Improving Question Answering with External Knowledge

Xiaoman Pan¹*  Kai Sun²*  Dian Yu³  Jianshu Chen³  
Heng Ji¹  Claire Cardie²  Dong Yu³  
¹University of Illinois at Urbana-Champaign, Champaign, IL, USA  
²Cornell University, Ithaca, NY, USA  
³Tencent AI Lab, Bellevue, WA, USA

Abstract

We focus on multiple-choice question answering (QA) tasks in subject areas such as science, where we require both broad background knowledge and the facts from the given subject-area reference corpus. In this work, we explore simple yet effective methods for exploiting two sources of external knowledge for subject-area QA. The first enriches the original subject-area reference corpus with relevant text snippets extracted from an open-domain resource (i.e., Wikipedia) that cover potentially ambiguous concepts in the question and answer options. As in other QA research, the second method simply increases the amount of training data by appending additional in-domain subject-area instances.

Experiments on three challenging multiple-choice science QA tasks (i.e., ARC-Easy, ARC-Challenge, and OpenBookQA) demonstrate the effectiveness of our methods: in comparison to the previous state-of-the-art, we obtain absolute gains in accuracy of up to 8.1%, 13.0%, and 12.8%, respectively. While we observe consistent gains when we introduce knowledge from Wikipedia, we find that employing additional QA training instances is not uniformly helpful: performance degrades when the added instances exhibit a higher level of difficulty than the original training data. As one of the first studies on exploiting unstructured external knowledge for subject-area QA, we hope our methods, observations, and discussion of the exposed limitations may shed light on further developments in the area.

1 Introduction

To answer questions relevant to a given text (e.g., a document or a book), human readers often rely on a certain amount of broad background knowledge obtained from sources outside of the text (McNamara et al., 2004; Salmerón et al., 2006). It is perhaps not surprising then, that machine readers also require knowledge external to the text itself to perform well on question answering (QA) tasks.

We focus on multiple-choice QA tasks in subject areas such as science, in which facts from the given reference corpus (e.g., a textbook) need to be combined with broadly applicable external knowledge to select the correct answer from the available options (Clark et al., 2016, 2018; Mihaylov et al., 2018). For convenience, we call these subject-area QA tasks.

Table 1: A sample problem from a multiple-choice QA task OpenBookQA (Mihaylov et al., 2018) in a scientific domain (✓: correct answer option).

<table>
<thead>
<tr>
<th>Question: a magnet will stick to</th>
<th>A. a belt buckle. ✓</th>
<th>B. a wooden table.</th>
<th>C. a plastic cup.</th>
<th>D. a paper plate.</th>
</tr>
</thead>
</table>

To correctly answer the question in Table 1, for example, scientific facts¹ from the provided reference corpus — {“a magnet attracts magnetic metals through magnetism” and “iron is always magnetic”}, as well as general world knowledge extracted from an external source such as {“a belt buckle is often made of iron” and “iron is metal”} are required. Thus, these QA tasks provide suitable testbeds for evaluating external knowledge exploitation and integration.

Previous subject-area QA methods (e.g., (Khot et al., 2017; Zhang et al., 2018; Zhong et al., 2018)) explore many ways of exploiting structured knowledge. Recently, we have seen that the framework of fine-tuning a pre-trained language model (e.g., GPT (Radford et al., 2018) and BERT (Devlin et al., 2019)) outperforms previous state-of-

* Equal contribution. This work was conducted when the two authors were at Tencent AI Lab, Bellevue, WA.

¹Ground truth facts are usually not provided in this kind of question answering tasks.
the-art methods (Mihaylov et al., 2018; Ni et al., 2019). However, it is still not clear how to incorporate different sources of external knowledge, especially unstructured knowledge, into this powerful framework to further improve subject-area QA.

We investigate two sources of external knowledge (i.e., open-domain and in-domain), which have proven effective for other types of QA tasks, by incorporating them into a pre-trained language model during the fine-tuning stage. First, we identify concepts in question and answer options and link these potentially ambiguous concepts to an open-domain resource that provides unstructured background information relevant to the concepts and used to enrich the original reference corpus (Section 2.2). In comparison to previous work (e.g., (Yadav et al., 2019)), we perform information retrieval based on the enriched corpus instead of the original one to form a document for answering a question. Second, we increase the amount of training data by appending additional in-domain subject-area QA datasets (Section 2.3).

