Improving Question Answering with External Knowledge

Xiaoman Pan[1*] **Kai Sun**[2*] **Dian Yu**[3] **Jianshu Chen**[3]
Heng Ji[1] **Claire Cardie**[2] **Dong Yu**[3]
[1]University of Illinois at Urbana-Champaign, Champaign, IL, USA
[2]Cornell University, Ithaca, NY, USA
[3]Tencent 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.

Question: a magnet will stick to __?			
A.	a belt buckle. ✓	**B**.	a wooden table.
C.	a plastic cup.	**D**.	a paper plate.

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

To correctly answer the question in Table 1, for example, scientific facts[1] 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 intergration.

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.

[1]Ground 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_i, we concatenate them with @ and # as the input sequence $@d_i\#q\#o_i\#$ 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**.

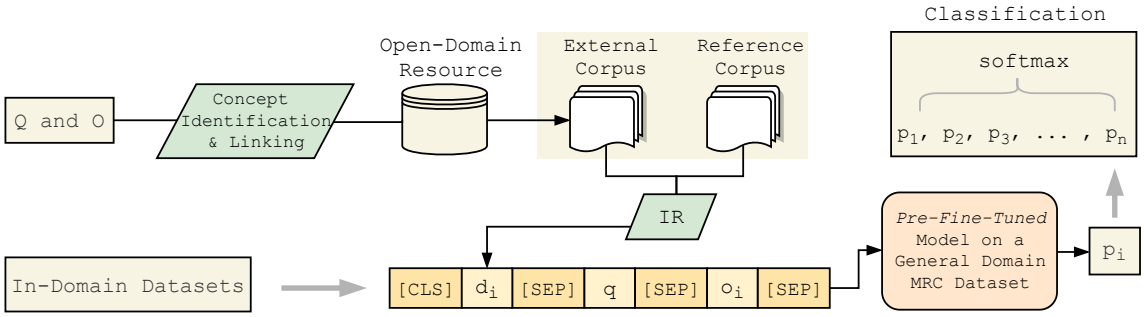

Figure 1: Overview of our framework (IR: information retrieval; MRC: machine reading comprehension). Q, O, q, o_i, d_i, and n denote the set of all questions, the set of all answer options, a question, one of the answer options associated with question q, the document (formed by retrieved sentences) associated with the (q, o_i) pair, and the number of answer options of q, respectively.

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 domain 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 $\text{BERT}_{\text{LARGE}}$ released by Devlin et al. (2019) as the pre-trained language model. We set the batch size to 24, learning rate to 2×10^{-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 $\text{BERT}_{\text{LARGE}}$ 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[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 $D = \{\text{ARC-Easy}, \text{ARC-Challenge}, \text{OpenBookQA}\}$. 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].

Dataset	Train	Dev	Test	Total
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

Table 3: The number of questions in RACE and the multiple-choice subject-area QA datasets for evaluation: ARC-Easy, ARC-Challenge, and OpenBookQA.

Dataset	Dev	Test
RACE-M	76.7	76.6
RACE-H	71.0	70.1
RACE-M + RACE-H	72.7	72.0

Table 4: Accuracy (%) of the pre-fine-tuned model on the RACE dataset, which contains two subsets: RACE-M and RACE-H, representing problems collected from **m**iddle and **h**igh school language exams, respectively.

3.3 Baselines

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

[2] https://spacy.io/.

[3] https://github.com/nlpdata/external.

