Relation-Guided Pre-Training for Open-Domain Question Answering

Ziniu Hu, Yizhou Sun, Kai-Wei Chang
University of California, Los Angeles
{bull, yzsun, kwchang}@cs.ucla.edu

Abstract

Answering complex open-domain questions requires understanding the latent relations between involving entities. However, we found that the existing QA datasets are extremely imbalanced in some types of relations, which hurts the generalization performance over questions with long-tail relations. To remedy this problem, in this paper, we propose a Relation-Guided Pre-Training (RGPT-QA) framework. We first generate a relational QA dataset covering a wide range of relations from both the Wikidata triplets and Wikipedia hyperlinks. We then pre-train a QA model to infer the latent relations from the question, and then conduct extractive QA to get the target answer entity. We demonstrate that by pre-training with proposed RGPT-QA technique, the popular open-domain QA model, Dense Passage Retriever (DPR), achieves 2.2%, 2.4%, and 6.3% absolute improvement in Exact Match accuracy on Natural Questions, TriviaQA, and WebQuestions. Particularly, we show that RGPT-QA improves significantly on questions with long-tail relations.

1 Introduction

Open domain question answering is a challenging task that answers factoid questions based on evidence in a large corpus (e.g., Wikipedia). Most open-domain QA systems follow retriever-reader pipeline (Chen et al., 2017), in which a retriever selects a subset of candidate entities and associated passages from the corpus that might contain the answer, then a reader extracts a text span from the passages as the answer. This process involves multiple entities that are relevant to answer the question. The QA system is required to extract these entities from the question and passages and identify the (latent) semantic relations between these entities in order to answer the question. For example, to answer the following question: “Where did Steph Curry play college basketball at?”, the QA model is required to reason the implicit relation triplet (Steph Curry, Educated At, Davidson College) to identify the correct answer.

To capture the relation knowledge required to answer questions, most QA systems rely on human-annotated supervised QA datasets. However, it is expensive and tedious to annotate a large set of QA pairs that cover enough relational facts for training a strong QA model. In addition, we showed that even for a large QA dataset like Natural Questions (Kwiatkowski et al., 2019), its training set only covers 16.4% of relations in WikiData (Vrandecic and Krötzsch, 2014) knowledge graph. Moreover, for those covered relations, the frequency distribution is imbalanced, i.e., 30% of relation types appear only once. Consequently, for the questions involving infrequent (a.k.a, long-tail) relations in the training set, the QA exact match accuracy is 22.4% lower than average. Such a biased relation distribution of existing QA datasets severely hurts the generalization of trained QA systems.

To improve the open-domain QA systems for questions with long-tail relations, in this paper, we propose RGPT-QA, a simple yet effective Relation-Guided Pre-training framework for training QA models with augmented relational facts from knowledge graph. The framework consists of two steps: 1) generate a relational QA dataset that covers a wide range of relations without human labeling; 2) pre-train a QA model to predict latent relations from questions and conduct extractive QA.

The key of our framework is to generate a relational QA dataset that align entities in Wikipedia passages with structured knowledge graph (e.g., WikiData). We call such a dataset Grounded Relational Wiki-Graph. In this graph, each edge indicates the relationship of two connected entities, and the edge is linked to a passage in Wikipedia describing this relationship. As WikiData knowledge
Figure 1: Cumulative distribution function (CDF) of relation frequency in Natural Question Training set.

Figure 2: Exact Match accuracy of a trained DPR model in validation set with different relation frequency in training set.

graph also suffers from low coverage of long-tail entities and relations, we further convert hyperlinks in Wikipedia into knowledge triplets without specifying relation labels. Next, we link each relation triplet to a Wikipedia passage to help generate natural questions. We assume that if one passage in the Wiki-page of source entity contains the target entity, then the context in this passage describes the relationship between the two entities. With the constructed graph, we use a template to synthesize question and answer pairs and then pre-train the QA model to capture the relational facts for answering complex open-domain questions.

As a pre-training method, RGPT-QA can be incorporated with any open-domain QA system. In this paper, we utilize the recently developed Dense Passage Retriever (DPR) (Karpukhin et al., 2020) as the base QA system to evaluate the proposed pre-training effectiveness. Experimental results show that RGPT-QA enhances DPR’s Exact Match accuracy by 2.2%, 2.4%, and 6.3% on Natural Questions, TriviaQA and WebQuestions respectively. Compared with the existing QA pre-training methods (Lee et al., 2019; Guu et al., 2020a; Lewis et al., 2019), RGPT-QA explicitly captures a wide range of relational facts and thus achieves better performance. Moreover, for the questions containing long-tail relations in Natural Questions, the performance is improved by 10.9%, showing that RGPT-QA alleviates the unbalanced relation distribution problem in the existing QA datasets.

The key contributions of this paper are:

- We propose RGPT-QA, a pre-training method to inject knowledge from relational facts in knowledge graph into QA models.
- RGPT-QA enhances the performance of a popular QA model, i.e., DPR, especially on the questions with long-tail relations.

2 Preliminary and Empirical Analysis

In this section, we firstly introduce the retriever-reader pipeline for open-domain QA, and then we analyze how the relation distribution in existing QA datasets influence generalization performance.

Open-Domain Question Answering. We focus on open-domain question answering that requires to extract answer from a large corpus (e.g. Wikipedia) $\mathcal{C} = \{p_i\}_{i=1}^N$ containing $N$ passages. Most open-domain QA systems follow a retriever-reader pipeline proposed by Chen et al. (2017). Given a factoid question $q$, the QA system first retrieves $K$ relevant passages $\{p_j\}_{j=1}^K$ from the corpus $\mathcal{C}$. Then a reading comprehension module extracts a text span $w_{\text{start}}, \ldots, w_{\text{end}}$ from one of these retrieved passages as the answer $a$ to the question. Some QA dataset annotated the passage where the answer $a$ is derived. We called this passage ground truth passage.

For the retriever, earlier systems utilize term-based retrieval methods, such as TF-IDF and BM25, which fails to capture the semantic relationship between question and passage beyond lexical matching. Recent studies (Lee et al., 2019; Karpukhin et al., 2020; Dhingra et al., 2020) use BERT-like pretrained language model to encode the question and passages independently into dense representations, and use maximum inner product search (MIPS) algorithms (Shrivastava and Li, 2014) to efficiently retrieve the most similar passage for each question. In this paper, we utilize Dense Passage Retriever (DPR) (Karpukhin et al., 2020) as the base QA model.

Relation Bias of Existing QA Datasets. We first explore how much relational knowledge between entities is required to answer the questions in the existing open-domain QA dataset. We con-
duct an empirical study to analyze the relation distribution in Natural Questions, one of the largest open-domain QA datasets, and how it influences QA model’s performance.

For each question in Natural Question training set, we first select the entity that the ground-truth passage is associated with. We then combine the entity with the answer as an entity pair, and check whether we can find a relation triplet in WikiData describing the relation between these two entities. Out of 58,880 training QA pairs, there are 23,499 pairs that could be aligned. The aligned QA pairs cover 329 relations, which accounts for 16.4% of the total 2,008 relations in WikiData. For most unaligned QA pairs, the answers are not entities and thus cannot be aligned to the graph.

In addition to the low relation coverage issue in Natural Question, we also find that the relation distribution is imbalanced. As showed in Figure 1, 90% of relations have frequency less than 41, and 30% of relations appear only once. On the contrary, the most frequent relation “P161 (cast member)” appears 1,915 times out of 9,238 aligned QA pairs. A complete list of all these relations with aligned QA pairs is shown in Table 6-9 in Appendix.

We then study whether the imbalanced relation distribution influences the performance of QA models trained on these datasets. We use a DPR model trained on training set of Natural Questions and then calculate the Exact Match accuracy in validation set of each aligned QA pairs. We then analyze the correlation of the accuracy with the relation frequency in training set. As illustrated in Figure 2, the validation set accuracy is overall proportional to the relation frequency in training set. For those relations with frequency less than 5, the average accuracy is only 20.3%, much lower than the average accuracy 42.7% over all samples in validation set. This shows that the relation bias in existing QA datasets severely influences the generalization of QA models to questions with long-tail relations.