We conduct experiments on three challenging multiple-choice science QA tasks where existing methods stubbornly continue to exhibit performance gaps in comparison with humans: ARC-Easy, ARC-Challenge (Clark et al., 2016, 2018), and OpenBookQA (Mihaylov et al., 2018), which are collected from real-world science exams or carefully checked by experts. We fine-tune BERT (Devlin et al., 2019) in a two-step fashion (Section 2.1). We treat entire Wikipedia as the open-domain external resource (Section 2.2) and all the evaluated science QA datasets (question-answer pairs and reference corpora) except the target one as in-domain external resources (Section 2.3). Experimental results show that we can obtain absolute gains in accuracy of up to 8.1%, 13.0%, and 12.8%, respectively, in comparison to the previous published state-of-the-art, demonstrating the effectiveness of our methods. We also analyze the gains and exposed limitations. While we observe consistent gains by introducing knowledge from Wikipedia, employing additional in-domain training data is not uniformly helpful: performance degrades when the added data exhibit a higher level of difficulty than the original training data (Section 3).

To the best of our knowledge, this is the first work to incorporate external knowledge into a pre-trained model for improving subject-area QA. Besides, our promising results emphasize the importance of external unstructured knowledge for subject-area QA. We expect there is still much scope for further improvements by exploiting more sources of external knowledge, and we hope the present empirical study can serve as a new starting point for researchers to identify the remaining challenges in this area.

2 Method

In this section, we first introduce our BERT-based QA baseline (Section 2.1). Then, we present how we incorporate external open-domain (Section 2.2) and in-domain (Section 2.3) sources of knowledge into the baseline.

2.1 Baseline Framework

Given a question $q$, an answer option $o_i$, and a reference document $d_j$, we concatenate them with @ and # as the input sequence $@d_j#@q#@o_i#$ to BERT (Devlin et al., 2019), where @ and # stand for the classifier token [CLS] and sentence separator token [SEP] in BERT, respectively. A segmentation A embedding is added to every token before $q$ (exclusive) and a segmentation B embedding to every other token, where A and B are learned during the language model pretraining of BERT. For each instance in the ARC (Easy and Challenge) and OpenBookQA tasks, we use Lucene (McCandless et al., 2010) to retrieve up to top $K$ sentences using the non-stop words in $q$ and $o_i$ as the query and then concatenate the retrieved sentences to form $d_i$ (Sun et al., 2019). The final prediction for each question is obtained by a linear plus softmax layer over the output of the final hidden state of the first token in each input sequence.

By default, we employ the following two-step fine-tuning approach unless explicitly specified. Following previous work (Sun et al., 2019) based on GPT (Radford et al., 2018), we first fine-tune BERT (Devlin et al., 2019) on a large-scale multiple-choice machine reading comprehension dataset RACE (Lai et al., 2017) collected from English-as-a-foreign-language exams, which provides a ground truth reference document instead of a reference corpus for each question. Then, we further fine-tune the model on the target multiple-choice science QA datasets. For convenience, we call the model obtained after the first fine-tuning phase as a pre-fine-tuned model.
Question: Mercury, the planet nearest to the Sun, has extreme surface temperatures, ranging from 465°C in sunlight to −180°C in darkness. Why is there such a large range of temperatures on Mercury?

A. The planet is too small to hold heat.
B. The planet is heated on only one side.
C. The planet reflects heat from its dark side.
D. The planet lacks an atmosphere to hold heat. ✓

Table 2: A sample problem from the ARC-Challenge dataset (Clark et al., 2018) (✓: correct answer option).

2.2 Utilization of External Knowledge from an Open-Domain Resource

Just as human readers activate their background knowledge related to the text materials (Kendeou and Van Den Broek, 2007), we link concepts identified in questions and answer options to an open-domain resource (i.e., Wikipedia) and provide machine readers with unstructured background information relevant to these concepts, used to enrich the original reference corpus.

Concept Identification and Linking: We first extract concept mentions from texts. Most mention extraction systems (e.g., Manning et al. (2014)) are trained using pre-defined classes in general domains such as PERSON, LOCATION, and ORGANIZATION. However, in ARC and OpenBookQA, the vast majority of mentions are from scientific domains (e.g., “rotation”, “revolution”, “magnet”, and “iron”). Therefore, we simply consider all noun phrases as candidate concept mentions, which are extracted by a noun phrase chunker. For example, in the sample problem in Table 2, we extract concept mentions such as “Mercury”.

