Contextualized Representations Using Textual Encyclopedic Knowledge

Mandar Joshi[*] [†] **Kenton Lee** [ϵ] **Yi Luan** [ϵ] **Kristina Toutanova** [ϵ]

[†] Allen School of Computer Science & Engineering, University of Washington, Seattle, WA
{mandar90}@cs.washington.edu

[ϵ] Google Research, Seattle
{kentonl,luanyi,kristout}@google.com

Abstract

We present a method to represent input texts by contextualizing them jointly with dynamically retrieved textual encyclopedic background knowledge from multiple documents. We apply our method to reading comprehension tasks by encoding questions and passages together with background sentences about the entities they mention. We show that integrating background knowledge from *text* is effective for tasks focusing on factual reasoning and allows direct reuse of powerful pretrained BERT-style encoders. Moreover, knowledge integration can be further improved with suitable pretraining via a self-supervised masked language model objective over words in background-augmented input text. On TriviaQA, our approach obtains improvements of 1.6 to 3.1 F1 over comparable RoBERTa models which do not integrate background knowledge dynamically. On MRQA, a large collection of diverse QA datasets, we see consistent gains in-domain along with large improvements out-of-domain on BioASQ (2.1 to 4.2 F1), TextbookQA (1.6 to 2.0 F1), and DuoRC (1.1 to 2.0 F1).

1 Introduction

Current self-supervised representations, trained at large scale from document-level contexts, are known to encode linguistic (Tenney et al., 2019) and factual (Petroni et al., 2019) knowledge into their deep neural network parameters. Yet, even large pretrained representations are not able to capture and preserve all factual knowledge they have "read" during pretraining due to the long tail of entity and event-specific information (Logan et al., 2019). Previous work has focused on specialized architectures to integrate background knowledge, typically from structured knowledge bases (Weissenborn et al., 2017; Bauer et al., 2018; Mihaylov and Frank, 2018; Yang et al., 2019; Zhang et al., 2019; Peters et al., 2019). While integrating structured knowledge can enable precise inference, the coverage of knowledge in structured resources is often limited.

[*]Work completed while interning at Google

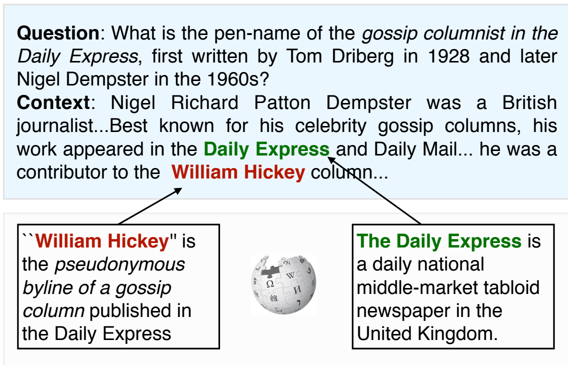

Figure 1: An example from TriviaQA showing how background sentences from Wikipedia help define the meaning of phrases in the context and their relationship to phrases in the question. The answer *William Hickey* is connected to the question phrase *pen-name of the gossip columnist in the Daily Express* through the background sentence.

We posit that representations should be able to integrate *textual* background knowledge since a wider scope of information is more readily available in textual form. Our method represents input texts by jointly encoding them with *dynamically retrieved* sentences from the Wikipedia pages of entities they mention. We term these representations TEK-enriched, for Textual Encyclopedic Knowledge (Figure 2 shows an illustration), and use them for reading comprehension by contextualizing questions and passages together with retrieved Wikipedia background sentences. Such background knowledge can help reason about the relationships between questions and passages. Figure 1 shows an example question from the TriviaQA dataset (Joshi et al., 2017) asking for *the*

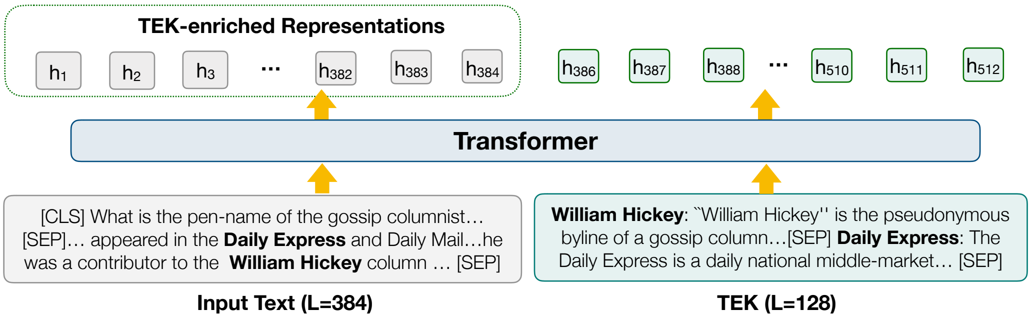

Figure 2: We contextualize the input text, in this case a question and a passage, together with textual encyclopedic knowledge (TEK) using a pretrained Transformer to create TEK-enriched representations.

pen-name of a gossip columnist. Encoding relevant background knowledge (*pseudonymous byline of a gossip column published in the Daily Express*) helps ground the vague reference to *the William Hickey column* in the given document context.