### 3 Method

In this section, we will discuss RGPT-QA framework in: 1) how to generate relational QA dataset for the pre-training purpose; and 2) how to construct a self-training task to empower QA model to capture relational facts.

| # of linked Entity | 5,640,366 |
|--------------------|-----------|
| # of relation labels | 2,008 |
| # of labelled triplet | 14,463,728 |
| # of unlabeled triplet (hyperlink) | 66,796,110 |
| # of grounded descriptions per triplet | 1.25 |

**Table 1: Statistics of Grounded Relational Wiki-Graph.**

**3.1 Construct QA Pre-Training Dataset**

To help QA model capture the knowledge from relation facts required to answer open-domain questions, we first focus on generating relational QA dataset, in which there exist relation connections between the source entity in questions to the target answer. Specifically, each QA pair datapoint \( d = \langle (s, r, t), q, p^+ \rangle \) consists of three components: 1) relational triplet \( (s, r, t) \), in which \( r \) denotes the relation between source entity \( s \) and target entity \( t \); 2) question \( q \) in natural language asking which entity has relation \( r \) to source entity \( s \), with target entity \( t \) as the correct answer; 3) positive context passage \( p^+ \in \mathbb{C}[s] \), a passage from source entity’s Wiki-page that contains the target answer \( t \).

**Grounded Relational Wiki-Graph.** To generate QA pre-training dataset, leveraging the relation triplets in knowledge graph, e.g., WikiData, is a natural choice to define questions that require relation reasoning. We therefore construct Grounded Relational Wiki-Graph, in which each relation triplet \( (s, r, t) \) is linked to a set of description passages \( \{\text{desc.}(s, t)\} \) in the Wiki-page of entity \( s \). These descriptions would be later utilized to generate questions \( q \) and positive context passages \( p^+ \).

To construct such a graph, we use the 2021 Jan. English dump of Wikidata and Wikipedia. For each Wikipedia hyperlink \( \langle s, ?, t \rangle \) (\( ? \) denotes the relation is unlabeled), the passage containing anchored text to \( t \) in the Wiki-page of \( s \) naturally fits our requirement for \( \text{desc.}(s, t) \). For each WikiData relation triplet \( \langle s, r, t \rangle \), if the two entities are linked by a hyperlink in Wikipedia, we label the relation of the aligned hyperlink as \( r \). For the other triplets \( \langle s, r, t \rangle \) without alignment with hyperlinks, we extract all mentioning of target entity \( t \) from the Wiki-page of \( s \), and use the context passage as \( \text{desc.}(s, t) \). The dataset statistics are shown in Table 1.

**Relational QA Pair Generation** In the following, we introduce the details to generate the relational QA pair from the constructed graph. Recent unsupervised QA studies (Li et al., 2020; Pan et al., 2020) revealed that if the question \( q \) and
context passage $p^+$ share a large lexical overlap, then the QA model could utilize low-level lexical patterns as shortcuts to find the answer. These shortcuts hinder the model from learning to comprehend the passages and answer the questions, hurting model’s generalizability. To avoid this lexical overlap issue, we aim to generate questions from a passage that is different from the context passage $p^+$.

We first select all the entity pairs $\langle s, t \rangle$ that have mutual links in the Grounded Relational Wiki-Graph, with $\text{desc.}(s, t)$ and $\text{desc}(t, s)$ in part of Wikipage of $s$ and $t$ respectively, describing the relationship between the two entities. Without loss of generality, we denote $s$ as source entity and $t$ as the target answer. The passage $\text{desc.}(s, t)$ containing target answer $t$ can be used as the positive passage $p^+$.

Next, we generate a question that is lexically different from $p^+$ using the following template:

$$q(s, r, t) = [\text{MASK}(r)] \text{ of } [s] \text{ which } [\text{desc.}(t, s)]?$$

in which $\text{MASK}(r)$ is a relation mask token. As $\text{desc.}(t, s)$ contains source entity $s$, it provides information to describe the relationship between $s$ and $t$, based on which the QA model should learn to infer the latent relation $r$, and retrieve positive passage $p^+ = \text{desc.}(s, t)$ and extract answer entity $t$. In addition, as $\text{desc.}(t, s)$ and $\text{desc.}(s, t)$ come from different Wiki-page, our question generation procedure can avoid the lexical overlap issue that often occur in prior Unsupervised QA methods.

**Mask Target Answer.** As description $\text{desc.}(t, s)$ is from target answer $t$’s wiki-page, it often contains the name of entity $t$. We thus need to mask $t$ from the question. Otherwise, the pre-trained model can simply identify the answer to a question based on the local patterns.

As an example, in Figure 3, we show how to generate question for triplet $\langle \text{Stephen Curry}, ?, \text{Splash Brothers} \rangle$. We firstly retrieve two descriptive passages $\text{desc.}(s, t)$ and $\text{desc.}(t, s)$ in two entities’ wiki pages. Using the template, we generate the question along with the ground-truth passage. We then mask out the target entity in question and source entity in true passage (will discuss later in retrieval pre-training) to avoid shortcut. A list of generated relational QA pairs are shown in Table 10 in Appendix.

![Diagram](image.png)

**Figure 3:** Example of a generated relational QA pair from Grounded Relational Wiki-Graph.

### 3.2 Relation-Guided QA Pre-Training

With the generated relational QA dataset, we introduce how to pre-train both retriever and reader components in the QA model.

#### 3.2.1 Relation Prediction Pre-Training

Our generated QA dataset contains the relation label $r$ between the source entity $s$ and the answer target $t$. Therefore, we design a self-training task to guide the model to predict the latent relation...
in question, which can benefit both retriever and reader. Specifically, we adopt a linear projection layer $L_R(\cdot)$ over the $\text{BERT}_{[\text{CLS}]}$ token embedding to predict the relation over the WikiData relation set. The pre-training loss of relation prediction is:

$$\mathcal{L}_{\text{rel}} = \frac{1}{B} \sum_q - \log P(r \mid q; \theta),$$

**Self-Distillation for Unlabelled Relation** The hyperlinks in wikipedia also provide valuable implicit information about the relations between entities. To leverage them, we use the trained relation predictor at each epoch with fixed parameter $\theta$ as teacher model to assign soft label and then progressively train the relation predictor as student model based on the assigned labels in the next epoch. This approach is referred to as self-distillation in the literature (Xie et al., 2020; Chen et al., 2020). We minimize this self-distillation loss as:

$$\mathcal{L}_{\text{distill}} = \frac{1}{B} \sum_q \sum_{\hat{r}} - \log P(\hat{r} \mid q; \theta) \cdot \text{sg}(P(\hat{r} \mid q; \hat{\theta})), $$

where $\text{sg}(\cdot)$ denotes the operation of stop gradient, which avoids back propagation to the teacher network with fixed parameter $\hat{\theta}$, $\hat{r}$ is enumerating all the relation labels.

As the relation predictor at early stages cannot give a reasonable prediction, we put a dynamic weight schedule to $\mathcal{L}_{\text{distill}}$ by a time-dependent weighting term $1 - e^{-\text{epoch}}$, which ramps up from zero to one. Combining the weighted self-distillation loss $\mathcal{L}_{\text{distill}}$ with the supervised relation loss $\mathcal{L}_{\text{rel}}$, we get the final relation loss $\mathcal{L}_{\text{rel}}$ to train the model capturing all relational facts covered in the Grounded Relational Wiki-Graph.

3.2.2 Dense Retrieval Pre-Training

The goal of dense retrieval pre-training is to get a question encoder $\text{Enc}_P$ and a passage encoder $\text{Enc}_p$ to map questions and all passages in the Wiki Corpus $\mathbb{C}$ into an embedding space, such that each question $q$ is close to its ground-truth positive context passage $p^+$ in the embedding space. The objective is as follows:

$$P_{\text{etr}}(p^+ \mid q, \mathbb{C}) = \frac{\exp \left( \text{sim}(q, p^+) \right)}{\sum_{p \in \mathbb{C}} \exp \left( \text{sim}(q, p) \right)}, \quad (1)$$

where $\text{sim}(q, p)$ is the cosine similarity between the normalized embeddings of question and passage.