Then each concept mention is disambiguated and linked to its corresponding concept (page) in Wikipedia. For example, the ambiguous concept mention “Mercury” in Table 2 should be linked to the concept Mercury (planet) rather than Mercury (element) in Wikipedia. For concept disambiguation and linking, we simply adopt an existing unsupervised approach (Pan et al., 2015) that first selects high quality sets of concept collaborators to feed a simple similarity measure (i.e., Jaccard) to link concept mentions.

Reference Corpus Enrichment: We apply concept identification and linking to the text of all questions and answer options. Then, for each linked concept, we extract Wikipedia sentences that contain this concept and all sentences from the Wikipedia article of this concept without removing redundant information. For example, the following sentence in the Wikipedia article of Mercury (planet) is extracted: “Having almost no atmosphere to retain heat, it has surface temperatures that vary diurnally more than on any other planet in the Solar System.”, which can serve as a reliable piece of evidence to infer the correct answer option D for the question in Table 2.

Most previous methods (Khashabi et al., 2017; Musa et al., 2018; Ni et al., 2019; Yadav et al., 2019) perform information retrieval on the reference corpus to retrieve relevant sentences to form reference documents. In contrast, we retrieve relevant sentences from the combination of an open-domain resource and the original reference corpus to generate a reference document for each (question, answer option) pair. We still keep up to top K sentences for each reference document (Section 2.1). See the framework overview in Figure 1.
2.3 Utilization of External Knowledge from In-Domain Data

Since there are a relatively small number of training instances available for a single subject-area QA task (see Table 3), instead of fine-tuning a pre-fine-tuned model on a single target dataset, we also investigate into fine-tuning a pre-fine-tuned model on multiple in-domain datasets simultaneously. For example, when we train a model for ARC-Challenge, we use the training set of ARC-Challenge together with the training, development, and test sets of ARC-Easy and OpenBookQA. We also explore two settings with and without merging the reference corpora from different tasks. We introduce more details and discussions in Section 3.2 and Section 3.6.

3 Experiments and Discussions

3.1 Datasets

In our experiment, we use RACE (Lai et al., 2017) — the largest existing multiple-choice machine reading comprehension dataset collected from real and practical language exams — in the pre-fine-tuning stage. Questions in RACE focus on evaluating linguistic knowledge acquisition of participants and are commonly used in previous methods (Wang et al., 2018a; Sun et al., 2019).

We evaluate the performance of our methods on three multiple-choice science QA datasets: ARC-Easy, ARC-Challenge, and OpenBookQA. ARC-Challenge and ARC-easy originate from the same set of exam problems collected from multiple sources. ARC-Challenge contains questions answered incorrectly by both a retrieval-based method and a word co-occurrence method, and the remaining questions form ARC-Easy. Questions in OpenBookQA are crowdsourced by turkers and then carefully filtered and modified by experts. See the statistics of these datasets in Table 3. Note that for OpenBookQA, we do not utilize the accompanying auxiliary reference knowledge bases to ensure a fair comparison with previous work.

3.2 Experimental Settings

For the two-step fine-tuning framework, we use the uncased BERT\textsubscript{LARGE} released by Devlin et al. (2019) as the pre-trained language model. We set the batch size to 24, learning rate to $2 \times 10^{-5}$, and the maximal sequence length to 512. When the input sequence length exceeds 512, we truncate the longest sequence among $q$, $o_i$, and $d_i$ (defined in Section 2.1). We first fine-tune BERT\textsubscript{LARGE} for five epochs on RACE to get the pre-fine-tuned model and then further fine-tune the model for eight epochs on the target QA datasets in scientific domains. We show the accuracy of the pre-fine-tuned model on RACE in Table 4.

We use the noun phrase chunker in spaCy\textsuperscript{2} to extract concept mentions. For information retrieval, we use the version 7.4.0 of Lucene (McCandless et al., 2010) and set the maximum number of the retrieved sentences $K$ to 50. We use the stop word list from NLTK (Bird and Loper, 2004).