Using dynamically retrieved text as background knowledge allows us to directly reuse powerful pretrained BERT-style encoders (Devlin et al., 2019). We show that an off-the-shelf RoBERTa (Liu et al., 2019b) model can be directly finetuned on minimally structured TEK-enriched inputs, which are formatted to allow the encoder to distinguish between the original passages and background sentences. This method considerably improves on current state-of-the art methods which only consider context from a single input document (Section 4). The improvement comes without an increase in the length of the input window for the Transformer (Vaswani et al., 2017).

Although existing pretrained models provide a good starting point for task-specific TEK-enriched representations, there is still a mismatch between the type of input seen during pretraining (single document segments) and the type of input the model is asked to represent for downstream tasks (document text with background Wikipedia sentences from multiple pages). We show that our Transformer model can be substantially improved by reducing this mismatch via self-supervised masked language model (Devlin et al., 2019) pretraining on TEK-augmented input texts.

Our approach records considerable improvements over state of the art base (12-layer) and large (24-layer) Transformer models for in-domain and out-of-domain document-level extractive QA, for tasks where factual knowledge about entities is important and well-covered by the background collection. On TriviaQA, we see improvements of 1.6 to 3.1 F1, respectively, over comparable RoBERTa models which do not integrate background information. On MRQA (Fisch et al., 2019), a large collection of diverse QA datasets, we see consistent gains in-domain along with large improvements out-of-domain on BioASQ (2.1 to 4.2 F1), TextbookQA (1.6 to 2.0 F1), and DuoRC (1.1 to 2.0 F1).

2 TEK-Enriched Representations

We follow recent work on pretraining bidirectional Transformer representations on unlabeled text, and finetuning the representations for downstream tasks (Devlin et al., 2019). Subsequent approaches have shown significant improvements over BERT by improving the training example generation, the masking strategy, the pretraining objectives, and the optimization methods (Liu et al., 2019b; Joshi et al., 2020). We build on these improvements to train TEK-enriched representations and use them for document-level extractive QA.

Our approach seeks to contextualize input text $X = (x_1, \ldots, x_n)$ jointly with relevant textual encyclopedic background knowledge B retrieved dynamically from multiple documents. We define a retrieval function, $f_{ret}(X, \mathcal{D})$, which takes X as input and retrieves from the corpus $\mathcal{D}$, a list of text spans $B = (B_1, \ldots, B_M)$. In our implementation, each of the text spans B_i is a sentence. The encoder then represents X by jointly encoding X with B using $f_{enc}(X, B)$ such that the output representations of X are cognizant of the information present in B (see Figure 2). We use a deep Transformer encoder for $f_{enc}(\cdot)$. The Transformer takes as input the sequence `[CLS]` X `[SEP]` B `[SEP]`.

We refer to inputs X we would like to represent generically as *contexts*, which could be either contiguous word sequences from documents (passages), or, for the QA application, question-passage pairs, which we refer to as *RC-contexts*. For a fixed

Transformer input length limit (which is necessary for computational efficiency), there is a trade-off between the length of the document context (the length of X) and the amount of background knowledge (the length of B). This trade-off is explored in detail in Section 5 where our experiments show that for an encoder input limit of 512, the values of $N_C = 384$ for the length of X and $N_B = 128$ for the length of the background provide an effective compromise.

We use a simple implementation of the background retrieval function $f_{ret}(X, \mathcal{D})$, using an off-the-shelf entity linker or Wikipedia hyperlinks, and a way to score the relevance of individual sentences using *ngram* overlap. We begin by specifying the retrieval function we use for RC-contexts followed by the QA model which uses TEK-enriched representations in Section 2.1. We then detail baseline and newly proposed self-supervised pretraining methods for our representations in Section 2.2.

2.1 Question Answering with TEK-Enriched Representations

The input X for the extractive QA task consists of the question Q and a candidate passage P. We use the following retrieval function $f_{ret}(X)$ to obtain relevant background B.

Background Knowledge Retrieval for QA

1. We detect entity mentions in X using a proprietary Wikipedia-based entity linker. [1]

2. We form a candidate pool of background segments B_i as the union of the sentences in the Wikipedia pages of the detected entities. We rank the sentences based on their number of overlapping ngrams with the question (equally weighted unigrams, bigrams, and trigrams).

To form the input for the Transformer encoder, each background sentence is minimally structured as B_i by prepending the name of the entity whose page it belongs to along with a separator ':' token. Each sentence B_i is followed by `[SEP]`. Appendix A shows an example of an RC-context with background knowledge segments.

QA Model Following BERT, we use a simple QA architecture by training two linear classifiers independently on top of the output representations of X for predicting the answer span boundary (start and end). We assume that the answer, if present, is contained only in the given passage, P, and do not consider potential mentions of the answer in the background B. For instances which do not contain the answer, we simply set the answer span to be the special token `[CLS]`. We use a fixed Transformer input window size of 512, and use a sliding window with a stride of 128 tokens to handle longer documents. Our TEK-enriched representations use document passages of length 384 while baselines use longer passages of length 512.

