

SITUATEDQA: Incorporating Extra-Linguistic Contexts into QA

Michael J.Q. Zhang and Eunsol Choi

Computer Science Department

The University of Texas at Austin

{mjzhang, eunsol}@utexas.edu

Abstract

Answers to the same question may change depending on the extra-linguistic contexts (when and where the question was asked). To study this challenge, we introduce SITUATEDQA, an open-retrieval QA dataset where systems must produce the correct answer to a question given the temporal or geographical context. To construct SITUATEDQA, we first identify such questions in existing QA datasets. We find that a significant proportion of information seeking questions have context-dependent answers (e.g. roughly 16.5% of NQ-Open). For such context-dependent questions, we then crowdsource alternative contexts and their corresponding answers. Our study shows that existing models struggle with producing answers that are frequently updated or from uncommon locations. We further quantify how existing models, which are trained on data collected in the past, fail to generalize to answering questions asked in the present, even when provided with an updated evidence corpus (a roughly 15 point drop in accuracy). Our analysis suggests that open-retrieval QA benchmarks should incorporate extra-linguistic context to stay relevant globally and in the future. Our data, code, and datasheet are available at <https://situatedqa.github.io/>.

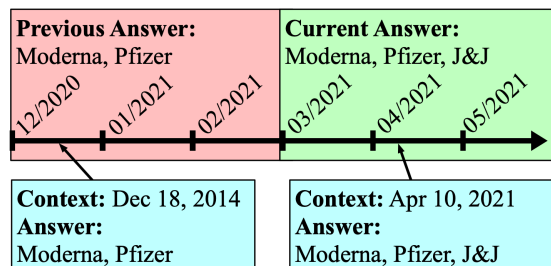
1 Introduction

Language reflects our ever-changing world; thus, the meaning of a sentence depends on a variety of extra-linguistic contexts, including *when* it was stated, *who* stated it, and *where* it was stated. Figure 1 depicts examples of this, where the answer to a question varies based on the temporal and geographical contexts. In this work, we study such questions by exploring the affects of temporal and geographical contexts in open-retrieval QA.

Ideally, QA systems should provide up-to-date, localized answers to questions like those in Figure 1 that are situated in the inquirer’s temporal

Context Type: Temporal

Question: Which COVID-19 vaccines have been authorized for adults in the US?



Context Type: Geographical

Question: Which COVID-19 vaccine was the first to be authorized by our government?

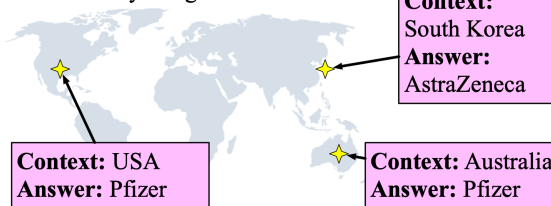


Figure 1: Examples of questions with answers that change depending on the temporal or geographical context.

and geographical contexts. The existing paradigm for evaluating QA systems, however, makes implicit assumptions about time and location and does not measure how well QA models adapt to new contexts. Recent work studying other NLP tasks have shown that this static paradigm misrepresents how models perform in practical settings (Osborne et al., 2014; Wang et al., 2008; Huang and Paul, 2018; Rijhwani and Preotiuc-Pietro, 2020). To address this challenge, we introduce SITUATEDQA, a dataset comprised of questions from existing open-retrieval QA datasets (Kwiatkowski et al., 2019) that have been re-annotated for their their context-dependence and answers across different situations (i.e., temporal or geographical contexts).¹

As we will see in Section 3, a significant pro-

¹In this work, we do not model temporal **and** geographical contexts jointly for simplicity.

Question q	Context Type c_t	Context Value c_v	Answer a
Who composed the music for the first Harry Potter film?	-	-	-
What’s the biggest country in Europe excluding Russia?	-	-	-
How many seasons are there for American Horror Story?	TEMP	Sep 18, 2019 Sep 13, 2017	10 9
Who made the most three point shots in the NBA?	TEMP	2014 2005	Ray Allen Reggie Miller
When was the last time states were created?	GEO	Nigeria United States	1 October 1996 1959
Where do we rank among the world’s largest cities?	GEO	Tokyo Shanghai	1st 3rd

Table 1: Examples of how questions interact with geographical and temporal context in SITUATEDQA. The first two questions are not identified as geographically nor temporally dependent.

portion of information-seeking questions are sensitive to the two extra-linguistic contexts we study. We annotate 9K questions from four existing datasets (Kwiatkowski et al., 2019; Berant et al., 2013; Clark et al., 2020; Campos et al., 2016) with their temporal dependence and 2K questions for their geographical dependence. Annotators find that up to 30% of questions in existing datasets have answers that change over time, and rule-based heuristics using an NER tagger identify that about 5% of questions specify a location. We collect answers from different temporal and geographic contexts for such context-dependent questions (see Table 1).