In addition, we design two slightly different settings for information retrieval. In setting 1, the original reference corpus of each dataset is independent. Formally, for each dataset $x \in D$, we perform information retrieval based on the corresponding original reference corpus of $x$ and/or the external corpus generated based on problems in $x$, where $D = \{\text{ARC-Easy, ARC-Challenge, OpenBookQA}\}$. In setting 2, all original reference corpora are integrated to further leverage external in-domain knowledge. Formally, for each dataset $x \in D$, we conduct information retrieval based on the given reference corpus of $D$ and/or the external corpus generated based on problems in $D$ instead of $x$.

3.3 Baselines

Here we only briefly introduce three baselines (i.e., GPT\textsuperscript{H}, RS\textsuperscript{H}, and BERT\textsuperscript{H}) that all fine-tune a

\begin{table}[h]
\centering
\begin{tabular}{l|c|c|c|c}
\hline
Dataset & Train & Dev & Test & Total \\
\hline
RACE & 87,866 & 4,887 & 4,934 & 97,687 \\
ARC-Easy & 2,251 & 570 & 2,376 & 5,197 \\
ARC-Challenge & 1,119 & 299 & 1,172 & 2,590 \\
OpenBookQA & 4,957 & 500 & 500 & 5,957 \\
\hline
\end{tabular}
\caption{The number of questions in RACE and the multiple-choice subject-area QA datasets for evaluation: ARC-Easy, ARC-Challenge, and OpenBookQA.}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{l|c|c|c}
\hline
Dataset & Dev & Test \\
\hline
RACE-M & 76.7 & 76.6 \\
RACE-H & 71.0 & 70.1 \\
RACE-M + RACE-H & 72.7 & 72.0 \\
\hline
\end{tabular}
\caption{Accuracy (%) of the pre-fine-tuned model on the RACE dataset, which contains two subsets: RACE-M and RACE-H, representing problems collected from middle and high school language exams, respectively.}
\end{table}

\textsuperscript{2}https://spacy.io/.
\textsuperscript{3}https://github.com/nlpdata/external.
<table>
<thead>
<tr>
<th>Method</th>
<th>ARC-E</th>
<th>ARC-C</th>
<th>OBQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR (Clark et al., 2018)</td>
<td>62.6</td>
<td>20.3</td>
<td>–</td>
</tr>
<tr>
<td>Odd-One-Out (Mihaylov et al., 2018)</td>
<td>–</td>
<td>–</td>
<td>50.2</td>
</tr>
<tr>
<td>DGEM (Khot et al., 2018)</td>
<td>59.0</td>
<td>27.1</td>
<td>24.4</td>
</tr>
<tr>
<td>KG² (Zhang et al., 2018)</td>
<td>–</td>
<td>31.7</td>
<td>–</td>
</tr>
<tr>
<td>AIR (Yadav et al., 2018)</td>
<td>58.4</td>
<td>26.6</td>
<td>–</td>
</tr>
<tr>
<td>NCRF+++ (Musa et al., 2018)</td>
<td>52.2</td>
<td>33.2</td>
<td>–</td>
</tr>
<tr>
<td>TriAN+++ (Zhang et al., 2018)</td>
<td>–</td>
<td>33.4</td>
<td>–</td>
</tr>
<tr>
<td>Two Stage Inference (Pirtoaca et al., 2019)</td>
<td>61.1</td>
<td>26.9</td>
<td>–</td>
</tr>
<tr>
<td>ET-RR (Ni et al., 2019)</td>
<td>–</td>
<td>36.6</td>
<td>–</td>
</tr>
<tr>
<td>GPTII (Radford et al., 2018; Sun et al., 2019)</td>
<td>57.0</td>
<td>38.2</td>
<td>52.0</td>
</tr>
<tr>
<td>RSII (Sun et al., 2019)</td>
<td>66.6</td>
<td>40.7</td>
<td>55.2</td>
</tr>
</tbody>
</table>

Table 5: Accuracy (%) on the test sets of ARC-Easy, ARC-Challenge, and OpenBookQA datasets. RACE is used in the pre-fine-tuning stage for all the tasks (Section 2.1). MD stands for fine-tuning on multiple target datasets simultaneously (Section 2.3). All results are single-model performance. GPTII, RSII, and BERTII are baselines that use two-step fine-tuning (Section 3.3). ARC-E: ARC-Easy; ARC-C: ARC-Challenge; OBQA: OpenBookQA.