2.2 Pretraining for TEK-Enriched Representations

Standard pretraining uses contiguous document-level natural language inputs. Since TEK-augmented inputs are formatted as natural language sequences, off-the-shelf pretrained models can be used as a starting point for creating TEK-enriched representations. As one of our approaches, we use the standard publicly released RoBERTa checkpoint as the pretraining method.

While the input format is the same, there is nonetheless a mismatch between contiguous document segments and TEK-augmented inputs sourced from multiple documents. We propose an additional pretraining stage—starting from the RoBERTa parameters, we resume pretraining using an MLM objective on TEK-augmented document text X, which encourages the model to integrate the knowledge from multiple background segments.

Background Knowledge Retrieval in Pretraining During pretraining, X represents a contiguous block of text from Wikipedia. The retrieval function $f_{ret}(X, \mathcal{D})$ returns $B = (B_1, \ldots, B_M)$ where each B_i is a sentence from the Wikipedia page of some entity linked from a span in X. For efficiency, in pretraining we do not use an entity linker, but follow Wikipedia hyperlinks. The background candidate sentences are ranked based on their ngram overlap with X. We use the top ranking sentences in B up to N_B tokens. If no entities are found in X, B is constructed from the context following X from the same document without any additional preprocessing.

Training Objective We pretrain a deep Transformer using a masked language model objective (Devlin et al., 2019) after initializing the parameters with pretrained RoBERTa weights. We follow improvements in SpanBERT (Joshi et al., 2020) and mask spans with lengths sampled from a geometric

[1] Work in progress aims to evaluate the impact of the entity linker and compare to publicly available resources.

Task	Train	Dev	Test
TQA Wiki	61,888	7,993	7,701
TQA Web	528,979	68,621	65,059
MRQA	616,819	58,221	9,633

Table 1: Data statistics for TriviaQA and MRQA.

distribution in the entire input (X and B). We use a single segment ID, and remove the next sentence prediction objective which has been shown to not improve performance (Joshi et al., 2020; Liu et al., 2019b) for multiple tasks including QA.

We evaluate two methods building textual-knowledge enriched representations for QA, differing in the pretraining approach used: (a) TEK_F which uses TEK-enriched representations, finetuned for QA starting from parameters pretrained with the standard RoBERTa method, and (b) TEK_{PF}, our full approach, which finetunes TEK-enriched QA representations initialized with our TEK-augmented text pretraining method. The subscript P stands for pretraining and F stands for finetuning. The next sections describe our experimental study, which focuses on assessing the novel aspects of the representations on multiple reading comprehension tasks.

3 Experimental Setup

We evaluate our models on two extractive question answering benchmarks, TriviaQA and MRQA.

TriviaQA TriviaQA (Joshi et al., 2017) is a large scale QA dataset of trivia questions paired with evidence collected via entity linking and web search. The dataset is *distantly supervised* in that the answers are contained in the evidence but the context may not support answering the questions. We experiment with both the Wikipedia and Web task formulations. For this benchmark, we follow the input preprocessing of Clark and Gardner (2018). The input to our model is the concatenation of the first four 400-token passages selected by their linear passage ranker. For training, we define the gold span to be the first occurrence of the gold answer(s) in the context (Joshi et al., 2017; Talmor and Berant, 2019).

MRQA The MRQA shared task (Fisch et al., 2019) consists of several widely used QA datasets unified into a common format aimed at evaluating out-of-domain generalization. The data consists of a training set, in-domain and out-of-domain development sets, and a private out-of-domain test set. The training set and the in-domain development set consist of modified versions of corresponding sets from SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2017), SearchQA (Dunn et al., 2017), TriviaQA Web (Joshi et al., 2017), HotpotQA (Yang et al., 2018) and Natural Questions (Kwiatkowski et al., 2019). The out-of-domain test evaluation, including access to questions and passages, is only available through Codalab[2]. Due to the complexity of our system which involves entity linking and retrieval, we perform development and model selection on the in-domain dev set and treat the out-of-domain dev set as the test set. The out-of-domain set we evaluate on has examples from BioASQ (Tsatsaronis et al., 2015), DROP (Dua et al., 2019), DuoRC (Saha et al., 2018), RACE (Lai et al., 2017), RelationExtraction (Levy et al., 2017), and TextbookQA (Kembhavi et al., 2017).

Table 1 shows the number of training, development, and test examples for each task, with TriviaQA Web (dev and test) and MRQA in-domain dev having well over 55K examples.

3.1 Implementation

We implemented all models in TensorFlow (Abadi et al., 2015). For pretraining, we used the 12-layer RoBERTa-base and 24-layer RoBERTa-large configurations, and initialized the parameters from their respective checkpoints. In TEK-augmented pretraining, we further pretrained the models for 200K steps with a batch size of 512 and BERT's triangular learning rate schedule with a warmup of 5000 steps on TEK-augmented contexts. We used a peak learning rate of 0.0001 for base and $5e^{-5}$ for large models. For finetuning hyperparameters, see Appendix B.