**Two-Level Negative Passage Sampling.** As we cannot enumerate all other passages in the denominator of Eq(1), we need to sample a set of negative passages for contrastive learning. Previous studies (Karpukhin et al., 2020) have revealed that it is essential that the sampled negative passages should be hard enough to train the retriever. As the question and passage embeddings are encoded independently, DPR can efficiently calculate the similarity of each question to all passages in the batch via dot product. Based on this property, as long as the passages within a batch are similar to each other, they serve the hard cases of negative passages to others. We thus propose a two-level negative passage sampling strategy to construct hard cases for training the retriever in the following.

We first sample at the level of entity. Given a set of randomly sampled $b$ entities, we adopt random walk from these seed entities over the Grounded Relational Wiki-Graph to get $B$ entities. As the connected entities have a relationship, their true passages are also semantically similar, and thus serve as good negative samples. We then conduct sampling at the level of passage. For each source entity $s_i$ with positive passage $p^+_i \in \mathbb{C}[s_i]$, we randomly pick $K$ other passages from the same Wiki-page to form a negative passage set $\{p^-_{i,j} \in \mathbb{C}[s_i], \text{s.t. } p^-_{i,j} \neq p^+_i \}_{j=1}^K$. These negative passages are similar to $p^+_i$, as they all describe the same entity $s_i$.

After we collect both the positive and $K$ negative passages for all the entities, we use the passage encoder $\text{Enc}_p$ to get a passage embedding matrix $P$ with dimension $(1 + K) \cdot B \times d$. We also use question encoder $\text{Enc}_Q$ to get question embedding matrix $Q$ with dimension $(B \times d)$. We then get a similarity matrix $S = QP^T$ with dimension $(B \times (1 + K) \cdot B)$, in which the diagonal entry corresponds to the similarity between question and its positive passage. We thus calculate the retrieval loss with in-batch negative samples via:

$$\mathcal{L}_{\text{etr}} = \frac{1}{B} \left( \sum_{i \in [1,B]} - \log \text{softmax}(S) \right)_{i,i}. \quad (2)$$

**Masking Source Entity.** As the true passage $p^+_i = \text{desc}.(s, t)$ might contain the name of source entity $s$. We mask out all the tokens of $s$ from the extracted passages, so that the model is required to understand the passages for correct retrieval instead of exploiting a shortcut.
3.2.3 Reading Comprehension Pre-Training

The goal of reading comprehension pre-training is to get a neural reader that re-ranks the top-k retrieved passages and extracts an answer span from each passage as the answer. The probability of a passage contains the target answer \(t\), and each token in the selected passage being the starting/ending positions of an \(t\) are defined as:

\[
P_{\text{rank}}(t \in p) = \frac{\exp(L_{\text{rank}}(\text{BERT}_{\text{CLS}}(q,p)))}{\sum_p \exp(L_{\text{rank}}(\text{BERT}_{\text{CLS}}(q,p)))},
\]

\[
P_{\text{start}}(i \mid p, q) = \frac{\exp(L_{\text{start}}(\text{BERT}_{\{i\}}(q,p)))}{\sum_j \exp(L_{\text{start}}(\text{BERT}_{\{j\}}(q,p)))},
\]

\[
P_{\text{end}}(i \mid p, q) = \frac{\exp(L_{\text{end}}(\text{BERT}_{\{i\}}(q,p)))}{\sum_j \exp(L_{\text{end}}(\text{BERT}_{\{j\}}(q,p)))},
\]

where \(L_{\text{start}}\) are linear project layers with different parameters. Note that the re-ranking module adopts cross-attention over questions and passages rather than the dot product of two independently encoded embedding used in retriever. For each QA pair \(d = \langle s, r, t, q, p^+ \rangle\), we select \(m\) other passages in wiki-page of entity \(s\) as negative passages, and maximize \(P_{\text{rank}}(t \in p^+)\). Then, we calculate \(P_{\text{start}}(i \mid p^+, q)\) and \(P_{\text{end}}(i \mid p^+, q)\) and maximize the probability for the ground-truth span of target answer \(t\). Combing the passage re-ranking and span extraction objectives, we get reading-comprehension loss \(L_{\text{read}}\).

4 Experiments

In this section, we evaluate RGPT-QA on three open-domain QA datasets: Natural Questions (NQ), Trivia QA and Web Questions (WQ).

4.1 Experiment Settings

We follow the pre-processing procedure described in DPR (Karpukhin et al., 2020) for a fair comparison. We use the English Wikipedia from Dec. 20, 2018 and split each article into passages of 100 disjoint words as the corpus. For each question in all the three datasets, we use a passage from the processed Wikipedia which contains the answer as positive passages. We evaluate the QA system by Exact Match (EM) Accuracy on the correct answer.

Our RGPT-QA could be integrated with any open-domain QA system. In this paper, we incorporate it with the recently developed QA system, Dense Passage Retriever (DPR) (Karpukhin et al., 2020) to evaluate our pre-training framework. The DPR model uses the RoBERTa-base (\(d=768, l=12\)) model as the base encoder. We first pre-train the retriever and reader in DPR using RGPT-QA. For retriever, we use the negative passage sampling strategy (c.f. Sec. 3.2.2), with initial entity size set to be 12, batch size of 128 and the hard negative passage number of 2. For reader, we randomly sample 64 source entities per batch to calculate the loss. For each entity, we sample 2 hard negative passages for re-ranking. We pre-train both the retriever and reader for 20 epochs using AdamW optimizer and a learning rate warm-up followed by linear decay. Pre-training is run on 8 Tesla V100 GPUs for two days. After the pre-training, we fine-tune the retriever and reader on each QA dataset following the same procedure and hyper-parameters described in DPR (Karpukhin et al., 2020).

QA Pre-Training Baselines. We compare RGPT-QA with three recently proposed pre-training methods for open-domain QA.

**T5** (Raffel et al., 2020) adopts multiple generative tasks to pre-train a generative model. The fine-tuned QA models directly generate answers without needing an additional retrieval step.

**ORQA** (Lee et al., 2019) adopts a Inverse Cloze Task (ICT) to pre-train retriever, which forces each sentence’s embedding close to context sentences.

**REALM** (Guu et al., 2020a) incorporates a retriever as a module into language model and trains the whole model over masked entity spans.

We directly report the results listed in their papers as they follow the same experiment settings.

We also add two knowledge-guided language models as baselines. Though not targeted at QA problem, these two methods are both designed to capture structured knowledge.

**KnowBERT** (Peters et al., 2019) adds entity embedding to each entity mention in text, and adopts the entity linking objective to pre-train the model.

**KEPLER** (Wang et al., 2019) uses Knowledge Embedding objective, i.e., TransE, to guide embedding encoded over entity description.

We initialize DPR base encoders by the released pre-trained models of these two work, and then fine-tune on each QA dataset with the same procedure.

We also add a Unsupervised Question Answering (Unsup.QA) (Lewis et al., 2019) as a baseline. For each entity as the answer, Unsup.QA selects a passage containing the entity as context passage and a cloze question. The cloze question is later rewritten by a machine translator to natural language. We use the generated QA dataset to pre-train both
Table 2: **End-to-end QA** Exact Match Accuracy (%) on test sets of three Open-Domain QA datasets, with the number of train/test examples shown in parentheses below. All the results except the last four rows are copied from the original papers. “–” denotes no results are available. Models in the first block are initialized by BERT/RoBERTa and then directly fine-tuned on the supervised QA datasets. While models in the second block are initialized by RoBERTa and then tuned on some QA pre-training tasks first, and then fine-tuned on the supervised QA datasets.

Table 3: **Retrieval (left)** accuracy over Top-20 results and **Reader (right)** Exact Match over Golden-Passages on validation sets of three Open-Domain QA datasets.

Table 4: **Ablation** of RGPT-QA components on validation sets of three Open-Domain QA datasets.

Table 5: **Ablation** of batch size and negative sampling for retrieval pre-training. B: Batch Size; K: Number of other passages as negative sample.

Comparing with other pre-training tasks for QA, RGPT-QA outperforms ORQA by 10.4%, 14.2% and 4.5% on the three datasets, and outperforms REALM by 3.3% and 0.2% on NQ and WQ. This demonstrates that the model performance can be enhanced by leveraging relational QA dataset guided by Grounded Relational Wiki-Graph. We provide a detailed analysis in Sec. 4.3.