We see consistent improvements in accuracy across all tasks after we enrich the reference corpus with relevant texts from Wikipedia to form new reference documents (i.e., RC + EC and IRC + IEC in Table 5). Moreover, using only the extracted external corpus to perform information retrieval for reference document generation can achieve reasonable performance compared to using the original reference corpus, especially on the OpenBookQA dataset (62.2% vs. 64.8% under setting 1 and 63.0% vs. 65.0% under setting 2). This indicates that we can extract reliable and relevant texts from external open-domain resources such as Wikipedia via linked concepts mentioned in Section 2.2. Moreover, using the integrated corpus (i.e., setting 2) consistently boosts the performance. Since the performance in setting 2 (integrated corpus) is better than that in setting 1 (independent corpus) based on our experiments, we take setting 2 by default for discussions unless explicitly specified.

We see further improvements on ARC-Challenge and OpenBookQA, by fine-tuning the pre-fine-tuned model on multiple target datasets (i.e., ARC-Easy, ARC-Challenge, and OpenBookQA). However, we do not see a similar gain on ARC-Easy by increasing the number of in-domain training instances. We will further discuss it in Section 3.6.

Our best models (i.e., IRC + IEC for ARC-
### Question

**What boils at the boiling point?**

- A. Kool-Aid. ✓
- B. Cotton.
- C. Paper Towel.
- D. Hair.

**Sentence(s) From Wikipedia**

*Kool-Aid* is known as Nebraska’s official soft drink. Common types of drinks include plain drinking water, milk, coffee, tea, hot chocolate, juice and *soft drinks*.

### Question

**Forest fires** occur in many areas due to **drought conditions**. If the drought conditions continue for a long period of time, which might cause the repopulation of trees to be threatened?

- A. a decrease in the thickness of soil. ✓
- B. a decrease in the amount of erosion.
- C. an increase in the bacterium population.
- D. an increase in the production of oxygen and fire.

**Sentence(s) From Wikipedia**

It is highly resistant to *drought conditions*, and provides excellent fodder; and has also been used in controlling *soil erosion*, and as revegetator; often after *forest fires*.

### Question

Juan and LaKeisha roll a few objects down a ramp. They want to see which object rolls the farthest. What should they do so they can repeat their investigation?

- A. Put the objects in groups.
- B. Change the height of the ramp.
- C. Choose different objects to roll.
- D. Record the details of the investigation. ✓

**Sentence(s) From Wikipedia**

The use of measurement developed to allow recording and comparison of observations made at different times and places, by different people.

### Question

Which statement best explains why the sun appears to *move across the sky* each day?

- A. The sun revolves around Earth. ✓
- B. Earth rotates around the sun.
- C. The sun revolves on its axis.
- D. Earth rotates on its axis. ✓

**Sentence(s) From Wikipedia**

*Earth’s rotation* about its *axis* causes the fixed stars to apparently *move across the sky* in a way that depends on the observer’s latitude.

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer Options</th>
<th>Sentence(s) From Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What boils at the boiling point?</strong></td>
<td>A. Kool-Aid. ✓, B. Cotton, C. Paper Towel, D. Hair</td>
<td><em>Kool-Aid</em> is known as Nebraska’s official soft drink. Common types of drinks include plain drinking water, milk, coffee, tea, hot chocolate, juice and <em>soft drinks</em>.</td>
</tr>
<tr>
<td><strong>Forest fires</strong> occur in many areas due to <strong>drought conditions</strong>. If the drought conditions continue for a long period of time, which might cause the repopulation of trees to be threatened?</td>
<td>A. a decrease in the thickness of soil. ✓, B. a decrease in the amount of erosion, C. an increase in the bacterium population, D. an increase in the production of oxygen and fire</td>
<td>It is highly resistant to <em>drought conditions</em>, and provides excellent fodder; and has also been used in controlling <em>soil erosion</em>, and as revegetator; often after <em>forest fires</em>.</td>
</tr>
<tr>
<td>Juan and LaKeisha roll a few objects down a ramp. They want to see which object rolls the farthest. What should they do so they can repeat their investigation?</td>
<td>A. Put the objects in groups, B. Change the height of the ramp, C. Choose different objects to roll, D. Record the details of the investigation. ✓</td>
<td>The use of measurement developed to allow recording and comparison of observations made at different times and places, by different people.</td>
</tr>
<tr>
<td>Which statement best explains why the sun appears to <em>move across the sky</em> each day?</td>
<td>A. The sun revolves around Earth. ✓, B. Earth rotates around the sun, C. The sun revolves on its axis, D. Earth rotates on its axis. ✓</td>
<td><em>Earth’s rotation</em> about its <em>axis</em> causes the fixed stars to apparently <em>move across the sky</em> in a way that depends on the observer’s latitude.</td>
</tr>
</tbody>
</table>

Table 6: Examples of corrected errors using the reference corpus enriched by the sentences from Wikipedia.

