Query.

4.2.1 Mention detection

Given a question Q , the goal of mention detection is to identify the subject mention m . For instance, the subject mention of the question in Figure 2 is “washington”. We treat this task as a common PLMs-based named entity recognition task. The sequence of question is

⁶ Our basic frameworks are trained using an NVIDIA GeForce RTX 2080 TI

encoded by PLMs and will then be fed into a linear classification layer. It will assign a label for each word in the question sequence, B for the beginning of mention, I for intermediate of mention, and O for non-mention. As PLMs adopt different tokenization methods for words, we only annotate the first token of each word and fill in the rest using the special character (pad).

4.2.2 Entity disambiguation

The mention m representing the entity name will be used to link to the grounded nodes in the KG. We pre-generate an inverted indexed dictionary which establishes a mapping of mentions to entities. We use m to look up the corresponding KG entities in the inverted index dictionary, which are regarded as candidate entities E_c . For instance, we obtain a set of candidate entities according to the mention “*washington*”, including the capital of the United States “*Washington, D.C.*”, the state “*Washington*” and the person “*George Washington*”. Besides, the adjacent relations r_{e_i} for each $e_i \in E_c$ retrieved from KG will be used for disambiguation. For example, the adjacent relations of the person “*George Washington*” are “*born_in*”, “*died_in*”, “*founded_organisation*”, etc.

Various PLMs are employed to score the entity $e_i \in E_c$, and the formula is $S_{e_i} = \mathcal{G}_{\text{PLM}}(Q|e_i|r_{e_i})$, where $|$ refers to the connection symbol, $\mathcal{G}_{\text{PLM}}()$ represents an PLM encoder. The loss function of the entity disambiguation model based on PLM is:

$$\mathcal{L}_{\mathcal{ED}} = -\log P(y = e_k), \quad (1)$$

$$P(y = e_k) = \frac{e^{S_{e_k}}}{e^{S_{e_k}} + \sum_{j=1}^N e^{S_{e_j}}}, \quad (2)$$

where e_k denotes the gold entity, N indicates the number of negative samples, e_j represents negative entities and $P(y = e_k)$ is the probability of e_k . In addition, we adopt a simple linguistic approach, fuzzy string matching, to initially rank E_c to select more challenging negative sample entities to train the model. We initial rank E_c according to the Levenshtein Distance score between the entity name and the mention m .

The entity set E_c with scores is obtained at the inference stage.

4.2.3 Relation detection

Relation detection is a PLMs-based classification task in this framework. Since simple KGQA only considers one-hop relations, a question Q corresponds to only one relation in KG. The model aims to map Q to a KG relation $r \in \mathbb{R}$. For instance, the question “*where is washington locate*” in Figure 2 corresponds to the relation “*located_in*” in KG. Specifically, PLMs are employed to encode the question sequence to obtain the vector h , and then h is fed into the linear classification layer to obtain the probability distribution of relations⁷. The goal of the model is to minimize:

$$\mathcal{L}_{\mathcal{RD}} = -\log P(y = \hat{r}|Q), \quad (3)$$

$$P(y = \hat{r}|Q) = \frac{e^{h_{\hat{r}}}}{\sum_{j=1}^M e^{h_{r_j}}}, \quad (4)$$

⁷ The dimension of h is 1×1 . Different PLMs obtain h in different ways, e.g. $h = w \cdot h_{[CLS]}^T$ in BERT.

where \hat{r} refers to the gold relation, $P(y = \hat{r} | Q)$ represents the probability of \hat{r} and M denotes to the number of relation categories. Finally, the relation set R_c with scores is obtained.

4.2.4 Answer query

This module does not involve any neural networks and aims to query the answer in KG using entity-relation pair. Given the set of candidate entities E_c and the set of candidate relations R_c obtained by the entity disambiguation and relation detection modules, we combine them into (e, r) pairs to be queried in KB, where $e \in E_c$ and $r \in R_c$. We rank each (e, r) pair, whose score is the weighted sum of its component scores, i.e., the entity disambiguation score and the relation detection score. The score of the (e, r) pair is

$$S_{(e,r)} = \lambda S_e + (1 - \lambda) S_r, \quad (5)$$

where $\lambda \in (0, 1)$, tuned according to the result of validation set. S_e and S_r are normalized entity score and relation score, respectively.

Note that the (e, r) pair may be invalid because such a combination does not exist in KG. We remove these pairs by querying and verifying them in KG. In addition, the popularity of entities is applied to further prune pairs for the same score. In our work, the popularity is derived from FACC1⁸ and the degree of entities.

4.3 The basic retrieval and ranking-based KGQA framework

The Retrieval and Ranking-based KGQA framework (KGQARR) is shown in Figure 3. It is a pipeline structure and consists of four modules, of which mention detection, entity disambiguation and answer query are identical to KGQACL, differing only in the relation detection module.

Different from KGQACL, the aim of relation detection in KGQARR is to select the relation in candidate relations R_c that has the highest semantic similarity score to the question pattern p . R_c consists of all adjacent relations searched by candidate entity E_c in KG. The question pattern p is obtained by using a special token $\langle e \rangle$ by mention m to replace the mention m in the question q ,

Following the way of Sentience-Bert [27], we employ two PLMs that share parameters to encode questions pattern and relations, respectively. This way of encoding significantly improves efficiency compared to cross-encoding. For each relation $r_i \in R_c$, we compute their similar score $Score(p, r_i)$. The final predicted relation \hat{r} is given by the following formula:

$$\hat{r} = \operatorname{argmax}_{r_i \in R_c} Score(p, r_i), \quad (6)$$

$$Score(q, r_i) = \cos(\operatorname{Pool}(h_p), \operatorname{Pool}(h_{r_i})), \quad (7)$$

where h_p and h_{r_i} are both obtained by PLMs, $\operatorname{Pool}()$ refers to the pooling layer. During training, we adopt the hinge loss to maximize the margin between the gold relation r^+ and the negative relation r^- in E_c .

$$\mathcal{L}_{RD} = \sum_{i=1}^k \max \{0, \gamma - Score(p, r^+) + Score(p, r_i^-)\}, \quad (8)$$

⁸ <http://lemurproject.org/clueweb12/FACC1/>

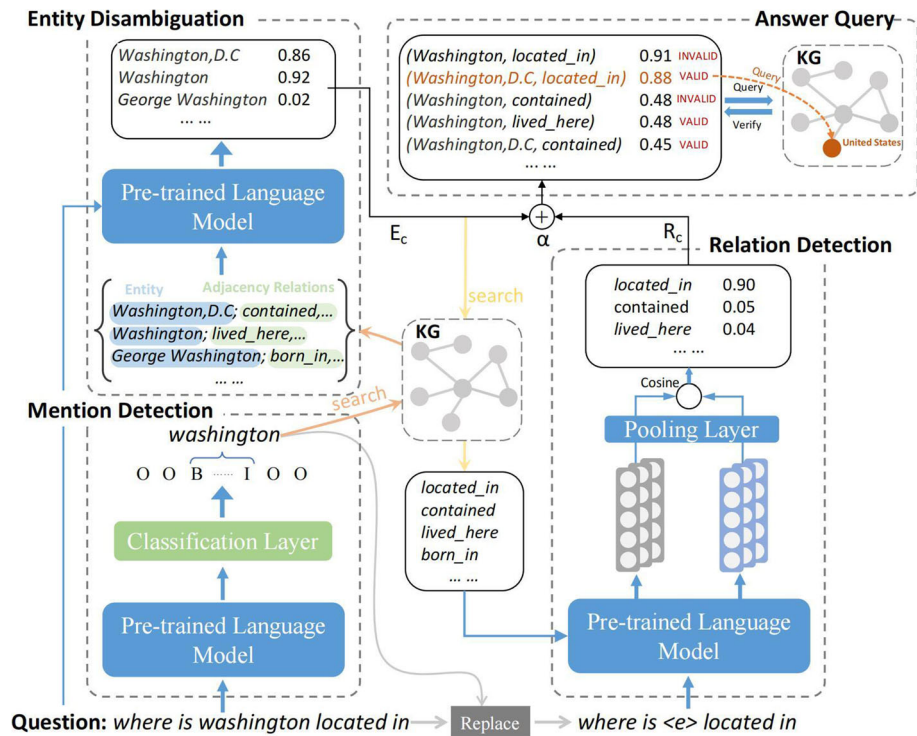


Figure 3 The basic retrieval and ranking-based KGQA framework

where γ is a constant parameter, where k is the number of negative relations. As with the KGQACL framework, the candidate relations R_c with scores and the candidate entities E_c with scores will be fed into the answer query module together.