KnowBERT and KEPLER encode structural knowledge into pre-trained language models. Both models focus on generating meaningful entity embedding, and are not designed to infer relations between entities for question answering. From the table, KEPLER trained via TransE performs slightly better than KnowBERT trained via entity linking, and RGPT-QA outperforms KEPLER by 2.8%, 2.1%, 5.7% on the three datasets.

Similar to RGPT-QA, Unsup.QA (Lewis et al., 2019) also generates QA data from Wikipedia. This baseline slightly improves DPR by 0.4%, 0.5%, 1.9% on the three datasets, while our RGPT-QA outperforms it by 1.8%, 1.9%, 4.4%. As discussed in Sec 3.1, one of the main reasons that
our graph-based QA generation strategy performs better is that we adopt grounded description passages $\text{desc.}(t,s)$ and $\text{desc.}(s,t)$ from different documents as questions and contexts. This avoids the lexical overlap problem in Unsup.QA and help model to capture relational facts.

We also show the retrieval and reader performance separately on validation sets in Table 3. Compared with DPR without pre-training, RGPT-QA improves top-20 accuracy of Retriever by 1.7%, 1.8%, and 3.5%, and improves EM accuracy of Reader by 1.5%, 1.1%, and 2.9%. Also, RGPT-QA outperforms all the other pre-training baselines. This shows that RGPT-QA improves both the retrieval and reader steps of open-domain QA.

**Ablation Studies.** We then analyze the importance of each model component in RGPT-QA. One key strategy is to mask out the target answer from questions and mask out source entities from passages during retrieval training. This can avoid the model using the entity surface to find the correct passage and answer. Without using masking strategy, the average EM performance drops 5.1%. This shows that it is essential to apply the mask strategy to avoid shortcut in QA pre-training. Next, we replace the hard negative passage sampling during retrieval pre-training with random batch sampling. The average EM performance drops 1.4%, showing the importance of hard negative samples. Finally, we study the unsupervised relation loss $L_{\text{distill}}$ and the supervised $L_{\text{rel}}$. Removing them leads to 0.5% and 1.3% performance drop, which shows the benefit of training the model to explicitly infer the relation from questions.

Another key component is the negative passage sampling for dense retrieval pre-training. We study how the batch size and number of negative sample influence the performance of trained retrieval. As is shown in Table 5, increasing batch size and negative sample size can improve the performance of retriever. Even with a small batch size and negative sample, our pre-training framework could still achieves better performance against non-pretrain baseline, showing that our approach is not sensitive to these two hyperparameters.

**Few-Shot QA Performance.** We analyze the improvement of RGPT-QA when only a few labelled training samples are available. We fine-tune DPR initialized by RGPT-QA on subset of Natural Questions with different percentages. As is shown in

![Figure 4: Few-shot QA experiment. Figure shows EM accuracy in validation set of DPR model with and without RGPT-QA pre-training, fine-tuned with different percentage of data on Natural Questions.](image)

Figure 4, RGPT-QA consistently outperforms DPR without pre-training, and the improvement is more significant with small data. Specifically, when only 0.5% (594) labelled QA pairs are provided, the DPR pre-trained by RGPT-QA can still achieve 26.0% Val EM accuracy, significantly higher than 9.4% achieved by the DPR without pre-training. The results show that RGPT-QA provides a good initialization for QA systems and reduce the requirement of large human-annotated QA dataset.

**4.3 Generalization for long-tail relations.**

As pointed out in Section 2, existing QA datasets suffer high relation bias, and thus a QA model trained on these datasets cannot generalize well to questions with long-tail relations. We thus analyze whether our RGPT-QA can remedy this issue. As is shown in Figure 5, the performance improvement of RGPT-QA against the supervised baseline is much more significant for the questions with infrequent relations. Specifically, for all relations appear less than 5 times in training set, the average EM accuracy of RGPT-QA is 33.3%, significantly higher than 22.4% achieved by DPR without pre-
training. This indicates that our relation QA generation method could indeed improve the performance on QA pairs with long-tail relations. Detailed prediction results are shown in Table 11 in Appendix.

5 Related Works

Unsupervised QA via Question Generation
To train a QA system without human annotation of QA pairs, Unsupervised QA has been proposed by Lewis et al. (2019) to generate synthetic \(\langle context, question, answer \rangle\) data for training QA models. Lewis et al. (2019) synthesize the QA data by: 1) run NER or noun chunkers over randomly sampled English Wikipedia paragraphs to extract answers; 2) Treat the paragraphs surrounding the answer as context; 3) Treat the context as cloze-style question and feed into a unsupervised machine translator to generate natural questions. Some follow-up works also utilize template (Fabbri et al., 2020) and pre-trained language model (Puri et al., 2020) over masked cloze-style questions for more human-readable questions. These cloze-style unsupervised QA methods achieve promising performance than previous heuristic QA baselines but underperform supervised ones. The main limitation is that the question is generated with the masked context as input, resulting in severe overlap of lexicon and word surface with the context. Consequently, the QA model might utilize the lexical pattern as a shortcut to find the answer. To address the problem of context-question lexical overlap, Dhingra et al. (2018) assume each article has an introductory paragraph, and use this paragraph to generate answer. Li et al. (2020) retrieve the Wikipedia cited document as context, Pan et al. (2020) leverage structured tables to extract key information from context, with which to synthesize questions.

To tackle the challenges in previous studies, our framework propose to leverage the Wikipedia hyperlinks and Wikidata relation triplets to construct Grounded Relational WikiGraph, based on which we generate relational QA dataset. We then pre-train a QA model to infer the latent relation from the question, and then conduct extractive QA to get the target answer entity. RGPT-QA improves the performance of the state-of-the-art QA frameworks, especially for questions with long-tail relations.

Acknowledgement

This work was partially supported by NSF III-1705169, NSF 1937599, DARPA HR00112090027, Okawa Foundation Grant, and Amazon Research Awards.
References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pages 1870–1879. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. Big self-supervised models are strong semi-supervised learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Bhuwan Dhingra, Danish Pruthi, and Dheeraj Rajagopal. 2018. Simple and effective semi-supervised question answering. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers), pages 582–587. Association for Computational Linguistics.

Bhuwan Dhingra, Manzil Zaheer, Vidhisha Balachandran, Graham Neubig, Ruslan Salakhutdinov, and William W. Cohen. 2020. Differentiable reasoning over a virtual knowledge base. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Alexander R. Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 4508–4513. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020a. REALM: retrieval-augmented language model pre-training. CoRR, abs/2002.08909.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020b. Retrieval augmented language model pre-training. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 3929–3938. PMLR.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. CoRR, abs/2004.04906.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Lijion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. Trans. Assoc. Comput. Linguistics, 7:452–466.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 6086–6096. Association for Computational Linguistics.

Patrick S. H. Lewis, Ludovic Denoyer, and Sebastian Riedel. 2019. Unsupervised question answering by cloze translation. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 4896–4910. Association for Computational Linguistics.

Zhongli Li, Wenhui Wang, Li Dong, Furu Wei, and Ke Xu. 2020. Harvesting and refining question-answer pairs for unsupervised QA. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 6719–6728. Association for Computational Linguistics.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019a. A discrete hard EM approach for weakly supervised question answering. 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 2019, Hong Kong, China, November 3-7, 2019, pages 2851–2864. Association for Computational Linguistics.

Sewon Min, Danqi Chen, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019b. Knowledge guided text retrieval and reading for open domain question answering. CoRR, abs/1911.03868.

Liangming Pan, Wenhui Chen, Wenhua Xiong, Min-Yen Kan, and William Yang Wang. 2020. Unsupervised multi-hop question answering by question generation. CoRR, abs/2010.12623.

Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 2019, Hong Kong, China, November 3-7, 2019, pages 43–54. Association for Computational Linguistics.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 5811–5826. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21:140:1–140:67.

Anshumali Shrivastava and Ping Li. 2014. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pages 2321–2329.

Haitian Sun, Pat Verga, Bhuwan Dhingra, Ruslan Salakhutdinov, and William W. Cohen. 2021. Reasoning over virtual knowledge bases with open predicate relations. CoRR, abs/2102.07043.

Pat Verga, Haitian Sun, Livio Baldini Soares, and William W. Cohen. 2020. Facts as experts: Adaptable and interpretable neural memory over symbolic knowledge. CoRR, abs/2007.00849.