Table 7: Percentage (%) of retrieved sentences from each source. Wiki: Wikipedia; Total: total number of retrieved sentences for all (question, answer option) pairs in a single task. ARC-Easy and ARC-Challenge share the same original reference corpus.

<table>
<thead>
<tr>
<th>Task</th>
<th>Wiki</th>
<th>OBQA</th>
<th>ARC</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARC-E</td>
<td>20.8</td>
<td>0.4</td>
<td>78.7</td>
<td>1,039,059</td>
</tr>
<tr>
<td>ARC-C</td>
<td>21.5</td>
<td>0.4</td>
<td>78.2</td>
<td>517,846</td>
</tr>
<tr>
<td>OBQA</td>
<td>20.6</td>
<td>1.1</td>
<td>78.3</td>
<td>1,191,347</td>
</tr>
</tbody>
</table>

Table 6 shows some examples of errors produced by IRC (Table 5) that do not leverage external knowledge from open-domain resources. These errors can be corrected by enriching the reference corpus with external sentences extracted from Wikipedia (IRC + IEC in Table 5). In the first example, the correct answer option “Kool-Aid” never appears in the original reference corpus. As a result, without external background knowledge, it is less likely to infer that “Kool-Aid” refers to liquid (can boil) here.

In addition to performing information retrieval on the enriched reference corpus, we investigate an alternative approach that uses concept identification and linking to directly enrich the reference document for each (question, answer option) pair. More specifically, we apply concept identification and linking to each (question, answer option) pair (q, o) and extract sentences from Wikipedia based on the linked concepts. These extracted sentences are appended to the reference documents d_i of (q, o_i) directly. We still keep up to K (i.e., 50) sentences per document. We observe that this direct appending approach generally cannot outperform the reference corpus enrichment approach described in Section 2.2.

We report the statistics of the sentences (without redundancy removal) extracted from each source in Table 7, used as inputs to our methods IRC + IEC and IRC + IEC + MD in Table 5. As the original reference corpus of OpenBookQA is made up of 1,326 sentences, fewer retrieved sentences are extracted from its reference corpus for all tasks compared to other sources.

### 3.5 Impact of External Knowledge from an Open-Domain Resource

Table 6 shows some examples of errors produced by IRC (Table 5) that do not leverage external knowledge from open-domain resources. These errors can be corrected by enriching the reference corpus with external sentences extracted from Wikipedia (IRC + IEC in Table 5). In the first example, the correct answer option “Kool-Aid” never appears in the original reference corpus. As a result, without external background knowledge, it is less likely to infer that “Kool-Aid” refers to liquid (can boil) here.

In addition to performing information retrieval on the enriched reference corpus, we investigate an alternative approach that uses concept identification and linking to directly enrich the reference document for each (question, answer option) pair. More specifically, we apply concept identification and linking to each (question, answer option) pair (q, o) and extract sentences from Wikipedia based on the linked concepts. These extracted sentences are appended to the reference documents d_i of (q, o_i) directly. We still keep up to K (i.e., 50) sentences per document. We observe that this direct appending approach generally cannot outperform the reference corpus enrichment approach described in Section 2.2.

We report the statistics of the sentences (without redundancy removal) extracted from each source in Table 7, used as inputs to our methods IRC + IEC and IRC + IEC + MD in Table 5. As the original reference corpus of OpenBookQA is made up of 1,326 sentences, fewer retrieved sentences are extracted from its reference corpus for all tasks compared to other sources.

### 3.6 Impact of External Knowledge from In-Domain Data

Compared to fine-tuning the pre-fine-tuned model on a single multiple-choice subject-area QA
dataset, we observe improvements in accuracy by fine-tuning on multiple in-domain datasets (MD) simultaneously (Section 2.3) for ARC-Challenge and OpenBookQA. In particular, we see a dramatic gain on the ARC-Challenge dataset (from 46.1% to 53.7%) as shown in Table 5.