5 Benchmarks

In this section, we will describe the four benchmarks utilized for the experiments and the method for constructing the benchmarks (Section 5.1). In addition, we introduce accuracy and efficiency evaluation metrics, as well as the method for evaluating the scalability of the KGQA system based on PLMs (Section 5.2).

5.1 Construction of the benchmarks

We construct experiments on four benchmarks. Apart from the popular simple KGQA benchmark SimpleQuestions [24], we construct three more benchmarks to explore the scalability of PLMs on KGQA. In particular, we increase the scale of the original KG of SimpleQuestion and propose the three KGQA benchmarks to investigate the performance changes of PLMs as the KG size increases. Note that the question-answering datasets of the four benchmarks are the same.

Table 1 The overall comparison of the KG for the four benchmarks

	Entities	Relations	Triples	Average Degree
SQS	1,951,909	6,701	7,188,636	88
SQM- A	61,558,084	11,127	105,003,132	176
SQM- B	90,282,560	12,227	202,002,048	293
SQL	108,183,882	12,794	292,088,464	407

Average degree represents the average number of adjacent relations of the entities appearing in the dataset

(a) The original SimpleQuestions with small-scale KG (SQS) [24]. The original benchmark contains more than 100,000 questions, divided into train/validation/test on a 7/1/2 split. The KG resource of this benchmark is FB2M, denoted as \mathbb{G}_S , which contains 2M entities and 6.7K relations. Some previous works pre-pruned KB to fit their methods because they assumed all questions were known. We do not preprocess KG to do experiments with all comparison models.

(b) SimpleQuestions with large-scale KG (SQL). Benchmark SQL requires getting a triple in a large-scale KG to answer questions. For the construction of the KG of SQL, denoted as \mathbb{G}_L , we retrieve all one-hop triples of the entities in FACC1 in the Freebase dump⁹. FACC1 provides the common names and the popularity of the entities. We then merge \mathbb{G}_{FACC1} with \mathbb{G}_S to obtain the KG \mathbb{G}_L : $\mathbb{G}_L = \mathbb{G}_{FACC1} \cup \mathbb{G}_S$. \mathbb{G}_L contains 108M entities, 12.7K relations and 292M triples, which completely covers \mathbb{G}_S .

Additionally, we construct two more benchmarks with KGs as \mathbb{G}_{M-A} and \mathbb{G}_{M-B} respectively. Their number of triples is between that of \mathbb{G}_S and \mathbb{G}_L , and the number of triples of these four KGs grows uniformly.

(c) SimpleQuestions with medium-scale KG_A (SQM- A). The KG \mathbb{G}_{M-A} of SQM- A includes 61M entities, 11.1K relations and 105M triples, which also completely covers \mathbb{G}_S .

(d) SimpleQuestions with medium-scale KG_B (SQM- B). The KG \mathbb{G}_{M-B} of SQM- B includes 90M entities, 12.2K relations and 202M triples, which also completely covers \mathbb{G}_S .

The overall comparison of the KG for the four benchmarks is shown in Table 1. Apart from the number of entities, relations and triples for the KG of each benchmark, we also count the average degree, i.e., the average number of adjacent relations of the entities appearing in the SimpleQuestions dataset. Average degree can reflect the challenge of the benchmark to some extent.

5.2 Evaluation metrics

We evaluate the overall performance of KGQA in terms of **accuracy** and **efficiency**. The accuracy metric follows the common evaluation method for SimpleQuestions, where we calculate the accuracy of inferred (s, r) pairs. Only a fact matching the ground truth answer in both subject \hat{s} and predicate \hat{r} is correct.

$$accuracy = \frac{\sum_{i=1}^N \mathbb{1}_{[(\hat{s}_i, \hat{r}_i) = (s_i, r_i)]}}{N}, \quad (9)$$

where N refers to the number of questions.

⁹ <https://developers.google.com/freebase>

We calculate the average training and test time of all KGQA systems to evaluate their efficiency. For a fair comparison, we set the same batch size and negative sampling number for each PLM in the same basic framework.

Additionally, we defined the Variation in Accuracy (VA) and the Variation in average test Time (VT) to evaluate the **scalability** of the KGQA system.

Definition 1 (VA) VA of a KGQA system on a benchmark represents the variation in the accuracy of the system on that benchmark compared to the accuracy of the system on the benchmark KGs. VA of a KGQA system in a benchmark KG_x represents the gap between the accuracy of the system on KG_x and the accuracy of the system on small-scale KG benchmarks KGs. $VA = accuracy_{KGs} - accuracy_{KG_x}$, where KG_x represents one of our four benchmarks. A **higher** VA means that the system performs worse in terms of scalability.

Definition 2 (VT) VT of a KGQA system on a benchmark KG_x represents the change in the average test time of the system on KG_x compared to the average test time of the system on small-scale KG benchmarks KGs. $VT = time_{KG_x} - time_{KGs}$, where KG_x represents one of our four benchmarks. A **higher** VT means that the system performs worse in terms of scalability.

6 Experiments

In this section, we first present all the PLMs-based KGQA systems (Section 6.1) and the experimental setup (Section 6.2). We then show the overall experimental results and discuss the results in light of the three research questions (Section 6.3). We further explore the sub-modules of KGQA and discuss them according to the three new research questions (Section 6.3). Besides, we also evaluate all systems on two other KGQA datasets apart from the SimpleQuestions family (Section 6.5). Finally, we compare the performance between ChatGPT and other PLMs on datasets SimpleQuestions, WebQuestionSP and FreebaseQA (Section 6.6).

6.1 All KGQA systems based on PLMs

In this work, 18 KGQA systems (9 PLMs * 2 basic KGQA frameworks) were implemented for evaluation. **Three** classes of **nine** PLMs were used for evaluation, including the common large-scale PLMs BERT¹⁰, ROBERTA¹¹, XLNET¹² and GPT2¹³, the lightweight PLMs ALBERT¹⁴, DISTILBERT¹⁵ and DISTILROBERTA¹⁶, and the knowledge-enhanced PLMs LUKE¹⁷ and KEPLER¹⁸. As common models serve as the backbone models for lightweight PLMs and knowledge-enhanced PLMs, we follow [54] to further classify the common PLMs

¹⁰ <https://huggingface.co/bert-base-uncased>.

¹¹ <https://huggingface.co/roberta-base>.

¹² <https://huggingface.co/xlnet-base-cased>.

¹³ <https://huggingface.co/gpt2>.

¹⁴ <https://huggingface.co/albert-base-v2>.

¹⁵ <https://huggingface.co/distilbert-base-uncased>.

¹⁶ <https://huggingface.co/distilroberta-base>.

¹⁷ <https://huggingface.co/studio-ousia/luke-base>.

¹⁸ <https://github.com/THU-KEG/KEPLER>.

according to pre-trained task categories, namely, Masked Language Modeling (MLM, i.e., BERT and RoBERTa), Language Modeling (LM, i.e., GPT2), and Permuted Language Modeling (PeLM, i.e., XLNET). Parameters of these PLMs¹⁹ are shown in Table 2.

BERT BERT is the most representative pre-trained language model that uses the encoder of the deep Transformer as its backbone. BERT uses Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) as self-supervised tasks for pretraining.

RoBERTa RoBERTa has almost the same architecture as BERT, while it differs in the parameter settings and training objectives. RoBERTa removes the NSP loss and creates the dynamic MLM mask instead of the static mask used in BERT to train a larger scale and a longer sequence model.

GPT2 Unlike BERT and Roberta, which are all masked language models, GPT2 is an autoregressive language model predicting one token at a time from left to right (i.e. LM). GPT2 is often used for natural language generation, whereas BERT and Roberta are mainly used for natural language understanding.

XLNET XLNET is known as a permuted language model [54]. Unlike GPT2 can not utilize the context from the backward side, XLNET resolves this problem by adopting a new objective called Permutation Language Modeling (PeLM), enabling the model to take advantage of both forward and backward contexts.

ALBERT ALBERT is a lite version of BERT. All its Transformer blocks share parameters and its embedding matrix is decomposed into two smaller matrices. Thus ALBERT has a much smaller number of parameters than BERT. Instead of NSP, Albert predicts the order of two consecutive text segments.

DistilBERT DistilBERT is a distilled version of BERT that is pre-trained on the same corpus in a self-supervised manner, using the BERT model as a teacher. This means that it only pre-trains on raw texts, with no humans labeling them in any way.

DistilRoBERTa DistilRoBERTa is a distilled version of RoBERTa. It follows the same training procedure as DistilBERT.