3.2 Baselines

Our full approach TEK_{PF} consists of two stages: (a) 200K steps of TEK-pretraining on Wikipedia starting from the RoBERTa checkpoint, and (b) finetuning and doing inference on RC-contexts augmented with TEK background. TEK_F replaces the first specialized pretraining stage with 200K steps for standard single-document-context pretraining for a fair comparison with TEK_{PF}. We compare TEK_{PF} and TEK_F with two baselines,

[2] https://worksheets.codalab.org/worksheets/0x926e37ac8b4941f793bf9b9758cc01be/

	TQA Wiki		TQA Web	
	EM	F1	EM	F1
Previous work				
Clark and Gardner (2018)	64.0	68.9	66.4	71.3
Weissenborn et al. (2017)	64.6	69.9	67.5	72.8
Wang et al. (2018)	66.6	71.4	68.6	73.1
Lewis (2018)	67.3	72.3	-	-
This work				
RoBERTa (Base)	66.7	71.7	77.0	81.4
RoBERTa++ (Base)	68.0	72.9	76.8	81.4
TEK$_F$ (Base)	70.0	74.8	78.2	83.0
TEK$_{PF}$ (Base)	**71.2**	**76.0**	**78.8**	**83.4**
RoBERTa (Large)	72.3	76.9	80.6	85.1
RoBERTa++ (Large)	72.9	77.5	81.1	85.5
TEK$_F$(Large)	74.1	78.6	82.2	86.5
TEK$_{PF}$ (Large)	**74.6**	**79.1**	**83.0**	**87.2**

Table 2: Test set performance on TriviaQA.

RoBERTa and RoBERTa++. Both use the same architecture as our approach, but use only original RC-contexts for finetuning and inference, and use standard single-document-context RoBERTa pretraining. TEK$_{PF}$ uses $N_C = 384$ and $N_B = 128$, while both baselines use $N_C = 512$ and $N_B = 0$.

RoBERTa We finetune the model on question answering data without knowledge augmentation starting from the same RoBERTa checkpoint that is used as an initializer for TEK-augmented pretraining.

RoBERTa++ For a fair evaluation of the new TEK-augmented pretraining method while controlling for the number of pretraining steps and other hyperparameters, we extend RoBERTa's pretraining for an additional 200K steps on single contiguous blocks of text (without background information). We use the same masking and other hyperparameters as in TEK-augmented pretraining.

4 Results

TriviaQA Table 2 compares our approaches with baselines and previous work. The 12-layer variant of our RoBERTa baseline outperforms or matches the performance of several previous systems including ELMo-based ones (Wang et al., 2018; Lewis, 2018) which are specialized for this task. We also see that RoBERTa++ outperforms RoBERTa, indicating that in spite of large scale pretraining, there is still room for improvement by simply pretraining for more steps on task-domain relevant text. Furthermore, the 12-layer and 24-layer variants of our TEK$_F$ approach considerably improve over a comparable RoBERTa++ baseline for both Wikipedia (1.9 and 1.1 F1 respectively) and Web (1.6 and 1.0 F1 respectively) indicating that TEK representations are useful even without additional TEK-pretraining. The base variant of our best model TEK$_{PF}$, which uses TEK-pretrained TEK-enriched representations records even bigger gains of 3.1 F1 and 2.0 F1 on Wikipedia and Web respectively over a comparable 12-layer RoBERTa++ baseline. We see similar trends for the 24-layer models with improvements of 1.6 F1 and 1.7 F1 over RoBERTa++.

MRQA Table 3 shows in-domain and out-of-domain evaluation on MRQA. As we observed for TriviaQA, 12-layer variants of our RoBERTa baselines are competitive with previous work, which includes D-Net (Li et al., 2019) and Delphi (Longpre et al., 2019), the top two systems of the MRQA shared task, while the 24-layer variants considerably outperform the current state of the art across all datasets. We also see that RoBERTa++ performs better than RoBERTa on all datasets except DROP and RACE. DROP is designed to test arithmetic reasoning, while RACE contains (often fictional and thus not groundable to Wikipedia) passages from English exams for middle and high school students in China. The drop in performance after further pretraining on Wikipedia could be a result of multiple factors including the difference in style of required reasoning or content; we leave further investigation of this phenomenon for future work. The base variants of TEK$_F$ and TEK$_{PF}$ outperform both baselines on all other datasets. Comparing the base variant of our full TEK$_{PF}$ approach to RoBERTa++ , we observe an overall improvement of 1.6 F1 with strong gains on BioASQ (4.2 F1), DuoRC (2.0 F1), and TextbookQA (2.0 F1). The 24-layer variants of TEK$_{PF}$ show similar trends with improvements of 2.1 F1 on BioASQ, 1.1 F1 on DuoRC, and 1.6 F1 on TextbookQA. Our large models see a reduction in the average gain mostly due to drop in performance on DROP. Like for TriviaQA, we observe that TEK-pretraining generally improves performance even further where TEK-finetuning is useful (with the exception of DuoRC where we see a small loss of 0.24 F1 due to TEK-pretraining for the large models [3]), with the biggest

[3] According to the Wilcoxon signed rank test of statistical significance, the large TEK$_{PF}$ is significantly better than TEK$_F$ on BioASQ and TextbookQA p-value $< .05$, and is not significantly different from it for DuoRC.