Denny Vrandecic and Markus Krötzsch. 2014. Wiki-data: a free collaborative knowledgebase. Commun. ACM, 57(10):78–85.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2019. KEPLER: A unified model for knowledge embedding and pre-trained language representation. CoRR, abs/1911.06136.

Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 10684–10695. IEEE.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2020. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 1441–1451. Association for Computational Linguistics.
Table 6: Relation with grounded QA pairs of Natural Questions Training Set (Top 1-82 by frequency).

| Relation       | Frequency | Question                                                                 | True Answer                  |
|----------------|-----------|---------------------------------------------------------------------------|------------------------------|
| P161 (cast member) | 1915     | what was the geeks name in 16 candles                                     | anthony michael hall         |
| P175 (performer)   | 1844     | who sang the original blinded by the light                                 | breck springsteen           |
| P676 (lyrics by)   | 519      | who sings the song i can see clearly now the rain is gone                   | johnny nash                 |
| P66 (composer)     | 442      | who made the beavis and butthed theme song                                | mike judge                  |
| P725 (lyricist)     | 315      | who plays the voice of tiana in princess and the frog                      | ant and rose                |
| P1346 (winner)     | 283      | who has won the 2017 womens singles wimbledon tennis tournament            | garbine muguru              |
| P50 (author)       | 263      | where does the saying standing on the shoulders of giants come from        | bernard of chartres         |
| P17 (country)      | 257      | where did the black panther party take place                             | united states               |
| P527 (part)        | 198      | the unit of area in mks system is                                          | mte                         |
| P162 (producer)    | 134      | who is in the video do n ’ t worry be happy                               | bobby mcferrin              |
| P276 (location)    | 117      | where will the summer olympics be held in 2020                           | tokyo                        |
| P940 (narrative location) | 103   | what state is a christmas story based in                                   | insinua                      |
| P915 (filming location) | 98  | where was the movie the english patient filmed                             | tunisia                      |
| P710 (participant) | 88       | who died at the gunfight at okay corral                                   | billy clanton               |
| P170 (creator)     | 87       | who came up with britain ’ s got talent                                     | simon cowell                |
| P1308 (officeholder) | 87      | who is the first lady of the usa                                          | melania trump               |
| P361 (part of)     | 74       | who sings if you want to destroy my sweater                                | weezer                      |
| P39, (R: position held) | 64     | who is the attorney general for new jersey                                | gurbir gwadwal               |
| P138 (named after) | 64       | who proved that mar ’ s orbit is elliptical not circular                   | nicolaas copernicus         |
| P112 (founded by)  | 61       | who created a settlement house with the help of other social reformers    | ellen gates starr            |
| P161, (R: cast member) | 60   | who is miss sue in the blind side                                         | kathy bates                 |
| P31 (subjects of)  | 57       | the world ’ s oldest settlement house built in a poem is called the epic  | epstein gamesh              |
| P58 (screenwriter) | 57       | who wrote the story for the shape of water                                | vanessa taylor               |
| P61 (discoveror or inventor) | 55   | who developed the analytical engine which had features of present day computers | charles babbage             |
| P206 (sport)       | 53       | who does young catherine marry in wuthering heights                       | harriet earnshaw            |
| P1923 (participating team) | 52 | who did the bengals play in the super bowl                                | san francisco 49ers          |
| P166, (R: award received) | 51   | which indian actor has won the most national awards                        | amitabh bachchan            |
| P674 (characters)  | 50       | who said to reign in hell than serve in heaven                            | satan                       |
| P729 (subclass of) | 46       | when does dna replication occur during the eukaryotic cell cycle           | mimos                       |
| P36, (R: part of)  | 49       | where does the transmission of electrical impulses in the heart begin      | sinoatrial node             |
| P131 (is located in) | 46     | where is saba university school of medicine located                        | saba                        |
| P279, (R: subclass of) | 45  | what are the names of the three pedals on a piano                           | soft pedal                  |
| P54 (member of sports team) | 41 | what team does steph curry brother play                                    | dallas mavericks            |
| P3344, (R: participant in) | 38 | who won rupauls drag race all stars three                                 | trixie mattel               |
| P405 (credited to) | 37       | where was the movie snow white and the huntsman filmed                     | united kingdom              |
| P93 (position held) | 34       | who is the present speaker of lok sabha 2018                               | sumitra mahajan             |
| P127 (owned by)    | 33       | who owns the independent newspaper in the uk                               | alexander lebedev           |
| P807, (R: conflict) | 32      | in the civil war who had more soldiers                                     | union army                  |
| P31 (instance of)  | 32       | what kind of bridge is the mackinac bridge                                 | suspension bridge           |
| P1441, (R: present in work) | 29 | who is the dads name in the Adams family                                   | gomez addams                |
| P175, (R: performer) | 28     | who does sean astin play in lord of the rings                              | samwise gamgee              |
| P156 (cast member) | 25       | what is the capital of adara and nag haravel                              | silvassa                    |
| P921 (main subject) | 24       | what disease did susannah have in brain on fire                           | anti-mnda receptor encephalitis |
| P186 (material used) | 22      | what is the liquid in a magic 8 ball                                       | alcohol                     |
| P179, (R: part of the series) | 22   | what is the second book in the mortal instruments series                   | puerto rico                 |
| P793, (R: significant event) | 21 | where which territories did the us gain in the spanish-american war        | moda center                 |
| P115 (home venue)  | 21       | where does portland ’ s nba basketball team the portland trailblazers play | bobby flay                   |
| P371 (presenter)   | 21       | who won beat bobby flay shrimp and grits                                   | abraham lincoln             |
| P180 (editors of)  | 18       | who r the 4 presidents on mt . rushmore                                    | piiri reis                  |
| P900, (R: notable work) | 19  | the explorer accurately mapped the coasts of europe and north africa      | ragtime                      |
| P136 (genre)       | 17       | scott joplin is best known as a composer of what kind of music             | jord stewart                |
| P1431 (executive producer) | 16 | who hosted the daily show before trevor noah                              | uttar pradesh               |
| P47 (shares border with) | 16     | what kind of bridge is the mackinac bridge                                 | neal broten                 |
| P54, (R: member of sports team) | 16 | who scored the first goal in dallas stars history                          | gospel of mathew             |
| P1145 (adaptation) | 16       | the tribute money depicts a scene from the                               | chakri toleti               |
| P97 (director)     | 15       | who is the director of welcome to new york                                 | tom perez                   |
| P488 (chairperson) | 15       | who is the leader of the democratic party now                              | mediterranean sea            |
| P403 (watercourse outflow) | 15 | what sea does the nile river flow into                                    | option key                  |
| P86 (composer)     | 14       | how to do all codes on a mac                                               | return of the jedi           |
| P1441 (present in work) | 14 | when does skywalker find out lea is his sister                             | johnson                     |
| P734 (family name) | 14       | who threw the first brick in the stonewall riots                           | plessey v. ferguson          |
| P1269 (facet of)   | 14       | which supreme court case established the separate but equal doctrine       | southern europe             |
| P706 (takes place in) | 13     | what region of the world is greece in                                      | newport news shipbuilding   |
| P176 (manufacturer) | 13       | who built the gerald r ford aircraft carrier                              | james gibbs                 |
| P94 (architect)    | 12       | scottish architect who developed st martins in the field                   | buenos aires                 |
| P1305 (contains of) | 12      | what is the name of capital of argentina                                  | portugal                    |
| P1532 (country for sport) | 12 | cristiano ronaldo what country does he play for                           | the pickwick papers         |
| P900 (notable work) | 12      | what was the first book that charles dickens published                     | american football           |
| P641 (sport)       | 11       | what is the number 1 sport in the usa                                      | united states               |
| P1001 (applies to jurisdiction) | 11 | who won the schenck v. united states case                                 | upper new york bay          |
| P206 (on lake)     | 11       | where is ellis island located in new york                                  | microsoft                   |
| P178 (developer)   | 11       | res office 2000 was developed by which company                            | gary oldman                 |
| P166 (award received) | 11 | who won best actor in the academy awards this year                        | mahatma gandhi              |
| P102, (R: party)   | 11       | who was known as the father of indian national congress                    | cb                      |
| P449 (original broadcaster) | 10 | what chbs channel is the late late show on                                 | nick caraway                |
| P2348 (manufacturer) | 10      | who was the main character in the great gatsby                            | mca records                 |
| P264 (record label) | 10       | who did the soundtrack for beverly hills cop                               | book of genesis             |
| P674, (R: characters) | 10     | where is the story of joseph in the bible found                            | beaty                       |
| P891 (signatory)   | 10       | who has started reducing emissions from deforestation and forest degradation | pluto                        |
| P138, (R: named after) | 10     | roman god of underworld also called orcus and pluto                        | georgetown university       |
| P69 (educated at)  | 10       | who did jaren jackson senior play college basketball                      | frank abagnale              |
| P1877 (after a work by) | 10 | the movie catch me if you can is based on who                             |                              |
Table 7: Relation with grounded QA pairs of Natural Questions Training Set (Top 83-164 by frequency).
| Relation                  | Frequency | Question                                                                 | Answer          |
|---------------------------|-----------|--------------------------------------------------------------------------|-----------------|
| P7047: (R: enemy of)      | 1         | who took out the governor’s eye on walking dead                          | michonne        |
| P59: (R: constellation)  |           | brightest star in the constellation lyra dan word                        | Vega            |
| P3902: (film crew member) |           | who pioneered animated movies with his short feature steamboat willie in 1928 | Walt Disney     |
| P2348: (R: time period)  |           | the main port of azum was the red sea city                              | Enid maria remaque |
| P7365: (sister)           |           | who wrote all quire on the western art by                               | Lake victoria   |
| P469: (lakes on river)    |           | where does the water from the nile come from                            | China           |
| P205: (basin country)     |           | the main port of azum was the red sea city                              | American political science review |
| P921: (R: main subject)   |           | who began the systematic study of political science                      | Janet janov   |
| P4914: (country from)     |           | a quantity 15 m / s to the north is a measure of                        | Klavering       |
| P1411: (R: nominated for) |           | who won the first Oscar for best actress                               | Janet janov    |
| P1447: (conjugate acid)   |           | give the name and formula for the acid derived from the following anion | Chlorous acid   |
| P9766: (R: location)      |           | the area between the tigris and euphrates rivers                         | Mesopotamia     |