Using our collected data, we evaluate whether existing open-retrieval QA systems adapt to new temporal or geographical contexts by providing answers that are up-to-date or from new locations. While it is often assumed that retrieval based systems for QA (Karpukhin et al., 2020; Guu et al., 2020) can adapt to updated facts during inference by simply updating the retrieval corpus, we find that this is not the case. State-of-the-art retrieval based systems with updated corpora are 15 percentage points less accurate on questions whose answers have been updated versus those that have remained constant since the time when their large-scale training dataset was collected. We also observe that models fail to generalize to answering questions from new locations (Shankar et al., 2017), with accuracy dropping by 10 percentage points when asked to provide the answer from a rare location versus a common one.

To support future research developing open-retrieval QA systems which situate questions within the inquirer’s extra-linguistic contexts, we propose two tasks for modeling *what* facts change

across different extra-linguistic contexts and *how* those facts change. We also provide fine-grained evaluations for measuring how well models adapt to new temporal and geographical contexts. We establish initial performance levels on our tasks by adapting state-of-the-art methods for open-retrieval QA (Karpukhin et al., 2020; Lewis et al., 2020a; Roberts et al., 2020). Finally, we provide rich analysis of existing QA systems, which suggest that benchmark construction should incorporate extra-linguistic contexts to remain relevant globally and in the future.

2 Definitions & Tasks

2.1 Defining Extra-Linguistic Contexts

We begin by defining the scope of contexts studied in this work. For a given question q , we say that a_i is its answer when asked in the context c_i . Each context consists of a type c_{t_i} and a value c_{v_i} . We study two context types: temporal (TEMP) and geographical (GEO). TEMP defines each context value as timestamp (e.g. a date or year) where a_i is the answer to q if it was asked at the time of c_{v_i} . GEO defines each context value as a geopolitical entity where a_i is the answer to the q in the location c_{v_i} . See Table 1 for examples from each context type. We limit our study to valid questions, ignoring presupposition cases (Kim et al., 2021) (e.g., asking who is the CEO of Google before Google was founded).

2.2 Situated Question Answering

Given a question q and context c_i , the task is to produce the corresponding answer a_i for the provided context. This task requires models to situate the question within the provided extra-linguistic context to produce an appropriate answer. Models

are evaluated on exact string match with annotated answer a_i and predicted answer \hat{a}_i after minor normalization (Rajpurkar et al., 2016).

By evaluating on different sets of extra-linguistic contexts of the same question, we can measure whether models are able to generalize to new contexts. For instance, we can evaluate how models perform on commonly versus rarely asked about locations, or on questions whose answers changed recently or long ago. Furthermore, this task allows us to train systems that explicitly model how facts change across different extra-linguistic contexts.

2.3 Context-Dependent Question Identification

In context-dependent question identification, we evaluate whether models can determine whether the answer to a given question depends on its extra-linguistic context. More formally, given a question q and context type c_t , models must determine whether there exists two distinct contexts values, (c_{v_i}, c_{v_j}) , with different respective answers, $a_i \neq a_j$. We cast this as binary classification and evaluate models on their classification accuracy, F1, precision, and recall.

Identifying what facts change depending on the extra-linguistic contexts is nontrivial even for human annotators that often requires extensive background knowledge on a subject. For example, determining whether the capital of Kazakhstan has changed requires specific knowledge of the nation’s history. Identifying these questions has many practical applications, such as identifying suitable questions for QA systems that can only provide static answers, an idea we will explore later in Section 6.

3 Data Collection

Here we describe the process for identifying context-dependent questions and collecting answers from alternate contexts. We split up data collection into three stages which are depicted in Figure 2: *identification*, *{Context / Answer} collection*, and *validation*. Additional data collection details with interface screenshots can be found in Appendix A.