Method	ARC-E	ARC-C	OBQA
IR (Clark et al., 2018)	62.6	20.3	–
Odd-One-Out (Mihaylov et al., 2018)	–	–	50.2
DGEM (Khot et al., 2018)	59.0	27.1	24.4
KG^2 (Zhang et al., 2018)	–	31.7	–
AIR (Yadav et al., 2018)	58.4	26.6	–
NCRF++ (Musa et al., 2018)	52.2	33.2	–
TriAN++ (Zhong et al., 2018)	–	33.4	–
Two Stage Inference (Pirtoaca et al., 2019)	61.1	26.9	–
ET-RR (Ni et al., 2019)	–	36.6	–
GPT$^{\text{II}}$ (Radford et al., 2018; Sun et al., 2019)	57.0	38.2	52.0
RS$^{\text{II}}$ (Sun et al., 2019)	66.6	40.7	55.2
Our BERT-Based Implementations			
Setting 1			
Reference Corpus (RC) (i.e., BERT$^{\text{II}}$)	71.9	44.1	64.8
External Corpus (EC)	65.0	39.4	62.2
RC + EC	73.3	45.0	65.2
Setting 2			
Integrated Reference Corpus (IRC)	73.2	44.8	65.0
Integrated External Corpus (IEC)	68.9	40.1	63.0
IRC + IEC	**74.7**	46.1	67.0
IRC + MD	69.4	50.7	67.4
IRC + IEC + MD	72.3	**53.7**	**68.0**
Human Performance	–	–	91.7

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 **m**ultiple target **d**atasets simultaneously (Section 2.3). All results are single-model performance. GPT$^{\text{II}}$, RS$^{\text{II}}$, and BERT$^{\text{II}}$ are baselines that use two-step fine-tuning (Section 3.3). ARC-E: ARC-Easy; ARC-C: ARC-Challenge; OBQA: OpenBookQA.

pre-trained language model on downstream tasks without substantial modifications to model architectures, which achieve remarkable success on many question answering tasks. Following the two-step fine-tuning framework (Section 2.1), **all** three strong baselines use RACE in the first fine-tuning stage for a fair comparison. We will discuss the impacts of pre-fine-tuning on baseline model performance in Section 3.8, noting that pre-fine-tuning is not the contribution of this work.

GPT$^{\text{II}}$: This baseline is based on fine-tuning a generative pre-trained transformer (GPT) language model (Radford et al., 2018) instead of BERT (Devlin et al., 2019).

RS$^{\text{II}}$: Based on GPT, general reading strategies (RS) (Sun et al., 2019) are applied during the fine-tuning stage such as adding a trainable embedding into the text embedding of tokens relevant to the question and candidate answer options.

BERT$^{\text{II}}$: Based on BERT, this baseline is an exact implementation described in Section 2.1.

3.4 Main Results

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	Answer Options	Sentence(s) From Wikipedia
What boils at the boiling point?	A. ***Kool-Aid***. ✓ B. Cotton. C. Paper Towel. D. Hair.	***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***.
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.	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***.
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***. ✓	The use of measurement developed to allow ***recording*** and comparison of ***observations*** made at different times and places, by different people.
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***. ✓	***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 6: Examples of corrected errors using the reference corpus enriched by the sentences from Wikipedia.

Easy and IRC + IEC + MD for ARC-Challenge and OpenBookQA) outperform the strong baseline BERT$^{\text{II}}$ introduced in Section 2.1 (74.7% vs. 71.9% on ARC-Easy, 53.7% vs. 44.1% on ARC-Challenge, and 68.0% vs. 64.8% on OpenBookQA), which already beats the previous state-of-the-art model RS$^{\text{II}}$. In the remaining sections, we analyze our models and discuss the impacts of external knowledge from various aspects.

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_i) 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.

Task	Wiki	OBQA	ARC	Total
ARC-E	20.8	0.4	78.7	1,039,059
ARC-C	21.5	0.4	78.2	517,846
OBQA	20.6	1.1	78.3	1,191,347

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.

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

First 4	Last 4	Accuracy	# Epochs
ARC-C	ARC-E	69.4	8
OBQA	ARC-E	70.9	8
ARC-C + OBQA	ARC-E	72.6	8
ARC-E	-	72.9	4
ARC-E	ARC-E	74.7	8

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.

Question Type	ARC-E		ARC-C	
	BERT^{II}	Ours	BERT^{II}	Ours
Word Matching	81.3	**85.4**	30.4	**73.9**
Paraphrasing	90.9	90.9	46.7	**66.7**
Knowledge	58.3	**83.3**	44.4	**55.6**
Math/Logic	100.0	100.0	33.3	33.3
Valid	80.0	**86.0**	36.1	**66.7**
Invalid	50.0	**80.0**	41.7	41.7
Easy	80.0	**90.0**	33.3	**53.3**
Hard	70.0	**80.0**	43.3	**60.0**

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).