	MRQA-In	BioASQ	TextbookQA	DuoRC	RE	DROP	RACE	MRQA-Out
				Shared task				
D-Net (Ensemble)	84.82	-	-	-	-	-	-	70.42
Delphi	-	71.98	65.54	63.36	87.85	58.9	53.87	66.92
				This work				
RoBERTa (Base)	82.98	68.8	58.32	62.56	86.87	**54.88**	**49.14**	68.17
RoBERTa++ (Base)	83.22	68.36	60.51	62.40	87.93	53.11	47.90	68.38
TEK$_F$ (Base)	83.44	69.71	62.19	63.43	87.49	51.04	46.43	68.46
TEK$_{PF}$ (Base)	**83.71**	**72.58**	**62.55**	**64.43**	**88.29**	54.58	47.75	**70.01**
RoBERTa (Large)	85.75	73.41	65.95	66.79	88.82	**68.63**	**56.84**	74.02
RoBERTa++ (Large)	85.80	74.73	67.51	67.40	**89.58**	67.62	55.95	74.58
TEK$_F$ (Large)	86.23	75.37	68.17	**68.80**	89.43	67.46	55.20	74.88
TEK$_{PF}$ (Large)	**86.33**	**76.80**	**69.10**	68.54	89.15	66.24	56.14	**75.00**

Table 3: In-domain and out-of-domain performance (F1) on MRQA. RE refers to the Relation Extraction dataset. MRQA-Out refers to the averaged out-of-domain F1.

gains seen on BioASQ.

Takeaways When are TEK-enriched representations most useful for question answering? Across all evaluation benchmarks, the strongest gains are on TriviaQA, BioASQ, and TextbookQA. All three datasets involve questions targeting the long tail of factual information, which has sizable coverage in Wikipedia, the encyclopedic collection we use. We hypothesize that enriching representations with textual encyclopedic knowledge could be particularly useful when factual information that might be difficult to "memorize" during pretraining is important. Current pretraining methods have a key advantage of being able to store a significant amount of world knowledge into model parameters (Petroni et al., 2019); it might enable the model to make correct predictions even from contexts with complex phrasing or partial information. TEK-enriched representations complement this strength via dynamic retrieval of factual knowledge. Finally, improvements on the science-based datasets BioASQ and TextbookQA further suggest that Wikipedia can be used as a *bridge corpus* for more effective domain adaptation for QA.

5 Ablation Studies

In the previous section, we saw consistent gains for our two-stage approach which involves first pretraining with encyclopedic knowledge on Wikipedia and then finetuning TEK-enriched representations on RC-contexts augmented with textual knowledge. In this section, we seek to isolate the contributions of the two components: tailored TEK pretraining for the Transformer encoder, and using TEK-enriched representations for QA tasks.

Setup For these experiments, we consider two variables: *pretraining* method and *finetuning* task representation. The pretraining method variable can take one of two values – Context-O, which refers to pretraining only on document-level contexts, and TEK, which refers to pretraining on contexts augmented with textual knowledge. Likewise, the finetuning task representation variable can also take two values: Context-O, which refers to finetuning and doing inference with task representations on given RC contexts only, and TEK, which refers to finetuning and making predictions with TEK-enriched representations of RC-contexts. A combination of Context-O pretraining and Context-O finetuning is equivalent to the RoBERTa++ baseline in our main results (Tables 2 and 3). Likewise a combination of TEK pretraining and TEK finetuning refers to our full approach, TEK$_{PF}$.

Table 4 shows results on the development sets of TriviaQA and MRQA (in-domain) for all four combinations of the pretraining and finetuning method variables. For these ablation studies, we use the 12-layer base models for all experiments, and report results, averaged over 3 random seeds.

Usefulness of TEK-Enriched Representations

We first compare the performance of TEK-enriched finetuning task representations which integrate textual knowledge to represent RC-contexts to Context-O finetuning task representations, which only encode the given question and passage without knowledge, for each pretraining regime (Table 4). For models finetuned from the same Context-O checkpoint (rows 1 and 2), we see consistent improvements on TriviaQA (1.4 and 1.2 F1 for

	Pretraining	Finetuning	Wiki	Web	MRQA
1	Context-O	Context-O	72.8	81.2	83.2
2	Context-O	TEK	74.2	82.4	83.4
3	TEK	Context-O	72.9	81.6	83.3
4	TEK	TEK	**75.1**	**82.8**	**83.7**

Table 4: F1 on TriviaQA and MRQA dev sets for different combinations of pretraining and finetuning.

N_C	N_B	Wiki	Web	MRQA
384	0	72.4	80.4	83.0
512	0	72.8	81.2	83.2
384	128	**74.2**	**82.4**	**83.4**
256	256	73.6	82.2	83.3
128	384	68.1	79.5	81.7

Table 5: Performance (F1) on TriviaQA and MRQA dev sets for varying lengths of context (N_C) and background (N_B). All models were finetuned from the same RoBERTa++ pretrained checkpoint.

Wikipedia and Web respectively) and MRQA (0.2 F1) after enriching representations with encyclopedic knowledge. *This underscores the advantage of* textual *encyclopedic knowledge in that it improves current models even without additional TEK-pretraining*. The improvements from TEK-enriching task representations are even more prominent for models that have been TEK pretrained (rows 2 and 4) with gains of 2.2 F1 and 1.2 F1 on TriviaQA Wikipedia and Web respectively, and 0.4 on MRQA.