| P5413: (R: fielding position) | 1         | who has the most sheets in the world                                     | Meya civilisation |
| P1710: (R: participant)   |           | where the region of conflict was between karthage and karthage           | Bear mountain bridge |
| P2563: (R: superpower)    |           | which american civilization was located in a rain forest                | Mayas           |
| P2596: (culture)          |           | where does the young ones develop in humans                             | Uterus          |
| P1071: (R: location of creation) | 1       | which programming language is used in microsoft access                    | Visual basic for applications |
| P1535: (R: used by)       |           | what consoles can you play star wars battlefront on                      | Xbox one        |
| P4100 (platform)          |           | what type of arabic is spoken in palestine                               | South levantine arabic |
| P4913: (dialect of)       |           | who is the minister during the regime of chandragupta                    | Chaknaya        |
| P1066: (student of)       |           | who went before michael jordan in the draft                              | Hakeem olajuwon |
| P3342: (R: significant person) |         | who wrote the power of love celine dion                                   | Candy derouge   |
| P896: (R: composer)       |           | where did the tour de france start in 1954                              | Adelaid         |
| P1427: (start point)      |           | who is the older brother marior luzario                                   | Family guy      |
| P3733: (sibling)          |           | which family came first family man or american dad                       | Family guy      |
| P2512: (R: series spin-off) |         | where does the Appalachian Trail cross the Hudson River                 | Bear mountain bridge |
| P2505: (R: cars)          |           | what type of port is used by flash drives                                 | Ubo mass storage device class |
| P5909: (complies with)    |           | what weight class did class cianctan di aijt in the world                | Heavy weight    |
| P2994: (R: different from) |         | what is given to fats that are liquid at room temperature               | Oil             |
| P7937: (form of creative work) |         | wagner’s tristan und isold is an example of                             | Opera           |
| P252: (type of orbit)     |           | which is the orbit of the international space station                    | Low earth orbit |
| P1103: (instrument)       |           | what kind of bass does john cooper play                                  | Bass guitar     |
| P373: (R: influenced by)  |           | who are the members of 3 s mafia                                        | Juicy j         |
| P206 (R: biological process) |         | where did euze spend most of his time                                   | Montana olympus |
| P201: (lake outflow)      |           | where does the water from Lake Okchobecho drain                         | Caloshaahatee river |
| P178: (R: developer)      |           | operating system developed in 1969 at at&k’s bell laboratories           | Unix            |
| P1312: (R: facet prototype) |         | what is the opposite side of a right angle triangle                      | Hypotenuous     |
| P20: (place of death)     |           | where did omi build his new political capital                            | Simba           |
| P2936 (language used)     |           | what is the national language of saudi arabia                           | Arabic          |
| P460: (R: said to be the same as) |         | what color is a school bus yellow or orange                              | Chrome yellow   |
| P686: (R: biological process) |         | which protein is responsible for the breakdown of a fibrin clot          | Plasmacytoma    |
| P3300: (musical conductor) |         | who did the music for ready player one                                  | Alan silvestri |
| P547: (commemorates)      |           | name of ship that landed at Plymouth Rock                                | Mayflower       |
| P2709: (fabrication method) |         | the medium of the artwork that decorates the sistine chapel ceiling      | Fresco          |
| P1107: (director / manager) |         | who led the red shirts to victory in sicily                             | Giuseppe giribaldi |
| P972: (R: catalog)        |           | who is number one on americas most wanted                               | Jason derrick brown |
| P263: (R: official residence) |       | which greek god ruled over a gloomy kingdom                              | Hades           |
| P2152 (antiparticle)      |           | a packet or unit of light energy is called a                           | Photon          |
| P1462: (standards body)   |           | who is responsible for creating the standards used on the internet      | Internet engineering task force |
| P664: (R: organizer)      |           | when did they start using gloves in ufc                                 | Ufc 14          |
| P937: (work location)     |           | where did beethoven live most of his life                                | Vienna          |
| P6475: (R: appears in the form of) |         | what was robin’s name in the draft                                       | Dick grayson    |
| P2596: (R: culture)       |           | a ruined city on crete centre of the minoan bronze age civilisation     | Don bluth      |
| P2553: (production designer) |         | who made the movie all dogs go to heaven                                | Donald duck    |
| P1108: (relative)         |           | what is the first name of huey’s ‘soule and louie’s uncle               | Nbc            |
| P3301 (broadcast by)      |           | who is broadcasting the super bowl on sunday                            | Rich skrenta    |
| P943: (programmer)        |           | who wrote the first computer virus called elk colonor                   | Puerto rico     |
| P361: (R: continent)      |           | puerto rico is in north or central america                              | Impressionism   |
| P135 (movement)           |           | what kind of art did claud monet paint                                   | Cerebrum        |
| P5051 (towards)           |           | which part of the cerebral hemisphere is supplied by the middle cerebral | Dear prudence   |
| P6766: (R: lyrics by)     |           | artery what beats songs does paul play drums on                         | Dari            |
| P364 (original language)  |           | when language do they speak in kite runner                              | Cortana         |
| P1071 (location of creation) |       | a town in the netherlands known for the production of a tin glazed earthenware | Retail          |
| P400: (R: platform)       |           | name of the windows phone 8.1 virtual assistant                          | Hoang van thi   |
| P452: (industry)          |           | what did the hudson bay company do for canada                           | Herbman         |
| P598 (commands)           |           | who controlled or ordered the viet cong in combat                        | Medial condyle of tuba |
| P1103: (R: instrument)    |           | who introduced the bass clarinet as a solo instrument in jazz           | Chile           |
| P3491 (muscle insertion)  |           | what is the origin and insertion of the seminembranosus                 | Syphilis        |
| P530 (diplomatic relation) |         | which two countries are on the western border of bolivia                | Argentina      |
| P1542: (has effect)       |           | what disease is caused by bacterium treponema pallidium                 | Microsoft office 2016 |
| P1366 (territory claimed by) |       | the Falkland islands are off the coast of south american country        | Honolu         |
| P4747 (editions)          |           | what is the latest offense for mac                                     | Martha Graham  |
| P7133 (significant place) |           | on which island is the us arizona memorial                              | District line   |
| P6101: (R: highest point) |           | the highest peak in north america mt. mckinley ( or denali ) is located at the state of | Constitutional monarchy |
| P1109 (choirleader)       |           | who danced the lead role in appalachian spring                          | Charlantagne    |
| P81 (connecting line)     |           | what line is parsons green on tube map                                  | Great dividing range |
| P122 (type of government) |           | what type of government did european settlers create in south africa in 1909 | With or without you |
| P971: (R: noble title)    |           | who was crowned the first holy roman emperor                            | National museum of american history |
| P4552: (mountain range)   |           | what mountain range is the blue mountains part of                       | Chicago         |
| P658: (tracklist)         |           | what is used 2’s lead single from 'the joshua tree'                     | Chicago        |
| P1906 (collection)        |           | where is the original star spangled banner located                      | Chicago        |
| P6909 (terminal locations) |       | where does route 66 start on the east coast                             | Chicago        |
| Triple | Question | True Passage | Pred. Relation | Pred. Answer |
|--------|----------|--------------|----------------|--------------|
| (<'edward heath', '?', 'admiral’s cup'>, Wiley (eponym of corsican nuthatch which he discovered a bird new to science, written by alvin schwartz and originally illustrated by stephen gammell... | in june 1883 when he shot a specimen while on a trip in the corsican mountains... | participant in: 0.33, winner: 0.19, participant: 0.04, victory: 0.03, sport: 0.03 | admiral’s cup (/) |
| (<'scary stories to tell in the dark', '?', 'stephen gammell'>, Wiley (eponym of corsican nuthatch which he discovered a bird new to science, written by alvin schwartz and originally illustrated by stephen gammell... | in june 1883 when he shot a specimen while on a trip in the corsican mountains... | illustrator: 0.13, creator: 0.11, author: 0.07, editor: 0.02, notable work: 0.02 | stephen gammell (/) |
| (<'heeley', '?', 'sheffield tramway'>, Wiley (eponym of corsican nuthatch which he discovered a bird new to science, written by alvin schwartz and originally illustrated by stephen gammell... | heeley travelled in malacca, north borneo, located in the administrative territorial entity: 0.21, located in the administrative territorial entity: 0.14, location: 0.07, shares border with: 0.04, terminus: 0.03 old tramway (/) | sibling: 0.29, relative: 0.13, spouse: 0.04, relative: 0.04, place served by transport hub: 0.02 | patricia miller (/) |
| (<'pablo goncalvez', '?', 'patricia miller (tennis)'>, Wiley (eponym of corsican nuthatch which he discovered a bird new to science, written by alvin schwartz and originally illustrated by stephen gammell... | heeley travelled in malacca, north borneo, located in the administrative territorial entity: 0.21, located in the administrative territorial entity: 0.14, location: 0.07, shares border with: 0.04, terminus: 0.03 old tramway (/) | sibling: 0.29, relative: 0.13, spouse: 0.04, relative: 0.04, place served by transport hub: 0.02 | patricia miller (/) |
| (<'chad prakkan district', '?', 'chad prakkan'>, Wiley (eponym of corsican nuthatch which he discovered a bird new to science, written by alvin schwartz and originally illustrated by stephen gammell... | bylool kom was a notable small town in the oil rich north borneo region... | <chad prakkan district, 'P131', (? located in the administrative territorial entity), 'chad prakkan'> | chad prakkan (/) |
| (<'gothic western', '?', 'lorin morgan-richards'>, Wiley (eponym of corsican nuthatch which he discovered a bird new to science, written by alvin schwartz and originally illustrated by stephen gammell... | england:DRE was a gothic western written by lorin morgan-richards, with an element of humor... | in the young adult series, 'the goodbye family' by lorin morgan-richards has been considered gothic western with an element of humor... | lorin morgan-richards (/) |
| (<'quentin bell', '?', 'virginia nicholson'>, Wiley (eponym of corsican nuthatch which he discovered a bird new to science, written by alvin schwartz and originally illustrated by stephen gammell... | has kid of quentin bell which her father was the writer and art historian quentin bell, nephew of... | they had three children: julian bell, an artist and muralist; cressida bell, a notable textile designer; and virginia nicholson, the writer of 'charleston: a bloomberg house... | virginia nicholson (/) |
| (<'u.s. route 441 in georgia', '?', 'lakemont, georgia'>, Wiley (eponym of corsican nuthatch which he discovered a bird new to science, written by alvin schwartz and originally illustrated by stephen gammell... | it was released as the eighth single from sheeran's fourth studio album "no.6 collaborations project" (2019)... | <u.s. route 441 in georgia, 'P361', (? part of), 'no.6 collaborations project'> | no.6 collaborations project (/) |
| (<'anjelica huston', '?', 'lorin morgan-richards'>, Wiley (eponym of corsican nuthatch which he discovered a bird new to science, written by alvin schwartz and originally illustrated by stephen gammell... | she had some success with an actors studio when she starred in the"agnes browne"... | <anjelica huston, 'P87', (? directed by), 'agnes browne'> | agnes browne (/) |
| (<'caddil eldorado', '?', 'oldsmobile toronado'>, Wiley (eponym of corsican nuthatch which he discovered a bird new to science, written by alvin schwartz and originally illustrated by stephen gammell... | it was released as the eighth single from sheeran's fourth studio album "no.6 collaborations project" (2019)... | <caddil eldorado, 'P272', (? production company), 'intrepid pictures'> | intrepid pictures (/) |
| (<'caddil eldorado', '?', 'oldsmobile toronado'>, Wiley (eponym of corsican nuthatch which he discovered a bird new to science, written by alvin schwartz and originally illustrated by stephen gammell... | it was released as the eighth single from sheeran's fourth studio album "no.6 collaborations project" (2019)... | <caddil eldorado, 'P272', (? production company), 'intrepid pictures'> | intrepid pictures (/) |