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 released by Sugawara et al. (2018) to analyze the impacts of external knowledge on instances in various categories. Sixty instances are annotated for each dataset. We refer readers to Sugawara et al. (2018) for detailed definitions of each category. We do not report the accuracy for *math/whole* as no annotated question in ARC belongs to this category.

We compare the BERT^{II} baseline in Table 5 that only uses the original reference corpus of a given end task with our best model. As shown in Table 9, by leveraging external knowledge from in-domain datasets (instances and reference corpora) and open-domain texts, we observe consistent improvements on most of the categories. Based on these experimental results on the annotated subset, we may assume it could be a promising direction to further improve challenging multiple-choice subject-area QA tasks through exploiting high-quality external knowledge besides designing task-specific models for different types of questions (Clark et al., 2016).

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"* 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

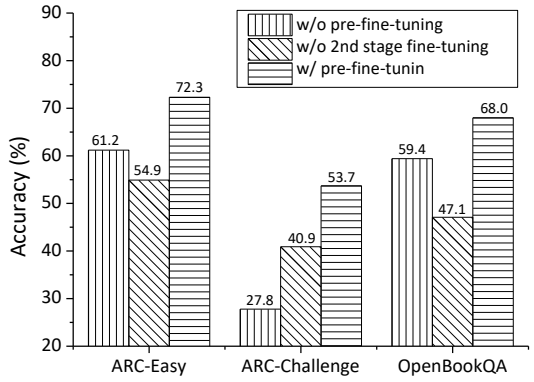

Figure 2: Accuracy (%) on the test sets of evaluation tasks with and without the pre-fine-tuning stage (2nd stage fine-tuning: fine-tune the pre-fine-tuned model on target science question answering datasets).

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 $\text{BERT}_{\text{LARGE}}$ 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^{II} 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 textual 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 Feuerriegel, 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

Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In *Proceedings of the ACL (Demonstrations)*, Barcelona, Spain.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the ACL*, Vancouver, Canada.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint*, cs.CL/1803.05457v1.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of the AAAI*, Phoenix, AZ.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proeedings of the NAACL-HLT*, Minneapolis, MN.

Bhuwan Dhingra, Kathryn Mazaitis, and William W Cohen. 2017. Quasar: Datasets for question answering by search and reading. *arXiv preprint*, cs.CL/1707.03904v2.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint*, cs.CL/1704.05179v3.

Minghao Hu, Yuxing Peng, Zhen Huang, Nan Yang, Ming Zhou, et al. 2019. Read+Verify: Machine reading comprehension with unanswerable questions. In *Proceedings of the AAAI*, Honolulu, HI.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint*, cs.CL/1705.03551v2.

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).

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2017. Learning what is essential in questions. In *Proceedings of the CoNLL 2017*.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2018. Question answering as global reasoning over semantic abstractions. In *Proceedings of the AAAI*, New Orleans, LA.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2017. Answering complex questions using open information extraction. In *Proceedings of the ACL*, Vancouver, Canada.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI*, New Orleans, LA.

Mio Kobayashi, Ai Ishii, Chikara Hoshino, Hiroshi Miyashita, and Takuya Matsuzaki. 2017. Automated historical fact-checking by passage retrieval, word statistics, and virtual question-answering. In *Proceedings of the IJCNLP*, Taipei, Taiwan.

Bernhard Kratzwald and Stefan Feuerriegel. 2018. Adaptive document retrieval for deep question answering. In *Proceedings of the EMNLP*, Brussels, Belgium.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A benchmark for question answering research. *TACL*.

Heeyoung Kwon, Harsh Trivedi, Peter Jansen, Mihai Surdeanu, and Niranjan Balasubramanian. 2018. Controlling information aggregation for complex question answering. In *Proceedings of the ECIR*, Grenoble, France.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations. In *Proceedings of the EMNLP*, Copenhagen, Denmark.