Comparing TEK Pretraining and Context-only Pretraining

Section 4 showed how TEK-pretraining further improves the model's ability to use the retrieved background knowledge to improve representations for the downstream RC task. We also compare the two pretraining setups for models which do *not* use background knowledge to form representations for the finetuning tasks (rows 1 and 3). We see marginal gains across all datasets for TEK pretraining indicating that pretraining with encyclopedic knowledge does not hurt QA performance even when such information is not available during finetuning and inference. While previous work (Liu et al., 2019b; Joshi et al., 2020) has shown that pretraining with single contiguous chunks of text clearly outperforms BERT's bi-sequence pipeline[4], our results suggest that using *background* sentences from other documents during pretraining has no adverse effect on the downstream tasks we consider.

Trade-off between Document Context and Knowledge

Our approach incorporates textual knowledge by using a part of our Transformer window for it, instead of additional context from the same document. Having established the usefulness of our background knowledge even without tailored pretraining, we now consider the trade-off between neighboring context and retrieved knowledge (Table 5). We first compare using a shorter window of 384 tokens for RC-contexts (question and passage) with using 512 tokens for RC-contexts (the first two rows). We see consistent gains from using longer document context, some of which our TEK-enriched representations need to sacrifice. We then consider the trade-off for varying values of context length N_C and background length N_B as indicated in rows 2-5. Our partitioning of 384 tokens for context and 128 for background outperforms other configurations. Moreover, adding up to 256 tokens of background knowledge improves performance over using only document context. *This suggests that relevant encyclopedic knowledge about multiple entities from outside of the current document is more useful than long-distance neighboring text from the same document for these benchmarks.*

6 Discussion

Our empirical results show that TEK-enriched representations consistently improve performance across multiple benchmarks. In this section, we analyze the source of these gains.

For 75% of the examples in the TriviaQA Wikipedia development set where our approach outperforms the context-only baselines, the answer string is mentioned in the background text. A qualitative analysis of these examples indicates that the retrieved background information typically falls into two categories – (a) where the background helps disambiguate between multiple answer candidates by providing partial pieces of information missing from the original context, and (b) where the background sentences help by providing a more direct, yet redundant, phrasing of the information need compared to the original context. Figure 3 provides examples of each category.

Note that even when the retrieved background contains the answer string, our model uses the back-

[4]BERT randomly samples the second sequence from a different document in the corpus with a probability of 0.5.

Question: Which river originates in the Taurus Mountains, and flows through Syria and Iraq?
Our Answer: Euphrates
Baseline Answer: Tigris
Context: The Southeastern Taurus mountains form the northern boundary of the Southeastern Anatolia Region and North Mesopotamia. They are also the source of the Euphrates River and Tigris River.
Background: Originating in eastern Turkey, the Euphrates flows through Syria and Iraq to join the Tigris in the Shatt al - Arab, which empties into the Persian Gulf.

Question: Who did Germany defeat to win the 1990 FIFA World Cup?
Our Answer: Argentina
Baseline Answer: Italy
Context: At the 1990 World Cup in Italy, West Germany won their third World Cup title, in its unprecedented third consecutive final appearance. Captained by Lothar Matthus, they defeated Yugoslavia (4-1), UAE (5-1), the Netherlands (2-1), Czechoslovakia (1-0), and England (1-1, 4-3 on penalty kicks) on the way to a final rematch against Argentina, played in the Italian capital of Rome.
Background: At international level, he is best known for scoring the winning goal for Germany in the 1990 FIFA World Cup Final against Argentina from an 85th - minute penalty kick.

Figure 3: Examples of questions, passages, and backgrounds from TriviaQA. The first example has background knowledge that provides information complementary to the context while the second provides a more direct, yet redundant, phrasing of the information need compared to the original context.

ground only to refine representations of the candidate answers in the original document context, as possible answer positions in the background are not considered in our model formulation. This highlights the strength of an encoder with full cross-attention between RC-contexts and background knowledge; this encoder is able to build representations for, and consider possible answers in all document passages, while integrating knowledge from multiple pieces of external textual evidence.

The exact form of background knowledge is dependent on the retrieval function. We have shown that contextualizing the input with textual background knowledge, especially after suitable pretraining, improves state of the art methods even with simple entity linking and *ngram*-match retrieval functions. We hypothesize that more sophisticated retrieval methods could further significantly improve performance (by for example, prioritizing for more complementary information).

7 Related Work

We categorize related work into three main groups: (*i*) integrating background knowledge for language understanding, (*ii*) pretraining of general purpose text representations, and (*iii*) question answering.