Table 10: Examples of generated Relational QA datapoints and the predicted relation and answer by DPR pre-trained via RGPT-QA.
| Relation Name | Freq | Question | True Answer | RGPT-QA Prediction | Supervised DPR Prediction |
|---------------|------|----------|-------------|-------------------|--------------------------|
| R: based on   | 0    | theme song to bridge on the river kwai | the river kwai march | the river kwai march | march                     |
| subject has role | 0    | antipsychotic where is the leaning tower of pisa built | mazesin | mazesin | medicotion               |
| practiced by  | 0    | what product or market does netflix deal with | streaming media | streaming media | netflix                   |
| industry      | 1    | who does the call to prayer in islam | pisa | pisa | mosque                   |
| made from     | 2    | what is the leaning tower of pisa built | ramiyo rao | ramiyo rao | netflix                  |
| offers view on | 2    | where was the leaning tower of pisa built | neil gorsuch | neil gorsuch | goat                     |
| R: residence  | 2    | who is the founder of ramoji film city | neil gorsuch | neil gorsuch | pisa                     |
| mother        | 2    | who bore abraham first son in the bible | hagar | hagar | pisa                     |
| R: employer   | 3    | who is the youngest judge currently sitting on the u.s. supreme court | neil gorsuch | neil gorsuch | pisa                     |
| R: has part   | 3    | corpora cavernosa and corpus spongiosum are anatomic structures of | penis | penis | jammu and kashmir         |
| indigenous to | 3    | urdu is the official language of which state | pakistant | pakistant | solomon                  |
| river source  | 4    | what is the source of the colorado river | la poudre pass | la poudre pass | jharkhand                |
| river tributary | 5    | river that joins the severn near chepstow crossword | river wye | river wye | river lugg               |
| R: family     | 5    | who was the second ruler of the davidic monarchy | solomon | solomon | jeroabo                 |
| R: genre      | 6    | who is considered by many to be the father of soul | james brown | james brown | musim                   |
| R: genre      | 6    | who brought surf music to a national audience | the beach boys | the beach boys | lugg                     |
| parent organization | 8 | who owns flying j and pilot truck stops | pilot corporation | pilot corporation | lugg | |
| narrator      | 10   | who plays the mom in cheaper by the dozen | bonnie hunt | bonnie hunt | lugg                     |
| educated at   | 10   | where did the gabbie show go to college | university of pittsburgh | university of pittsburgh | jammun and kashmir        |
| director      | 15   | who did the movie i can only imagine | ajit wadekar | ajit wadekar | solomon                  |
| executive producer | 16 | who stars in the movie the quiet place | muke dinka | muke dinka | jeroabo            |
| R: player of  | 16   | pitt players in the nfl hall of fame | ruben brown | ruben brown | musim                   |
| R: player of  | 16   | who was the captain when india played its first-ever odi | muke dinka | muke dinka | musim                   |
| shares border with | 16  | what state is directly west of north dakota | austin ryan | austin ryan | musim                   |
| depicts       | 19   | who raised the american flag on two jima | abraham lincoln | abraham lincoln | jeroabo             |
| narrates      | 19   | faces of the presidents on mt. rushmore | nadiya hussain | nadiya hussain | jeroabo             |
| R: notable work | 19  | who won the 2015 great british baking show | toby buckland | toby buckland | jeroabo             |
| presenter     | 21   | who presented gardeners world from 2008 to 2010 | neil fielding | neil fielding | jeroabo             |
| presenter     | 21   | who are the new hosts of bachelor off | bo jones | bo jones | jeroabo             |
| material used | 22   | what kind of meat is on a t-bone | beef | beef | solomon                  |
| main subject  | 24   | new york times co v Sullivan held that there must be proof of | actual malice | actual malice | musim                   |
| instance of   | 29   | what kind of money do they use in russia | keelingwood | keelingwood | musim                   |
| instance of   | 29   | how does a plane wing create lift which physics concept applies | force | force | musim                   |
| country of origin | 37  | who used the springfield rifle in the civil war | united states | united states | musim                   |
| R: subclass of | 44   | what is the most abundant neurotransmitter in the nervous system | serotonin | serotonin | musim                   |
| R: subclass of | 44   | when a blood vessel is injured the first phase in hemostasis to occur is | coagulation | coagulation | musim                   |
| R: part of    | 49   | what part of south asia have the highest population densities | philippines | philippines | musim                   |
| characters    | 50   | what the settlement of the israelites in canaan is the theme of which book | philippines | philippines | musim                   |
| R: award received | 51  | most number of national awards for best actress | philippines | philippines | musim                   |
| participating team | 52  | who did melbourne beat in the 1964 grand final | battle of cajamarca | battle of cajamarca | musim                   |
| spouse        | 53   | who does jack danius end up with in sons of anarchy | joshua | joshua | musim                   |
| R: instance of | 56   | which of the following is the si unit for length | metre | metre | musim                   |
| R: instance of | 56   | what is the most abundant neurotransmitter in the nervous system | meter | meter | musim                   |
| R: instance of | 56   | pricing tactics lower the price of a product below cost | meter | meter | musim                   |
| named after   | 64   | who was saint patrick’s day named after | saint patrick | saint patrick | musim                   |
| part of       | 70   | arabian sea is the part of which ocean | indian ocean | indian ocean | musim                   |
| of              | 70   | what led to the downfall of the incan empire | lego | lego | musim                   |
| officeholder  | 88   | what is the name of the governor of new jersey | phil murphy | phil murphy | musim                   |
| participant    | 88   | what two groups were fighting in the chinese civil war | phil murphy | phil murphy | musim                   |
| participant    | 88   | who played the superbowl halftime show last year | phil murphy | phil murphy | musim                   |
| participant    | 88   | who came second in the overall ranked of the tour de france last year | phil murphy | phil murphy | musim                   |
| filming location | 98   | what city does the terminator take place in | los angeles | los angeles | musim                   |
| filming location | 98   | where was back to the future three filmed | monument valley | monument valley | musim                   |