LUKE LUKE is based on RoBERTa and adds entity embeddings as well as an entity-aware self-attention mechanism. The entity-aware self-attention mechanism is an extension of the self-attention mechanism of the Transformer and considers the types of words or entities when computing attention scores.

KEPLER KEPLER is a unified model for knowledge embedding (KE) and PLM representation. It encodes textual entity descriptions with a PLM and then optimizes KE and language modeling objectives jointly.

6.2 Experimental setup

All KGQA systems were trained using an NVIDIA GeForce RTX 2080 TI. We performed a grid search for all KGQA systems, choosing the hyperparameter configuration that achieves the highest final accuracy. We adopted an early stop strategy in training and set the patience to 3. Because the PLMs-based entity disambiguation model is too large, the batch size can only be set to 1. Nevertheless, this may cause some PLMs to be difficult to converge, so we used the gradient accumulation method to increase the gradient accumulation step instead of increasing the batch size. Note that the value of the batch size multiplied by the gradient accumulation step is the same for all PLMs in the same sub-module task to ensure a fair comparison of their training times. In addition, KGQA systems output Top-50 results for entity disambiguation and Top-5 results for relation detection to combine (subject, relation) pairs.

¹⁹ These data are from <https://huggingface.co/>.

Table 2 Parameters of various PLMs

PLMs	BERT	ROBERTA	XLNET	GPT2	ALBERT	DISTILBERT	DISTILROBERTA	LUKE	KEPLER
Parameters	110M	125M	117M	125M	11M	66M	82M	253M	125M

6.3 Overall results and discussions

In this section, we show the overall results of the 18 KGQA systems based on various PLMs in terms of accuracy and efficiency (i.e., average training time and average testing time). Based on these experimental results, we will discuss three questions: (1) What PLMs have the best accuracy or efficiency performance? (2) What are the differences in accuracy and efficiency between the two basic KGQA frameworks? (3) How scalable are the various PLMs, i.e., how do their accuracy and efficiency vary as the size of KG increases?

6.3.1 Discussion on the accuracy of KGQA systems

The accuracy performance of all the studied PLMs-based KGQA frameworks and benchmarks are summarised in Table 3.

The bold numbers in Table 3 represent the highest accuracy of all KGQA system results using one benchmark. ROBERTA and the two knowledge-enhanced PLMs LUKE and KEPLER achieves the best accuracy results. Both LUKE and KEPLER are based on ROBERTA for knowledge augmentation. The results demonstrate the powerful performance of ROBERTA and that knowledge enhancement is beneficial for knowledge-intensive tasks, i.e., KGQA. LUKE and KEPLER perform better on small-scale KG benchmarks while ROBERTA perform better on large-scale KG benchmarks. This may be due to the fact that the introduction of additional knowledge-enhancing pre-training objects affects the robustness of the model itself. Comparing the performance of the same PLM on the two basic frameworks, we found that the KGQARR framework significantly outperforms the KGQAACL framework on small-scale KG benchmark SQS. However, as the size of KG increases, the accuracy of KGQARR becomes inferior to that of KGQAACL. We will investigate the reason for this in Section 6.4 by analyzing the performance variation of their submodules. In addition, we noted that XLNET and GPT2 are worse than the other PLMs in terms of accuracy in almost all settings, even for the three lightweight PLMs. In particular, the GPT2-based KGQARR system performed extremely poorly in all benchmarks. We believe that the modelling way of PLMs influences it. All PLMs are modelled by the auto-encoding way (i.e. MLM) except XLNET and GPT2. GPT2 is modeled via auto-regressive way (i.e. LM), while XLNET combines the idea of auto-encoding with auto-regressive modeling (i.e. PeLM). Auto-encoding modelling is skilled in solving natural language understanding (NLU) tasks, while auto-regressive modelling is proficient in solving natural language generation (NLG) tasks. Therefore, XLNET and GPT2 perform poorly on KGQA because traditional approaches treat it as an NLU task to solve. There has been some recent work [46] to convert KGQA to the NLG task for handling, and we will explore this approach in the future.

We investigated the scalability of the KGQA system with two metrics, VA and VT, defined in Section 5.2. As shown in Figure 4, the VA of all PLMs-based KGQA systems under both basic frameworks shows an increasing trend, indicating that scalability gradually worsens as KG size increases. We excluded the analysis of GPT2 due to its terrible accuracy. Among the KGQAACL and KGQARR frameworks, XLnet exhibits the worst scalability performance, especially for the benchmark SQL with the largest KG size. In addition, the knowledge-enhanced PLMs LUKE and KEPLER perform inferiorly to other PLMs in terms of scalability on the larger scale KG benchmarks (SQM- B and SQL). In contrast, the lightweight PLMs ALBERT, DISTILBERT and DISTILROBERT are more robust to KG scale variations and perform better in terms of scalability. We will further analyze which sub-modules in the framework primarily affect scalability in Section 6.4.

Table 3 Overall accuracy (%) of different PLMs-based KGQA systems on four benchmarks

Frameworks PLMs	Benchmarks	KGQA _{CL}				KGQA _{RR}			
		SQS	SQM- A	SQM- B	SQL	SQS	SQM- A	SQM- B	SQL
<i>Common PLMs</i>									
BERT(<i>MLM</i>)		74.79	68.02	66.03	64.81	76.76	67.75	66.54	65.39
ROBERTA (<i>MLM</i>)		75.32	68.97	66.33	65.63	77.12	68.20	66.81	65.66
XLNET (<i>PeLM</i>)		74.30	66.47	65.31	59.92	76.26	66.41	65.40	58.55
GPT2 (<i>LM</i>)		73.39	66.52	64.53	64.30	23.18	20.04	19.88	19.75
<i>Lightweight PLMs</i>									
ALBERT		74.26	67.01	65.37	64.66	76.20	66.80	65.60	65.10
DISTILBERT		74.29	67.87	65.45	64.36	75.11	68.07	65.55	64.46
DISTILROBERTA		74.04	67.32	64.67	64.39	73.24	67.37	65.48	64.00
<i>Knowledge-enhanced PLMs</i>									
LUKE		75.67	69.25	65.26	64.86	77.00	67.97	65.11	64.26
KEPLER		74.87	68.41	65.23	65.15	77.41	68.41	66.20	64.94

Bolded numbers indicate the highest accuracy

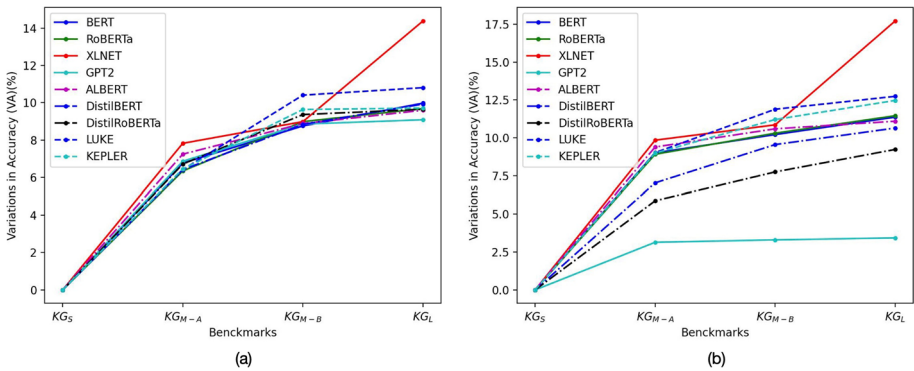


Figure 4 Scalability of all KGQA frameworks (a) and KGQA frameworks (b) in terms of accuracy variation

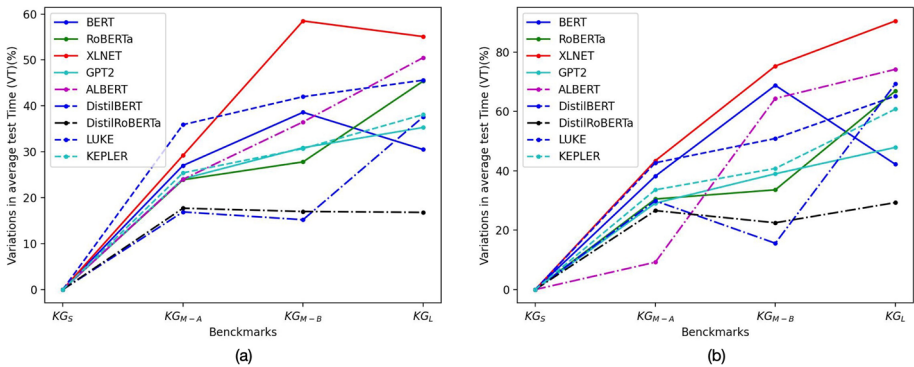


Figure 5 Scalability of all KGQA frameworks (a) and KGQA frameworks (b) in terms of average test time variation

6.3.2 Discussion on the efficiency of KGQA systems

The efficiency performance of all the studied PLMs-based KGQA frameworks and benchmarks are summarised in Table 4. We set the same patience for all KGQA systems so that each model was trained to converge. Due to the different convergence rates of the different PLMs, the variation of the training time of the PLMs did not coincide with the variation of the testing time.