Integrating Background Knowledge for Natural Language Understanding (NLU)

Many NLU tasks require the use of multiple kinds of background knowledge (Fillmore, 1976; Minsky, 1986). Earlier work combined features over the given task data with hand-engineered features over knowledge repositories to define models for end tasks (Ratinov and Roth, 2009; Nakashole and Mitchell, 2015), *inter alia*. Subsequent work used single-word embeddings such as Glove (Pennington et al., 2014) and Word2vec (Mikolov et al., 2013), trained following the Distributional Hypothesis (Harris, 1954). Other forms of external knowledge include relational knowledge between word or entity pairs, typically integrated via embeddings from structured KBs (Yang and Mitchell, 2017; Bauer et al., 2018; Mihaylov and Frank, 2018; Wang and Jiang, 2019) or via word pair embeddings trained from text (Joshi et al., 2019). Weissenborn et al. (2017) used a specialized architecture to integrate background knowledge from ConceptNet and Wikipedia entity descriptions. For open-domain QA, recent works (Sun et al., 2019; Xiong et al., 2019) jointly reasoned over text and KGs, building a graph of text and KG candidate answers, and using a specialized architecture to define the flow of information between them.

These methods did not take advantage of large scale unlabeled text to pre-train deep contextualized representations like BERT (Devlin et al., 2019), which have the capacity to encode even more knowledge in their parameters. Most relevant to ours is work building upon these powerful pretrained representations, and further integrating external knowledge. Recent work focuses on refining pretrained contextualized representations using entity or triple embeddings from structured knowledge graphs (KGs) (Peters et al., 2019; Yang et al., 2019; Zhang et al., 2019). The KG embeddings are trained separately (often to predict links in the KG), and knowledge from KG is fused with deep Transformer representations via special-purpose architectures. Some of these prior works also pre-train the knowledge fusion layers from unlabeled text through self-supervised objectives (Zhang et al., 2019; Peters et al., 2019). Instead of separately encoding structured KBs, and then attending to their single-vector embeddings, we explore directly using wider-coverage textual encyclopedic background knowledge. This enables direct application of a pretrained deep Transformer (RoBERTa) for jointly contextualizing input text and background knowledge. We showed background knowledge integration can be further improved by additional knowledge-augmented self-supervised pretraining.

In concurrent work, Liu et al. (2019a) augment text with relevant triples from a structured KB. They process triples as word sequences using BERT with a special-purpose attention masking strategy. This allows the model to partially re-use BERT for encoding and integrating the structured knowledge. Our work uses wider-coverage textual sources instead and shows the power of additional knowledge-tailored self-supervised pretraining.

Pretraining Contextualized Representations

We have heavily built upon recent work in pretraining general purpose text representations (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019b; Joshi et al., 2020) which encode contiguous segments from documents. Our pretrained TEK-enriched representations encode contiguous texts jointly with dynamically retrieved textual encyclopedic knowledge from multiple documents as background. Other pretrained knowledge integration methods (Zhang et al., 2019; Peters et al., 2019) refine input text representations by integrating structured KB embeddings instead.

Question Answering

For extractive text-based question answering with deep pretrained models, prior work has integrated background knowledge in the form of structured KBs (Yang et al., 2019; Liu et al., 2019a). For open-domain QA (Chen et al., 2017), where documents known to answer the question are not given as input (e.g. OpenBookQA (Mihaylov et al., 2018)), methods exploring retrieval of relevant textual knowledge are a necessity. Recent work in these areas has focused on improving the evidence retrieval components (Lee et al., 2019; Banerjee et al., 2019; Guu et al., 2020), and has used Wikidata triples with textual descriptions of Wikipedia entities as a source of evidence (Min et al., 2019). Other approaches use pseudo-relevance feedback (PRF) (Xu and Croft, 1996) style multi-step retrieval of passages by query reformulation (Buck et al., 2018; Nogueira and Cho, 2017), entity linking (Das et al., 2019b), and more complex reader-retriever interaction (Das et al., 2019a). When multiple candidate contexts are retrieved for open-domain QA, they are sometimes jointly contextualized using a specialized architecture (Min et al., 2019). To our knowledge, we are the first to explore pretraining of representations which can integrate background from multiple source documents, and hypothesize that these representations could be further improved by more sophisticated retrieval approaches.

8 Conclusion

We presented a method to build text representations by jointly contextualizing the input with dynamically retrieved textual encyclopedic background knowledge from Wikipedia. We applied our method to reading comprehension tasks and showed consistent improvements, in-domain and out-of-domain, across multiple benchmarks that require factual reasoning and knowledge well represented in the background collection.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2019. Careful selection of knowledge to solve open book question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6120–6129, Florence, Italy. Association for Computational Linguistics.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230, Brussels, Belgium. Association for Computational Linguistics.

Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Andrea Gesmundo, Neil Houlsby, Wojciech Gajewski, and Wei Wang. 2018. Ask the right questions: Active question reformulation with reinforcement learning. In *International Conference on Learning Representations*, Vancouver, Canada.

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 (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia. Association for Computational Linguistics.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2019a. Multi-step retriever-reader interaction for scalable open-domain question answering. In *International Conference on Learning Representations*.

Rajarshi Das, Ameya Godbole, Dilip Kavarthapu, Zhiyu Gong, Abhishek Singhal, Mo Yu, Xiaoxiao Guo, Tian Gao, Hamed Zamani, Manzil Zaheer, and Andrew McCallum. 2019b. Multi-step entity-centric information retrieval for multi-hop question answering. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 113–118, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

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 arXiv:1704.05179*.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Eunsol Choi, Omer Levy, Daniel Weld, and Luke Zettlemoyer. 2019. pair2vec: Compositional word-pair embeddings for cross-sentence inference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3597–3608, Minneapolis, Minnesota. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion 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. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Patrick Lewis. 2018. *Setting the TriviaQA SoTA with Contextualized Word Embeddings and Horovo*.