However, MD leads to a performance drop on ARC-Easy. We hypothesize that other commonly adopted approaches may also lead to performance drops. To verify that, we explore another way of utilizing external knowledge for ARC-Easy by first fine-tuning the pre-fine-tuned model for four epochs on external in-domain data (i.e., ARC-Challenge, OpenBookQA, or ARC-Challenge + OpenBookQA) and then further fine-tuning for four epochs on ARC-Easy. As shown in Table 8, we also observe that compared to only fine-tuning on ARC-Easy, fine-turning on external in-domain data hurts the performance. The consistent performance drops across the two methods of using MD on ARC-Easy are perhaps due to an intrinsic property of the tasks themselves – the question-answer instances in ARC-Easy are relatively simpler than those in ARC-Challenge and OpenBookQA. Introducing relatively complex problems from ARC-Challenge and OpenBookQA may hurt the final performance on ARC-Easy. As mentioned earlier, compared to questions in ARC-Easy, questions in ARC-Challenge are less likely to be answered correctly by retrieval-based or word co-occurrence methods. We argue that questions in the ARC-Challenge tend to require more external knowledge for reasoning, similar to the observation of Sugawara et al. (2018) (30.0% vs. 20.0%).

3.7 Discussions about Question Types and Remaining Challenges

We use the human annotations such as required reasoning skills (i.e., word matching, paraphrasing, knowledge, meta/whole, and math/whole) and validity of questions in ARC-Easy and ARC-Challenge, which are released by Sugawara et al. (2018).

We also analyze the instances that our approach fails to answer correctly in the OpenBookQA development set to study the remaining challenges. It might be promising to identify the relations among concepts within an answer option. For example, our current model mistakenly selects the answer option "the sun orbits the earth" as associated with the question "Revolution happens when?" probably because "sun", "orbits", and "earth" frequently co-occur in our generated reference document, though these concepts such as "revolution" are successfully linked to their corresponding Wikipedia pages in the astronomy field.

Besides, we might also need to identify causal

<table>
<thead>
<tr>
<th>First 4</th>
<th>Last 4</th>
<th>Accuracy</th>
<th># Epochs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARC-C</td>
<td>ARC-E</td>
<td>69.4</td>
<td>8</td>
</tr>
<tr>
<td>OBQA</td>
<td>ARC-E</td>
<td>70.9</td>
<td>8</td>
</tr>
<tr>
<td>ARC-C + OBQA</td>
<td>ARC-E</td>
<td>72.6</td>
<td>8</td>
</tr>
<tr>
<td>ARC-E</td>
<td></td>
<td>72.9</td>
<td>4</td>
</tr>
<tr>
<td>ARC-E</td>
<td></td>
<td>74.7</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 8: Accuracy (%) on the ARC-Easy test set. The first four epochs are fine-tuned using the dataset(s) in the first column. The last four epochs are fine-tuned using the dataset in the second column. # Epochs: the total number of epochs.

Table 9: Accuracy (%) by different categories on the annotated test sets of ARC-Easy and ARC-Challenge, which are released by Sugawara et al. (2018).
relations between events. For example, given the question “The type of climate change known as anthropogenic is caused by this”, our model mistakenly predicts another answer option “forest fires” with its associated contexts “climate change has caused the island to suffer more frequent severe droughts, leading to large forest fires”, instead of the real cause “humanity” supported by “the problem now is with anthropogenic climate change—that is, climate change caused by human activity, which is making the climate change a lot faster than it normally would”.

3.8 Discussions about Pre-Fine-Tuning

Previous work (Devlin et al., 2019) has shown that fine-tuning BERT\textsubscript{LARGE} on small datasets can be sometimes unstable. Additionally, Sun et al. (2019) show that fine-tuning GPT (Radford et al., 2018) that is pre-fine-tuned on RACE can dramatically improve the performance of relatively small multiple-choice tasks. Here we only use the BERT\textsuperscript{11} baseline for a brief discussion. We have a similar observation: we can obtain more stable performance on the target datasets by first fine-tuning BERT on RACE (language exams), and we see consistent performance improvements on all the evaluated science QA datasets. As shown in Figure 2, we see that the performance drops dramatically without using pre-fine-tuning on the RACE dataset.