As shown in Table 4, the two lightweight PLMs DISTILBERT and DISTILROBERTA exhibit the highest efficiency in training and testing. DISTILBERT is up to 3.1x faster than XLNET on training (2325.5 ms vs. 749.7 ms), and DISTILBERT is up to 3.2x faster than XLNET on testing (85.3 ms vs. 26.3 ms). Another lightweight model, ALBERT, has the least number of parameters, but its efficiency does not have any advantage over other PLMs. Therefore, the knowledge distillation approach is an effective way to improve efficiency, while the matrix parameter sharing strategy only reduces GPU memory consumption, with no improvement in efficiency. In addition, the time consumption of all PLMs tends to increase as the size of KG increases. Comparing the efficiency of different basic frameworks for the same PLM, KGQARR is always more time-consuming than KGQACL. According to the analysis in Section 6.3.1, the KGQARR framework is only more accurate than the KGQACL framework for small-scale KG benchmark SQS. Therefore, the KGQACL framework is a better choice for large-scale KG benchmarks.

The efficiency scalability of all PLMs based on both frameworks is shown in Figure 5. All PLMs show an increasing trend in VT as the KG size increases. Among them, DISTILROBERTA has the best scalability as it has the smallest VT on all benchmarks. Another lightweight PLM, DITILBERT, also shows good scalability. Section 6.3.1 also demonstrates DISTILROBERTA and DITILBERT have the equivalent accuracy performance as the other PLMs on large-scale KG. These findings indicate that the two knowledge distillation PLMs have excellent scalability. The knowledge distillation is a promising approach for PLMs applied to KGQA.

6.3.3 Summary and new research questions

Some important conclusions can be drawn from the above discussion. ROBERTA, LUKE and KEPLER perform best in terms of overall accuracy. Nevertheless, Luke and Kepler have slightly poorer scalability, with a greater variation in accuracy as the KG size increases. The two lightweight PLMs DISTILBERT and DISTILROBERTA exhibit the best scalability in accuracy and efficiency. Their accuracy on large-scale KG is the same as other PLMs, and their inference time is up to 3.3x faster than other PLMs. For the KGQA framework, the KGQARR framework is significantly less efficient than the KGQACL framework. Furthermore, the higher accuracy of KGQARR-based systems than KGQACL-based systems is only at small KG scales. As the KG size increases, the KGQACL-based system gradually outperforms the KGQARR-based system, which indicates the poor scalability of the KGQARR framework.

These findings lead us to explore the following questions further. (1) What sub-modules of the KGQA system are primarily responsible for the differences in accuracy and efficiency? (2) What sub-modules in the KGQA system are most susceptible to the variation in the size of KG? (3) Why does the KGQARR framework have worse scalability? We explore these three questions by examining the performance of the submodules of all KGQA systems.

Table 4 Efficiency of all PLMs-based KGQA frameworks on four benchmarks

Frameworks Benchmarks PLMs	KGQACL				KGQARR			
	SQS	SQM-A	SQM-B	SQL	SQS	SQM-A	SQM-B	SQL
<i>Common PLMs</i>								
BERT (MLM)	Tr	489.1	1042.8	584.5	1296.8	1190.4	2538.0	1563.7
	Te	45.9	84.5	76.4	70.4	108.6	139.2	112.6
ROBERTA (MLM)	Tr	523.6	701.0	772.6	1450.1	1341.8	1470.2	1906.1
	Te	45.8	73.6	91.2	70.0	100.5	103.6	136.9
XLNET(PeLM)	Tr	850.4	631.3	655.6	2325.5	2500.7	2308.9	1685.3
	Te	85.3	143.8	140.4	117.5	160.9	192.8	208.0
GPT2 (LM)	Tr	467.2	484.9	502.1	854.2	1341.1	1078.0	1076.9
	Te	48.4	79.3	83.7	75.6	104.7	114.6	123.5
<i>Lightweight PLMs</i>								
ALBERT	Tr	661.8	581.8	896.6	2679.6	1609.0	1500.8	1969.8
	Te	54.9	91.4	105.4	105.6	114.8	170.0	179.8
DISTILBERT	Tr	452.3	553.2	463.3	749.7	970.7	1169.4	1120.8
	Te	26.3	41.5	63.9	48.4	78.3	64.0	117.7
DISTILROBERTA	Tr	477.5	871.4	901.0	801.7	1422.6	1504.1	1290.6
	Te	28.0	45.0	44.8	42.9	69.5	65.4	72.2
<i>Knowledge-enhanced PLMs</i>								
LUKE	Tr	462.7	590.5	634.1	1738.6	2430.3	2213.9	1609.2
	Te	41.5	83.5	87.1	67.9	110.6	118.8	133.1
KEPLER	Tr	960.5	753.4	1130.7	2164.4	2522.4	2100.8	2011.5
	Te	48.4	79.1	86.5	69.0	102.6	109.8	129.8

Tr denotes average training time (ms) and Te denotes average test time (ms). Underlined ones indicate the shortest time spent

Table 5 Results of mention detection of all PLMs-based KGQA systems

PLMs	Precision	Recall	F1	Te
<i>Common PLMs</i>				
BERT (<i>MLM</i>)	95.12	95.50	95.31	2.50
ROBERTA (<i>MLM</i>)	94.62	95.77	95.21	2.27
XLNET (<i>PeLM</i>)	95.05	95.28	95.16	4.00
GPT2 (<i>LM</i>)	85.85	88.15	86.98	<u>0.73</u>
<i>Lightweight PLMs</i>				
ALBERT	94.32	94.68	94.51	2.43
DISTILBERT	94.58	95.37	94.97	1.15
DISTILROBERTA	94.41	94.92	94.66	1.18
<i>Knowledge-enhanced PLMs</i>				
LUKE	94.02	94.40	94.21	1.22
KEPLER	94.60	95.01	94.80	2.63

Te refers to the average test time (ms). Bolded numbers indicate the best performance and underlined ones indicate the shortest time spent

6.4 Study of the KGQA sub-modules

6.4.1 Results and discussion on KGQA sub-modules

We further compared the sub-module performance of each KGQA system in this section to explore the primary influencers of accuracy and efficiency for each PLMs. Additionally, we compared the two KGQA base frameworks to explore the reasons for their large variability.

Tables 5, 6 and 7 show the overall results for Mention Detection (MD), Entity Disambiguation (ED) and Relation Detection (RD) respectively. For efficiency, we only compare the average test time. We do not analyze Answer Query further as it is irrelevant to PLMs. The final result of MD is not affected by the KGQA basic frameworks and benchmarks, and it is only relevant to PLMs. Table 5 shows that all PLMs except GPT2 have similar accuracy and efficiency on MD. It indicates that GPT2, based on auto-regressive modelling (i.e. LM), is not good at solving NER tasks. Bert has the highest F1 value but poor efficiency. Notice that BERT's distilled version DISTILBERT improves efficiency by almost double and has only a slight performance penalty.

The ED of both frameworks is the same. As shown in Table 6, ROBERTA exhibits the best accuracy performance, and DISTILBERT and DISTILROBERTA have the shortest test time. It is worth to note that the accuracy and efficiency of all PLMs in the ED task are greatly affected by the KG size. This is because as the KG size increases, the number of candidate entities and the degree of entities increases, as shown in Table 1. But, as the KG size gets larger, the impact on accuracy becomes smaller. XLNET shows the most severe decrease in accuracy (27.59% decrease), and ALBERT shows the most significant increase in test time (50.3ms increase).