Table 11: Comparison of the prediction of DPR initialized by RGPT-QA with DPR without pre-training. These are all samples that two models made different predictions, and the relation frequency in the training set is less than 100.
| Question                                                                 | Predicted Answer               | True Answer               | Match? | Predicted Relation |
|-------------------------------------------------------------------------|--------------------------------|---------------------------|--------|--------------------|
| Who played mr darling on andy griffith show                             | Denver Pyle                    | Denver Pyle               | ✓      | P175 (performer)   |
| Who voices flik in a bug’s life                                         | Dave Foley                     | Dave Foley                | ✓      | P725 (voice actor) |
| Who’s the dad of blair waldorf’s baby                                   | Chuck                          | Chuck                     | ✓      | P26 (spouse)       |
| Where do you think glaciers can be found today                          | rocky mountains                | mountain ranges on every continent | ✗      | P31, (R: instance of) |
| When did ginny weasley join the quidditch team                          | half-blood prince              | half-blood prince         | ✓      | P674, (R: characters) |
| When does far cry 5 for ps4 come out                                    | 2018                           | march 27, 2018            | ✗      | P400 (game platform) |
| When do millennials end and gen z start                                 | mid-1990s to mid-2000s         | mid-1990s                 | ✗      | P155 (preceded by)  |
| Who killed hotchner’s wife in criminal minds                             | George Foyet                   | George Foyet              | ✓      | P7047 (enemy of)    |
| Who said walk tall and carry a big stick                                | u.s. president theodore roosevelt | theodore roosevelt      | ✗      | P170 (creator)     |
| Who does the voice of sheen from jimmy neutron                          | Jeffrey Garcia                 | Jeffrey Garcia            | ✓      | P734 (family name)  |
| How many seasons of gossip girl are there                               | 6                              | 6                         | ✓      | P527 (has part)     |
| What can be used to detect the charge of particles                      | ionization detectors           | particle detector         | ✗      | P279, (R: subclass of) |
| Who was robin in the original batman series                             | Burt Ward                      | Burt Ward                 | ✓      | P161 (cast member)  |
| What is the song funky cold medina about                                 | a love potion                  | a fictional aphrodisiac   | ✗      | P138 (named after)  |
| What do you call a quarter pounder in france                            | royal cheese                   | royal cheese              | ✓      | P1889 (different from) |
| Who developed the first alternating current electric system             | Galileo Ferraris               | Nikola Tesla              | ✗      | P61 (inventor)      |
| Who won s5 of rupaul’s drag race                                        | Jinkx Monsoon                  | Jinkx Monsoon             | ✓      | P1346 (winner)      |
| When was the svalbard global seed vault built                            | 2006                           | 2006                      | ✓      | P88 (built for)     |
| Who was the mother of dragons married to                                 | Khal Drogo                     | Dothraki Horselord Khal Drogo | ✗      | P26 (spouse)       |
| Which organization sets monetary policy for the united states            | the federal reserve            | the federal reserve       | ✓      | P1001 (jurisdiction) |
| What season of the voice was miley cyrus on                              | eleventh                       | season 11                 | ✗      | P179 (part of series) |
| Upon which document in american history is the language of the declaration of sentiments based | united states declaration of independence | united states declaration of independence | ✓      | P144 (based on)     |
| What kind of car does dale earnhardt jr drive                           | chevrolet camaro               | chevrolet                 | ✗      | P54 (played for)    |
| How many times did brazil win the fifa world cup                         | five                           | five                      | ✓      | P1344 (participant in) |
| Second life is an example of a                                           | massively multiplayer online role-playing games | an online virtual world | ✗      | P31 (instance of)   |
| What percentage of the world’s population lives in east asia             | 22%                            | 22%                       | ✓      | P276, (R: located in) |
| From which body part shurpnakha drive her name                           | fingernails                    | fingernails               | ✓      | P186 (ingredient)   |
| How many chapters does the gospel of john have                           | four                           | 21                        | ✗      | P527 (has part)     |
| Who sang the original always on my mind                                  | bj. thomas                     | gwen mccrae               | ✗      | P175 (performer)    |
| Where does the amazon river start and finish                            | atlantic ocean                 | atlantic ocean            | ✓      | P403 (watercourse outflow) |
| Who did dwayne wade play for last year                                   | the miami heat                 | miami heat                | ✗      | P647 (drafted by)   |
| Who owns the rights to the power rangers                                 | Hasbro                         | Hasbro                    | ✓      | P127 (owned by)     |

Table 12: Predicted relations for those QA pairs in Natural Questions Valid Set that cannot be aligned to WikiData.