4 Related Work

4.1 Subject-Area QA Tasks and Methods

As there is not a clear distinction between QA and machine reading comprehension (MRC) tasks, for convenience we call a task in which there is no reference document provided for each instance as a QA task. In this paper, we focus on multiple-choice subject-area QA tasks, where the in-domain reference corpus does not provide sufficient relevant content on its own to answer a significant portion of the questions (Clark et al., 2016; Kobayashi et al., 2017; Welbl et al., 2017; Clark et al., 2018; Mihaylov et al., 2018). In contrast to other types of QA scenarios (Nguyen et al., 2016; Dhingra et al., 2017; Joshi et al., 2017; Dunn et al., 2017; Kwiatkowski et al., 2019), in this setting: (1) the reference corpus does not reliably contain text spans from which the answers can be drawn, and (2) it does not provide sufficient information on its own to answer a significant portion of the questions. Thus they are suitable for us to study how to exploit external knowledge for QA.

Our work follows the general framework of discriminatively fine-tuning a pre-trained language model such as GPT (Radford et al., 2018) and BERT (Devlin et al., 2019) on QA tasks (Radford et al., 2018; Devlin et al., 2019; Hu et al., 2019; Yang et al., 2019). As shown in Table 5, the baseline based on BERT already outperforms previous state-of-the-art methods designed for subject-area QA tasks (Yadav et al., 2018; Pirtoaca et al., 2019; Ni et al., 2019; Sun et al., 2019).

4.2 Utilization of External Knowledge for Subject-Area QA

Previous studies have explored many ways to leverage structured knowledge to solve questions in subject areas such as science exams. Many researchers investigate how to directly or indirectly use automatically constructed knowledge bases/graphs from reference corpora (Khot et al., 2017; Kwon et al., 2018; Khashabi et al., 2018; Zhang et al., 2018) or existing external general knowledge graphs (Li and Clark, 2015; Sachan et al., 2016; Wang et al., 2018a,c; Zhong et al., 2018; Musa et al., 2018) such as ConceptNet (Speer et al., 2017). However, for subject-area QA, unstructured knowledge is seldom considered in previous studies, and it is still not clear the usefulness of this kind of knowledge.

As far as we know, for subject-area QA tasks, this is the first attempt to impart sources of external unstructured knowledge into one state-of-the-art pre-trained language model, and we are among the first to investigate the effectiveness of the ex-
ternal unstructured texts in Wikipedia (Pirtoaca et al., 2019) and additional in-domain QA data.

4.3 Utilization of External Knowledge for Other Types of QA and MRC

For both QA and MRC tasks in which the majority of answers are extractive such as SQuAD (Rajpurkar et al., 2016) and TriviaQA (Joshi et al., 2017), previous work has shown that it is useful to introduce external open-domain QA instances and textural information from Wikipedia by first retrieving relevant documents in Wikipedia and then running a MRC model to extract a text span from the documents based on the question (Chen et al., 2017; Wang et al., 2018b; Kratzwald and Feurerriegel, 2018; Lee et al., 2018; Lin et al., 2018).

Based on Wikipedia, we apply concept identification and linking to enrich QA reference corpora, which has not been explored before. Compared to previous data argumentation studies for other types of QA tasks (Yu et al., 2018), differences exist in: 1) we focus on in-domain data and discuss the impacts of the difficulties of additional in-domain instances on a target task; 2) we are the first to show it is useful to merge reference corpora from different in-domain subject-area QA tasks.

5 Conclusion and Future Work

We focus on how to incorporate external knowledge into a pre-trained model to improve subject-area QA tasks that require background knowledge. We exploit two sources of external knowledge through: enriching the original reference corpus with relevant texts from open-domain Wikipedia and using additional in-domain QA datasets (instances and reference corpora) for training. Experimental results on ARC-Easy, ARC-Challenge, and OpenBookQA show the effectiveness of our simple method. The promising results also demonstrate the importance of unstructured external knowledge for subject-area QA. In the future, we plan to jointly exploit various types of external unstructured and structured knowledge.

Acknowledgments

We thank the anonymous reviewers for their constructive and helpful feedback.

References


Panayiota Kendeou and Paul Van Den Broek. 2007. The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. Memory & cognition, 35(7).


Yang Li and Peter Clark. 2015. Answering elementary science questions by constructing coherent scenes using background knowledge. In Proceedings of the EMNLP, Lisbon, Portugal.


Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In Proceedings of the ACL (Demonstrations), Baltimore, MD.


Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the EMNLP, Austin, TX.