As shown in Table 7, there are significant differences in performance on RD between the two KGQA frameworks, which leads to differences in the final accuracy and efficiency of the two frameworks. All rows in Table 7 show that the KGQACL is more efficient than the KGQARR because KGQARR needs to encode all candidate relations to and questions to calculate similarity, whereas KGQACL only needs to encode questions. This is also why the

Table 6 Results of entity disambiguation of all PLMs-based KGQA systems on four benchmarks

PLMs		Benchmarks			
		SQS	SQM- A	SQM- B	SQL
<i>Common PLMs</i>					
BERT (<i>MLM</i>)	Top@1	82.26	73.58	72.79	71.39
	Te	41.8	67.6	80.2	72.1
ROBERTA (<i>MLM</i>)	Top@1	82.57	75.32	73.85	71.80
	Te	42.0	65.8	69.3	87.3
XLNET (<i>PeLM</i>)	Top@1	82.46	69.12	67.75	54.87
	Te	78.5	107.6	136.8	87.3
GPT2 (<i>LM</i>)	Top@1	80.17	71.49	69.4	69.18
	Te	46.5	70.4	77.0	81.5
<i>Lightweight PLMs</i>					
ALBERT	Top@1	82.0	73.21	71.3	71.43
	Te	50.6	74.3	86.9	100.9
DISTILBERT	Top@1	81.80	72.59	70.41	67.7
	Te	<u>24.3</u>	<u>41.1</u>	<u>39.0</u>	61.6
DISTILROBERTA	Top@1	81.59	72.75	68.76	69.14
	Te	26.0	44.5	42.7	<u>42.6</u>
<i>Knowledge-enhanced PLMs</i>					
LUKE	Top@1	81.85	74.27	67.45	69.08
	Te	38.9	74.5	80.6	83.9
KEPLER	Top@1	82.53	74.43	69.59	71.33
	Te	44.0	69.2	74.5	81.6

Top@1 denotes top-1 recall. Te refers to the average test time (ms). Bolded numbers indicate the best performance and underlined ones indicate the shortest time spent

increase in KG size significantly affects the accuracy and efficiency of KGQARR, yet it does not affect KGQACL. Although the accuracy of KGQARR is significantly higher than that of KGQACL on small-scale KG benchmark SQS, the former is less scalable than the latter. In addition, the knowledge-enhanced PLMs LUKE and KEPLER show the highest accuracy performance, which indicates the effectiveness of the knowledge-enhanced approach.

In general, both ED and RD modules significantly impact the final accuracy. ED and RD based on the KGQARR framework have a primary effect on the final efficiency, and they are most susceptible to changes in KG size. KGQARR has worse scalability than KGQACL due to their different approaches to solving RD.

6.4.2 Entity disambiguation using the vanilla method

The analysis in Section 6.4.1 demonstrates that PLMs-based entity disambiguation takes up the most time in the whole KGQA system. Given the high computational complexity of PLMs, we attempt to solve entity disambiguation using a vanilla method without any neural networks. We use only a simple linguistic approach, fuzzy matching, to rank all candidate entities (mentioned in Section 4.2.2). Specifically, we rank all candidate entities according to the Levenshtein Distance score between the entity name and the subject mention.

Table 7 Results of relation detection of all PLMs-based KGQA systems on four benchmarks

Frameworks PLMs	Benchmarks	KGQACL				KGQARR			
		SQS	SQM-A	SQM-B	SQL	SQS	SQM-A	SQM-B	SQL
<i>Common PLMs</i>									
BERT (MLM)	Top@1	81.22	80.79	80.50	80.43	85.14	81.86	81.69	81.02
	Te	1.5	1.6	1.5	1.5	26.1	33.4	56.3	37.8
ROBERTA (MLM)	Top@1	81.38	81.40	81.40	81.16	84.84	82.35	81.41	81.07
	Te	1.4	1.6	1.4	1.5	25.7	32.2	30.4	47.1
XLNET (PeLM)	Top@1	81.26	81.00	81.07	80.83	85.11	82.88	81.68	80.52
	Te	2.7	2.7	2.8	1.5	35.0	49.2	51.2	70.4
GPT2 (LM)	Top@1	81.19	81.13	81.10	81.70	17.77	16.24	15.92	15.09
	Te	1.1	1.3	1.3	1.5	28.3	33.6	34.1	37.2
<i>Lightweight PLMs</i>									
ALBERT	Top@1	79.64	79.33	79.59	78.96	83.86	80.95	79.88	79.23
	Te	1.9	1.9	1.9	2.0	52.6	38.0	80.5	76.2
DISTILBERT	Top@1	81.57	81.11	80.97	81.52	85.41	82.59	81.74	81.33
	Te	<u>0.7</u>	0.9	<u>1.0</u>	<u>0.8</u>	<u>14.5</u>	<u>19.4</u>	23.6	54.6
DISTILROBERTA	Top@1	81.05	80.68	80.67	81.30	81.31	82.69	82.06	80.62
	Te	0.8	<u>0.8</u>	<u>1.0</u>	0.9	15.7	20.7	<u>21.4</u>	<u>28.2</u>
<i>Knowledge-enhanced PLMs</i>									
LUKE	Top@1	81.77	82.12	82.22	81.74	85.50	83.00	82.00	80.60
	Te	1.2	1.5	1.5	1.7	27.7	34.9	37.0	47.9
KEPLER	Top@1	81.12	81.09	80.96	80.75	85.60	83.00	81.70	80.30
	Te	1.8	1.8	1.8	2.0	22.4	30.7	32.6	45.6

Top@1 denotes top-1 recall. Te refers to the average test time (ms)



Figure 6 Comparison of top-1 recall of entity disambiguation and final accuracy results for various PLMs (a-i) based methods and vanilla method. Van ED and Van Acc denote entity disambiguation and the whole KGQA system using vanilla method, and ED denotes entity disambiguation and the whole KGQA system using PLM

Table 6 shows the comparison results regarding the performance of vanilla methods and PLMs on entity disambiguation on all benchmarks, and their impact on the final accuracy. Compared to the vanilla method, all PLMs significantly improved the performance of entity disambiguation on all benchmarks (Van ED vs. ED). However, the improvements in the final accuracy of PLMs are not as significant in most cases (Van Acc vs. Acc). This is because the answer query module performs a weighted combination of candidate entities with scores and candidate relations with scores, which also screens out ambiguous entities to some extent. It is worth noting that XLNET improves entity disambiguation on the large-scale KG benchmark KGL (Figure 6(c)), yet is worse than the vanilla method in terms of final accuracy. This is because the large-scale KG contains too many noisy relations, leading the XLNET-based entity disambiguation model to assign a lower score to the ambiguous entities that some answer query modules can filter out. In addition, Figure 7 shows the efficiency of the vanilla method compared to that of PLMs on entity disambiguation, the vanilla method takes much less time than PLMs. Therefore, PLMs-based entity disambiguation is time costly and has limited improvement in the final accuracy of KGQA. More importantly, using the XLNET-based entity disambiguation model can even reduce the final accuracy on large-scale KG benchmark KGL.

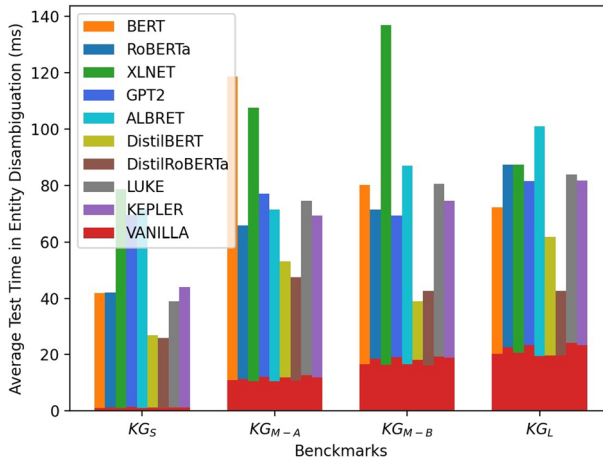


Figure 7 The average test time of the various PLMs-based methods compared to the vanilla method used to resolve entity disambiguation in all benchmarks

Table 8 Overall accuracy and efficiency of different PLMs-based KGQA systems on WebQuestionSP (WB) and FreebaseQA (FBQ)

Frameworks		KGQA _{CL}		KGQA _{RR}	
PLMs	Benchmarks	WQ	FBQ	WQ	FBQ
		<i>Common PLMs</i>			
BERT (<i>MLM</i>)	Acc	61.76	40.12	64.11	41.48
	Te	44.5	44.9	72.2	72.4
ROBERTA (<i>MLM</i>)	Acc	62.32	40.40	64.59	41.97
	Te	44.4	44.9	72.2	72.3
XLNET (<i>PeLM</i>)	Acc	61.46	39.80	63.87	41.15
	Te	85.1	85.1	119.8	117.7
GPT2 (<i>LM</i>)	Acc	60.49	39.03	18.11	5.09
	Te	48.3	48.5	76.3	76.5
<i>Lightweight PLMs</i>					
ALBERT	Acc	61.47	39.83	63.89	41.16
	Te	55.1	55.3	104.9	104.9
DISTILBERT	Acc	61.49	39.84	63.59	40.88
	Te	<u>26.7</u>	<u>26.9</u>	48.4	48.5
DISTILROBERTA	Acc	61.05	39.43	62.57	39.36
	Te	28.9	28.0	<u>43.2</u>	<u>43.0</u>
<i>Knowledge-enhanced PLMs</i>					
LUKE	Acc	62.31	40.62	64.52	42.08
	Te	41.2	41.3	67.5	67.7
KEPLER	Acc	62.02	40.29	64.46	42.02
	Te	48.8	48.8	68.8	69.0