Jinhyuk Lee, Seongjun Yun, Hyunjae Kim, Miyoung Ko, and Jaewoo Kang. 2018. Ranking paragraphs for improving answer recall in open-domain question answering. In *Proceedings of the EMNLP*, Brussels, Belgium.

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

Yankai Lin, Haozhe Ji, Zhiyuan Liu, and Maosong Sun. 2018. Denoising distantly supervised open-domain question answering. In *Proceedings of the ACL*, Melbourne, Australia.

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.

Michael McCandless, Erik Hatcher, and Otis Gospodnetic. 2010. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co., Greenwich, CT.

Danielle S McNamara, Irwin B Levinstein, and Chutima Boonthum. 2004. iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, 36(2).

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the EMNLP*, Brussels, Belgium.

Ryan Musa, Xiaoyan Wang, Achille Fokoue, Nicholas Mattei, Maria Chang, Pavan Kapanipathi, Bassem Makni, Kartik Talamadupula, and Michael Witbrock. 2018. Answering science exam questions using query rewriting with background knowledge. *arXiv preprint*, cs.AI/1809.05726v1.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint*, cs.CL/1611.09268v2.

Jianmo Ni, Chenguang Zhu, Weizhu Chen, and Julian McAuley. 2019. Learning to attend on essential terms: An enhanced retriever-reader model for open-domain question answering. In *Proceedings of the NAACL-HLT*, Minneapolis, MN.

Xiaoman Pan, Taylor Cassidy, Ulf Hermjakob, Heng Ji, and Kevin Knight. 2015. Unsupervised entity linking with abstract meaning representation. In *Proceedings of the NAACL-HLT*, Denver, CO.

George-Sebastian Pirtoaca, Traian Rebedea, and Stefan Ruseti. 2019. Improving retrieval-based question answering with deep inference models. *arXiv preprint*, cs.CL/1812.02971v2.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. In *Preprint*.

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.

Mrinmaya Sachan, Avinava Dubey, and Eric P Xing. 2016. Science question answering using instructional materials. In *Proceedings of the ACL*, Berlin, Germany.

Ladislao Salmerón, Walter Kintsch, and José J Caãs. 2006. Reading strategies and prior knowledge in learning from hypertext. *Memory & Cognition*, 34(5).

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI*, San Francisco, CA.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the EMNLP*, Brussels, Belgium.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019. Improving machine reading comprehension with general reading strategies. In *Proceedings of the NAACL-HLT*, Minneapolis, MN.

Liang Wang, Meng Sun, Wei Zhao, Kewei Shen, and Jingming Liu. 2018a. Yuanfudao at SemEval-2018 Task 11: Three-way attention and relational knowledge for commonsense machine comprehension. In *Proceedings of the SemEval*, New Orleans, LA.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerald Tesauro, Bowen Zhou, and Jing Jiang. 2018b. R^3: Reinforced reader-ranker for open-domain question answering. In *Proceedings of the AAAI*, New Orleans, LA.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, et al. 2018c. Improving natural language inference using external knowledge in the science questions domain. *arXiv preprint*, cs.CL/1809.05724v2.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the W-NUT*, Copenhagen, Denmark.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019. Alignment over heterogeneous embeddings for question answering. In *Proceedings of the NAACL-HLT*, Minneapolis, MN.

Vikas Yadav, Rebecca Sharp, and Mihai Surdeanu. 2018. Sanity check: A strong alignment and information retrieval baseline for question answering. In *Proceedings of the ACM SIGIR*, Ann Arbor, MI.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *arXiv preprint*, cs.CL/1902.01718v1.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining local convolution with global self-attention for reading comprehension. In *Proceedings of the ICLR*, Vancouver, Canada.

Yuyu Zhang, Hanjun Dai, Kamil Toraman, and Le Song. 2018. KGˆ 2: Learning to reason science exam questions with contextual knowledge graph embeddings. *arXiv preprint*, cs.LG/1805.12393v1.

Wanjun Zhong, Duyu Tang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2018. Improving question answering by commonsense-based pre-training. *arXiv preprint*, cs.CL/1809.03568v1.