Hongyu Li, Xiyuan Zhang, Yibing Liu, Yiming Zhang, Quan Wang, Xiangyang Zhou, Jing Liu, Hua Wu, and Haifeng Wang. 2019. D-NET: A pre-training and fine-tuning framework for improving the generalization of machine reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 212–219, Hong Kong, China. Association for Computational Linguistics.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019a. K-bert: Enabling language representation with knowledge graph. *arXiv preprint arXiv:1909.07606*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach.

Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, Zhucheng Tu, and Chris DuBois. 2019. An exploration of data augmentation and sampling techniques for domain-agnostic question answering. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 220–227, Hong Kong, China. Association for Computational Linguistics.

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 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832, Melbourne, Australia. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Sewon Min, Danqi Chen, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Knowledge guided text retrieval and reading for open domain question answering. *arXiv preprint arXiv:1911.03868*.

Marvin Minsky. 1986. *The Society of Mind*. Simon & Schuster, Inc., New York, NY, USA.

Ndapandula Nakashole and Tom M. Mitchell. 2015. A knowledge-intensive model for prepositional phrase attachment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 365–375, Beijing, China. Association for Computational Linguistics.

Rodrigo Nogueira and Kyunghyun Cho. 2017. Task-oriented query reformulation with reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 574–583, Copenhagen, Denmark. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Robert Logan, 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)*, Hong Kong, China. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia. Association for Computational Linguistics.

Haitian Sun, Tania Bedrax-Weiss, and William Cohen. 2019. PullNet: Open domain question answering with iterative retrieval on knowledge bases and text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2380–2390, Hong Kong, China. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2019. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Chao Wang and Hui Jiang. 2019. Explicit utilization of general knowledge in machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2263–2272, Florence, Italy. Association for Computational Linguistics.

Wei Wang, Ming Yan, and Chen Wu. 2018. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1705–1714, Melbourne, Australia. Association for Computational Linguistics.

Dirk Weissenborn, Tomáš Kočiskỳ, and Chris Dyer. 2017. Dynamic integration of background knowledge in neural nlu systems. *arXiv preprint arXiv:1706.02596*.

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

Jinxi Xu and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In *SIGIR*.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2346–2357, Florence, Italy. Association for Computational Linguistics.

Bishan Yang and Tom Mitchell. 2017. Leveraging knowledge bases in LSTMs for improving machine reading. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1436–1446, Vancouver, Canada. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

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 Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

[CLS] The River Thames known alternatively in parts as the Isis, is a river that flows through southern England including **London**. At 215 miles (346 km), it is the longest river entirely in England and the second-longest in the United Kingdom, after the River Severn. It flows through Oxford (where it is called **the Isis**), Reading, Henley-on-Thames and Windsor. [SEP] **London** : The city is split by the River Thames into North and South, with an informal central London area in its interior . [SEP] **The Isis** : The Isis" is an alternative name for the River Thames, used from its source in the Cotswolds until it is joined by the Thame at Dorchester in Oxfordshire [SEP]	[CLS] Which English rowing event is held every year on the River Thames for 5 days (Wednesday to Sunday) over the first weekend in July ? [SEP] Each year the **World Rowing Championships** is held by FISA ... Major domestic competitions take place in dominant rowing nations and include The Boat Race and **Henley Royal Regatta** in the United Kingdom , the Australian Rowing Championships in Australia , ... [SEP] **Henley Royal Regatta** : The regatta lasts for five days (Wednesday to Sunday) ending on the first weekend in July . [SEP] **World Rowing Championships** : The event then was held every four years until 1974 , when it became an annual competition ... [SEP]

Table 6: Pretraining (left) and QA finetuning (right) examples which encode contexts with background sentences from Wikipedia. The input is minimally structured by including the source page of each background sentence, and separating the sentences using special [SEP] tokens. Background is shown in blue and entities are indicated in **bold**.

Appendices

A Input Examples

Table 6 shows pretraining (left) and QA finetuning (right) examples which encode contexts with background sentences from Wikipedia.

B Finetuning Procedure

We apply the following finetuning hyperparameters to all methods, including the baselines. For each method, we chose the best model based on dev set performance.

TriviaQA We choose learning rates from {1e-5, 2e-5} and finetune for 5 epochs with a batch size of 32.

MRQA We choose learning rates from {1e-5, 2e-5} and number of epochs from {2, 3, 5} with a batch size of 32.

For both benchmarks, especially for large models, we found higher learning rates to perform sub-optimally on the development sets. Table 7 reports best performing hyperparamter configurations for each benchmark.

Dataset	Epochs	LR
12-layer Models		
MRQA	3	2e-5
TriviaQA Wiki	5	1e-5
TriviaQA Web	5	1e-5
24-layer Models		
MRQA	2	1e-5
TriviaQA Wiki	5	2e-5
TriviaQA Web	5	1e-5

Table 7: Hyperparameter configurations for TEK$_{PF}$