Acc denotes accuracy (%). Te refers to the average test time (ms). Bolded numbers indicate the highest accuracy. Underlined ones indicate the shortest time spent

6.5 Validation beyond the simple questions benchmarks

In addition to the four benchmarks of the SimpleQuestions family (Section 5.1), we evaluated the accuracy and efficiency of all systems on the WebQuestionSP [57] and FreebaseQA [56] datasets. Both datasets adopt the large-scale KG, Freebase, as the resource and include a high proportion of simple questions (71.3% in WebQuestionSP and 66.4% FreebaseQA). As these two datasets include numerous questions with multi-hop paths or multiple constraints, such as “*What character did Natalie Portman play in Star Wars?*”, we followed [55] to pre-process these two datasets²⁰. Specifically, we kept only simple questions that can be answered by a triple and questions with entities or predicates within FB2M.

Table 8 demonstrates the overall accuracy and efficiency results for all systems on WebQuestionSP and FreebaseQA. Based on these results, we get similar conclusions to the experiments on SimpleQuestions. ROBERTA and LUKE have the best performance. GPT2 perform the worst in terms of performance, especially in the KGQARR framework. Almost all PLMs based on the KGQARR framework have higher accuracy performance than those based on the KGQACL framework but are more time-consuming. The two distillation-based PLMs, DISTILBERT and DISTILROBERTA, are far more efficient than the other PLMs. Furthermore, all systems performed poorly on FreebaseQA, with even the best, LUKE, only achieving 42.08% accuracy. After performing an error analysis, we found that FreebaseQA contained many mislabelled and unanswerable questions.

6.6 ChatGPT for zero-shot KGQA

We conducted experiments to compare the performance between ChatGPT²¹ and other PLMs on 300 sampling questions from SimpleQuestions, WebQuestionSP and FreebaseQA. Note that ChatGPT was under the zero-shot KGQA setting, while other PLMs were fine-tuned using the training set in a better-performing framework (KGQACL framework with GPT2 and KGQARR frameworks with the other PLMs). The input of ChatGPT consists of the instruction (“*Please answer the given question based on the context. The answers should be factual answers.*”) and the question, inspired by [62]. After reading the entire input, the model generates the answer in the form of a piece of text²². For each question, the answers generated by ChatGPT were evaluated and cross-validated by two professionals with reference to gold answers.

Table 9 demonstrates that ChatGPT outperforms other PLMs by up to 12% on WebQuestionSP and far surpasses other PLMs by up to 54% on FreebaseQA. However, ChatGPT performs miserably on SimpleQuestions with an accuracy of only 29.3%. We speculate that the discrepancy is caused by the different construction methods of these datasets. WebQuestionSP was derived from Google Suggest API, while FreebaseQA was scraped from trivia and quiz-league websites, which are still accessible. In contrast, SimpleQuestions were constructed by humans based on Freebase triples. Therefore, it is possible that ChatGPT has seen these questions or related texts due to its extremely large training corpus.

We further categorized the error cases of ChatGPT on SimpleQuestions as shown in Table 10. We consider the ENUMERATION type questions, which account for 38.2%, as a

²⁰ The pre-processed datasets are available at <https://github.com/aistairc/simple-qa-analysis>.

²¹ The version of ChatGPT is Jan 30 Version, and the user’s access times are limited. We have released the script for accessing ChatGPT.

²² We have also tried to generate concise answers like entity names by specific instructions, but it leads to worse performance.

Table 9 Accuracy (%) of ChatGPT, KGQACL framework with GPT2 and KGQARR frameworks with the other PLMs on SimpleQuestions (SQ), WebQuestionSP (WB) and FreebaseQA (FBQ)

PLMs	Benckmarks		
	SQ	WQ	FBQ
BERT	76.7	68.7	41.7
ROBERTA	77.7	69.0	42.0
XLNET	75.7	68.0	41.0
GPT2	72.7	64.3	39.0
ALBERT	76.3	66.7	41.3
DISTILBERT	76.7	67.7	41.3
DISTILROBERTA	76.0	67.0	40.3
LUKE	78.0	68.3	42.0
KEPLER	77.3	68.7	41.7
CHATGPT	29.3	76.3	93.7

Bolded numbers indicate the highest accuracy

category of errors due to the difficulty of verifying that all enumeration items are correct. Note that when enumeration items include the gold answers, we consider the answer as correct. That is, In type 1 ENUMERATION, ChatGPT's answer does not contain a golden answer. Even though we regard all the questions of Type 1 as correct, the accuracy of ChatGPT on SimpleQuestions is 56.3%, which is still significantly inferior to other PLMs. WRONG ANSWERS (account for 26.9%) indicate that ChatGPT answers differ from gold answers. Besides, we noticed that ChatGPT even generates incorrect facts, also known as the hallucination problem [63]. For example, in the second example in Table 10, this politician and revolutionary is actually *Felix Dzerzhinsky* rather than *Ivan Dzerzhinsky*. In addition, 22.6% of the error cases are due to a lack of knowledge about the subject entity (i.e. LACK OF KNOWLEDGE), and 10.0% of the error cases are due to a lack of additional information to disambiguate the subject entity (i.e. AMBIGUOUS ENTITIES). 1.4% of the errors are due to ChatGPT misunderstanding the semantics of the question (i.e. MISUNDERSTANDING). The poor quality of the question itself causes 0.9% of the errors (i.e. DATASET PROBLEM). These cases demonstrate that ChatGPT may generate factual errors and still lacks extensive factual knowledge since many subject entities cannot be identified.

7 Conclusion and future works

Due to the improved performance of PLMs on most NLP tasks, it has become a consensus to use PLMs as a skeleton to solve NLP tasks. In this paper, we investigate the application of PLMs to solve a knowledge-intensive task, namely knowledge graph question answering. We conduct comprehensive experiments to explore the accuracy and efficiency performance of PLMs on KGQA, as well as the scalability of PLMs as KG size increases. In addition, we compare the performance between ChatGPT and other PLMs on three KGQA datasets. We present a detailed analysis of these experimental results and draw some important conclusions regarding the use of PLMs in KGQA.

1. ROBERTA and the knowledge-enhanced PLMs LUKE and KEPLER achieve the highest accuracy performance in the KGQA task. LUKE and KEPLER performed better on the small-scale KG benchmarks, and ROBERTA performed better on the large-scale KG benchmarks.

Table 10 Six error types of ChatGPT on SQ**1. ENUMERATION (38.2%)**

Question: *what album is from the dance music genre?*

ChatGPT: *There are many albums that belong to the dance music genre, as it is a broad and diverse genre. Some popular dance music albums include: * Daft Punk – Homework (1997) * The Chemical Brothers – Dig Your Own Hole (1997)...*

Gold: *love's crazy*

2. WRONG ANSWERS (26.9%)

Question: *what is ivan dzerzhinsky profession ?*

ChatGPT: *Ivan Dzerzhinsky was a Soviet **statesman** and Bolshevik **revolutionary**...*

Gold: *composer*

3. LACK OF KNOWLEDGE (22.6%)

Question: *which country does elizabeth larner have nationality in?*

ChatGPT: *I'm sorry, but I don't have any information on an individual named Elizabeth Larner and their nationality. Can you provide more context or details?*

Gold: *kingdom of great britain and northern ireland*

4. AMBIGUOUS ENTITIES (10.0%)

Question: *who is the lyricist of forever?*

ChatGPT: *Forever is a popular song and there are several songs with the title "Forever." Can you provide more context or specify which "Forever" you're referring to? This would help me better answer your question.*

Gold: *ferr*

5. MISUNDERSTANDING (1.4%)

Question: *who plays defender in soccer?*

ChatGPT: *In soccer, players who play in the defensive positions are referred to as defenders. The main responsibility of a defender is to prevent the opposing team from scoring goals by blocking shots, tackles, and interceptions...*

Gold: *bogdan bucurică*

6. DATASET PROBLEM (0.9%)

Question: *what is book was published by ?*

ChatGPT: *I'm sorry, but you haven't specified which book you are referring to. Can you please provide the name of the book so I can assist you better?*

Gold: *thirteenth tribe*

2. Lightweight PLMs DISTILBERT and DISTILROBERTA with knowledge distillation technology significantly improve efficiency and have lower accuracy than other PLMs on the small-scale KG benchmarks. However, DISTILBERT and DISTILROBERTA exhibit the best scalability. As KG size increases, the gap between their accuracy and that of other PLMs is gradually eliminated.
3. The accuracy of XLNET with permuted language modelling and GPT2 with language modelling is worse than that of PLMs with masked language modelling, especially the KGQARR framework based on GPT2.
4. The combined overall accuracy and efficiency results of KGQA show that PLMs-based entity disambiguation has no advantage over fuzzy matching-based entity disambiguation. Although the former is significantly better than the latter in the performance of entity disambiguation, the gap in accuracy between the two KGQA systems based on them is insignificant because the answer queryf module has the ability to disambiguate.

5. ChatGPT shows superior performance on zero-shot WebQuestions and FreebaseQA, even significantly outperforming other PLMs with fine-tuning. We speculate that this is due to the fact that ChatGPT has seen a similar corpus during training, as it performs extremely poorly on manually constructed SimpleQuestions. The error case analysis suggests that ChatGPT may generate answers with incorrect facts and still lack knowledge since many subject entities cannot be identified.

Further, we examine the overall results of the various PLMs on the subtasks of KGQA and obtain similar conclusions. ROBERTA and BERT exhibit the best performance on the entity detection and entity disambiguation tasks, while the knowledge-enhanced PLMs LUKE and KEPLER show strong capabilities on the relation detection task. DISTILBERT and DISTIL-ROBERTA have a clear efficiency advantage and perform well on all tasks except for the entity disambiguation task, which is slightly inferior. In addition, we find that the KGQARR-based systems are significantly less efficient than the KGQACL-based systems. Furthermore, the higher accuracy of KGQARR-based systems than KGQACL-based systems is only when the KG scale is small. As the KG scale increases, the former is gradually inferior to the latter, which indicates the poor scalability of the KGQARR framework.

In future work, we will extend the proposed simple KGQA framework to the multi-hop complex KGQA framework. We will also keep investigating the application of knowledge distillation and knowledge-enhanced PLMs in KGQA, as our experiments show them to be promising. In addition, we will follow up with artificial general intelligence models like ChatGPT and test them more carefully, especially in terms of efficiency.

Author Contributions Nan Hu: Conceptualization, Methodology, Software, Writing - Original Draft. Yike Wu: Investigation, Software, Validation. Guilin Qi: Conceptualization, Methodology, Writing - review & editing. Dehai Min: Software. Jiaoyan Chen: Writing - review & editing. Jeff Z. Pan: Writing - review & editing. Zafar Ali: Validation.

Funding This work is supported by National Nature Science Foundation of China (No. U21A20488).

Data Availability All datasets and codes in this paper can be accessed from <https://github.com/aanonymouss/PLMs-in-Practical-KBQA>.

Declarations

Ethical Approval Not applicable

Competing Interests The authors declare that there are no competing interest regarding the publication of this article.

References

1. Manning, C.D.: Human language understanding & reasoning. In: *Daedalus*, pp. 127–138 (2022). https://doi.org/10.1162/daed_a_01905
2. Mohammed, S., Shi, P., Lin, J.J.: 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*, pp. 291–296 (2018). <https://doi.org/10.18653/v1/n18-2047>
3. Lukovnikov, D., Fischer, A., Lehmann, J.: Pretrained transformers for simple question answering over knowledge graphs. In: *18th International Semantic Web Conference*, pp. 470–486 (2019). https://doi.org/10.1007/978-3-030-30793-6_27

4. Golub, D., He, X.: Character-level question answering with attention. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1598–1607 (2016). <https://doi.org/10.18653/v1/d16-1166>
5. Petrochuk, M., Zettlemoyer, L.: Simple questions nearly solved: a new upperbound and baseline approach. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 554–558 (2018). <https://doi.org/10.18653/v1/d18-1051>
6. Türe, F., Jojic, O.: No need to pay Attention: simple recurrent neural networks work! In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2866–2872 (2017). <https://doi.org/10.18653/v1/d17-1307>
7. Yu, M., Yin, W., Hasan, K.S., Santos, C.N., Xiang, B., Zhou, B.: Improved neural relation detection for knowledge base question answering. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 571–581 (2017). <https://doi.org/10.18653/v1/P17-1053>
8. Cui, H., Peng, T., Feng, L., Bao, T., Liu, L.: Simple question answering over knowledge graph enhanced by question pattern classification. In: Knowl. Inf. Syst., pp. 2741–2761 (2021). <https://doi.org/10.1007/s10115-021-01609-w>
9. Lukovnikov, D., Fischer, A., Lehmann, J., Auer, S.: Neural network-based question answering over knowledge graphs on word and character level. In: Proceedings of the 26th International Conference on World Wide Web, pp. 1211–1220 (2017). <https://doi.org/10.1145/3038912.3052675>
10. Hao, Y., Liu, H., He, S., Liu, K., Zhao, J.: Pattern-revising enhanced simple question answering over knowledge bases. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3272–3282 (2018). <https://aclanthology.org/C18-1277/>
11. Yin, W., Yu, M., Xiang, B., Zhou, B., Schütze, H.: Simple question answering by attentive convolutional neural network. In: Proceedings of the 26th International Conference on Computational Linguistics, pp. 1746–1756 (2016). <https://aclanthology.org/C16-1164/>
12. Zhao W., Chung T., Goyal AK., Metallinou A.: 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, pp. 324–334 (2019). <https://doi.org/10.18653/v1/n19-1029>
13. Hao, Y., Zhang, Y., Liu, K., He, S., Liu, Z., Wu, H., Zhao, J.: 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, pp. 221–231 (2017). <https://doi.org/10.18653/v1/P17-1021>
14. Luo, D., Su, J., Yu, S.: A BERT-based approach with relation-aware attention for knowledge base question answering. In: 2020 International Joint Conference on Neural Networks, pp. 1–8 (2020). <https://doi.org/10.1109/IJCNN48605.2020.9207186>
15. Dai, Z., Li, L., Xu, W.: CFO: conditional focused neural question answering with large-scale knowledge bases. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 800–810 (2016). <https://doi.org/10.18653/v1/p16-1076>
16. Lan, Y., Wang, S., Jiang, J.: knowledge base question answering with a matching-aggregation model and question-specific contextual relations. In: IEEE ACM Trans. Audio Speech Lang. Process., pp. 1629–1638 (2019). <https://doi.org/10.1109/TASLP.2019.2926125>
17. Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
18. Hochreiter, S., Schmidhuber, J.: Long short-term memory. In: Neural Computation, pp. 1735–1780 (1997)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Communications of the ACM, pp. 84–90 (2017). <https://doi.org/10.1145/3065386>
20. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 4171–4186 (2019). <https://doi.org/10.18653/v1/n19-1423>
21. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. In: [ArXiv:1907.11692](https://arxiv.org/abs/1907.11692) (2019)
22. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A lite BERT for self-supervised learning of language representations. In: 8th International Conference on Learning Representations (2020)
23. Huang, X., Zhang, J., Li, D., Li, P.: Knowledge graph embedding based question answering. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 105–113 (2019). <https://doi.org/10.1145/3289600.3290956>
24. Bordes, A., Usunier, N., Chopra, S., Weston, J.: Large-scale simple question answering with memory networks. In: [arXiv:1506.02075](https://arxiv.org/abs/1506.02075) (2015)

25. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*, pp. 3111–3119 (2013)
26. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543 (2014). <https://doi.org/10.3115/v1/d14-1162>
27. Reimers N., Gurevych I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 3980–3990 (2019). <https://doi.org/10.18653/v1/D19-1410>
28. Li, B.Z., Min, S., Iyer, S., Mehdad, Y., Yih, W.: Efficient one-pass end-to-end entity linking for questions. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6433–6441 (2020). <https://doi.org/10.18653/v1/2020.emnlp-main.522>
29. Wu, L.Y., Petroni, F., Josifoski, M., Riedel, S., Zettlemoyer, L.: Scalable zero-shot entity linking with dense entity retrieval. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6397–6407 (2020). <https://doi.org/10.18653/v1/2020.emnlp-main.519>
30. Chen, S., Wang, J., Jiang, F., Lin, C.: Improving entity linking by modeling latent entity type information. In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pp. 7529–7537 (2020)
31. Oliya, A., Saffari, A., Sen, P., Ayoola, T.: End-to-end entity resolution and question answering using differentiable knowledge graphs. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4193–4200 (2021). <https://doi.org/10.18653/v1/2021.emnlp-main.345>
32. Wang, Z., Ng, P.K., Nallapati, R., Xiang, B.: Retrieval, re-ranking and multi-task learning for knowledge-base question answering. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 347–357 (2021). <https://doi.org/10.18653/v1/2021.eacl-main.26>
33. Yamada, I., Asai, A., Shindo, H., Takeda, H., Matsumoto, Y.: LUKE: deep contextualized entity representations with entity-aware self-attention. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6442–6454 (2020). <https://doi.org/10.18653/v1/2020.emnlp-main.523>
34. Zhang, T., Wang, C., Hu, N., Qiu, M., Tang, C., He, X., Huang, J.: DKPLM: decomposable knowledge-enhanced pre-trained language model for natural language understanding. In: *Thirty-Sixth AAAI Conference on Artificial Intelligence*, pp. 11703–11711 (2022)
35. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: ERNIE: enhanced language representation with informative entities. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pp. 1441–1451 (2019). <https://doi.org/10.18653/v1/p19-1139>
36. Wang, X., Gao, T., Zhu, Z., Liu, Z., Li, J., Tang, J.: KEPLER: a unified model for knowledge embedding and pre-trained language representation. In: *Transactions of the Association for Computational Linguistics*, **9**, pp. 176–194. (2021). https://doi.org/10.1162/tacl_a_00360
37. Peters, M.E., Neumann, M., Robert L. Logan, I., Schwartz, R., Joshi, V., Singh, S., Smith, N.A.: Knowledge enhanced contextual word representations. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 43–54 (2019). <https://doi.org/10.18653/v1/D19-1005>
38. Bollacker, K.D., Evans, C., Paritosh, P.K., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 1247–1250 (2008). <https://doi.org/10.1145/1376616.1376746>
39. Danny Sullivan.: A reintroduction to our knowledge graph and knowledge panels. <https://blog.google/products/search/about-knowledge-graph-and-knowledge-panels/>(2020). Accessed 3 Oct 2022
40. Lan, Y., Jiang, J.: Query graph generation for answering multi-hop complex questions from knowledge bases. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 969–974 (2020). <https://doi.org/10.18653/v1/2020.acl-main.91>
41. Gu, Y., Kase, S.E., Vanni, M.T., Sadler, B.M., Liang, P., Yan, X., Su, Y.: Beyond I.I.D.: Three levels of generalization for question answering on knowledge bases. In: *Proceedings of the Web Conference*, pp. 3477–3488 (2021). <https://doi.org/10.1145/3442381.3449992>
42. Ye, X., Yavuz, S., Hashimoto, K., Zhou, Y., Xiong, C.: RNG-KBQA: generation augmented iterative ranking for knowledge base question answering. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 6032–6043 (2022). <https://doi.org/10.18653/v1/2022.acl-long.417>
43. Gu, Y., Su, Y.: ArcaneQA: dynamic program induction and contextualized encoding for knowledge base question answering. In: *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 1718–1731 (2022)

44. Chen, S., Liu, Q., Yu, Z., Lin, C., Lou, J., Jiang, F.: ReTraCk: a flexible and efficient framework for knowledge base question answering. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, pp. 325–336 (2021). <https://doi.org/10.18653/v1/2021.acl-demo.39>
45. Qin, K., Li, C., Pavlu, V., Aslam, J.A.: Improving query graph generation for complex question answering over knowledge base. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 4201–4207 (2021). <https://doi.org/10.18653/v1/2021.emnlp-main.346>
46. Xie, T., Wu, C., Shi, P., Zhong, R., Scholak, T., Yasunaga, M., Wu, C., Zhong, M., Yin, P., Wang, S.I., Zhong, V., Wang, B., Li, C., Boyle, C., Ni, A., Yao, Z., Radev, D., Xiong, C., Kong, L., Zhang, R., Smith, N.A., Zettlemoyer, L., Yu, T.: UnifiedSKG unifying and multi-tasking structured knowledge grounding with text-to-text language models. In: [arXiv:2201.05966](https://arxiv.org/abs/2201.05966) (2022)
47. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: TinyBERT Distilling BERT for natural language understanding. In: Findings of the Association for Computational Linguistics, pp. 4163–4174 (2020). <https://doi.org/10.18653/v1/2020.findings-emnlp.372>
48. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT smaller, faster, cheaper and lighter. In: [arXiv:1910.01108](https://arxiv.org/abs/1910.01108) (2019)
49. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (2018). Accessed 4 Oct 2022
50. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: XLNet generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems, pp. 5754–5764 (2019)
51. Zhang, C., Lai, Y., Feng, Y., Zhao, D.: A review of deep learning in question answering over knowledge bases. In: AI Open, pp. 205–215 (2021)
52. Lan, Y., He, G., Jiang, J., Jiang, J., Zhao, W.X., Wen, J.: Complex knowledge base question answering a survey. In: IEEE TKDE (2021)
53. Gu, Y., Pahuja, V., Cheng, G., Su, Y.: Knowledge base question answering: a semantic parsing perspective. [ArXiv, In: arXiv:2209.04994\(2022\)](https://arxiv.org/abs/2209.04994)
54. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: Pre-trained models for natural language processing a survey. In: Science China Technological Sciences, pp. 1872–1897 (2020)
55. Han, N., Topic, G., Noji, H., Takamura, H., Miyao, Y.: An empirical analysis of existing systems and datasets toward general simple question answering. In: COLING, pp. 5321–5334 (2020)
56. Jiang, K., Wu, D., Jiang, H.: FreebaseQA: a new factoid QA data set matching trivia-style question-answer pairs with freebase. In: North American Chapter of the Association for Computational Linguistics, pp. 318–323 (2019)
57. Yih, W., Richardson, M., Meek, C., Chang, M., Suh, J.: The value of semantic parse labeling for knowledge base question answering. In: Annual Meeting of the Association for Computational Linguistics (2016)
58. Hu, N., Bi, S., Qi, G., Wang, M., Hua, Y., Shen, S.: Improving core path reasoning for the weakly supervised knowledge base question answering. In: DASFAA, pp. 162–170 (2022)
59. Zhang, J., Zhang, X., Yu, J., Tang, J., Tang, J., Li, C., Chen, H.: Subgraph retrieval enhanced model for multi-hop knowledge base question answering. In: Annual Meeting of the Association for Computational Linguistics, pp. 5773–5784 (2022)
60. Das, R., Zaheer, M., Thai, D.N., Godbole, A., Perez, E., Lee, J., Tan, L., Polymenakos, L., McCallum, A.: Case-based reasoning for natural language queries over knowledge bases. In: Conference on Empirical Methods in Natural Language Processing, pp. 9594–9611 (2021)
61. Ye, X., Yavuz, S., Hashimoto, K., Zhou, Y., Xiong, C.: RNG-KBQA: generation augmented iterative ranking for knowledge base question answering. In: Annual Meeting of the Association for Computational Linguistics, pp. 6032–6043 (2021)
62. Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., Yang, D.: Is ChatGPT a general-purpose natural language processing task solver? In: [arXiv:2302.06476](https://arxiv.org/abs/2302.06476) (2023)
63. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q.V., Xu, Y., Fung, J.: A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In: [arXiv:2302.04023](https://arxiv.org/abs/2302.04023) (2023)
64. Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., Wu, Y.: How close is ChatGPT to human experts? comparison corpus, evaluation, and detection. In: [arXiv:2301.07597](https://arxiv.org/abs/2301.07597) (2023)
65. Christiano, P.F., Leike, J., Brown, T.B., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: Neural Information Processing Systems, pp. 4299–4307 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Nan Hu¹ · Yike Wu¹ · Guilin Qi¹ · Dehai Min¹ · Jiaoyan Chen² · Jeff Z Pan³ · Zafar Ali¹

Nan Hu
nanhu@seu.edu.cn

Yike Wu
wuyike@seu.edu.cn

Dehai Min
zhishanq@seu.edu.cn

Jiaoyan Chen
jiaoyan.chen@manchester.ac.uk

Jeff Z Pan
j.z.pan@ed.ac.uk

Zafar Ali
zafarali@seu.edu.cn

¹ School of Computer Science and Engineering, Southeast University, 2 Dongda Rd, 211189 Nanjing, Jiangsu, China

² Department of Computer Science, The University of Manchester, Oxford Rd, M13 9PL Manchester, UK

³ School of Informatics, The University of Edinburgh, 10 Crichton St, 2EH8 9AB Edinburgh, UK