3.1 Identification

We source questions from a variety of datasets for open-retrieval QA: Natural Questions (NQ-Open) (Kwiatkowski et al., 2019), WebQuestions (Berant et al., 2013), TyDi-QA (Clark et al.,

2020), and MS-MARCO (Campos et al., 2016). All of these datasets are in English and contain questions that are answerable by Wikipedia documents.

Temporally dependent questions are abundant in existing datasets, as they are unambiguous during annotation and annotators simply provide the current answer at the time of collection. In contrast, geographically dependent questions are often rejected during annotation due to their ambiguity, as there is no consensus among annotators on the assumed geographical context. To correct for this scarcity, we *generate* geographically dependent questions by modifying existing questions from NQ-Open. We identify questions with phrases that specify a location by running an NER tagger (Peters et al., 2017) and remove the location entity using heuristics based on its syntactic role as identified by a dependency parser (Dozat and Manning, 2017). See Appendix A for full details.

For each question, we collect 3-way yes / maybe / no annotations for whether they are temporally or geographically dependent. We label questions as context dependent if at least two of annotators label so. We discard questions with split annotations and majority maybe annotations, and map the labels to a binary yes/no label.²

3.2 {Context / Answer} Collection

After identifying context-dependent questions, we then move on to collecting multiple {Context / Answer} pairs via crowdsourcing. For this part, we exclusively use questions from NQ-Open dataset. We allow crowdworkers to query an up-to-date version of English Wikipedia using the Google Search API.³ We outline the annotation process for each context type below:

TEMP To construct SITUATEDQA examples with temporal context, we take two steps: (1) crowd sourcing timeline for each question and (2) generating (q, c_v, a) from the annotated timeline.

Crowdworkers are asked to provide a brief timeline of answers for a given question, consisting of the current answer, the previous answer, as well as the start and end transition timestamps for each answer given as dates or years (see Figure 2). From annotated timeline and query pairs, we construct examples of (q, c_v, a) pairs in one of three ways

²These usually are highly ambiguous questions, such as “What are the measurements of a vintage oil can?”, and total discarded questions only accounts for 11% of the data.

³We use the February 20, 2021 dump of English Wikipedia.

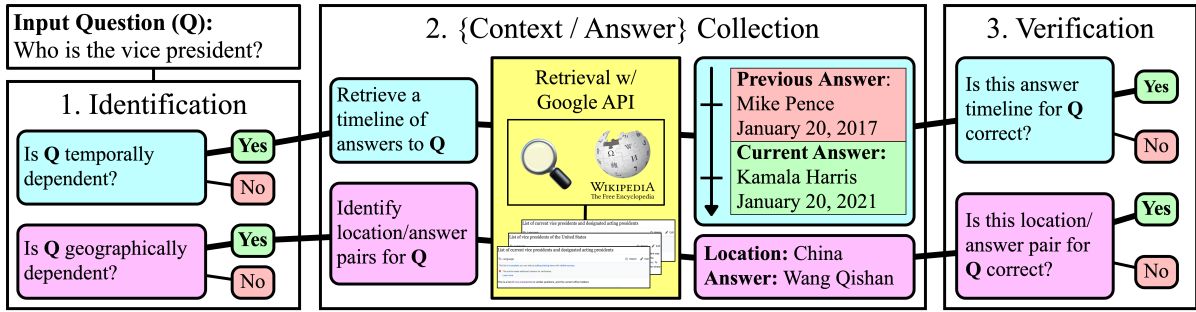


Figure 2: Data collection pipeline: Crowdworkers are first asked to identify context dependent questions. We then collect brief answer timelines for temporally dependent questions and location/answer pairs geographically dependent questions, each of which is then verified by another worker.

which we refer to as **Start**, **Sampled**, and **Static**. We describe each of these methods below:

Start examples simply use each answer’s start transition timestamp as c_v . This is intended to simulate the common scenario of asking information that is new or has recently changed. **Sampled** examples use a timestamp that lies between an answer’s start and end transition as c_v . As each answer in a timeline can result in many valid (q, c_v, a) pairs, we limit the number of values of c_v for each answer to a maximum of two. We uniformly sample up to two timestamps between each answers start and end transition, using each sampled timestamp to create a new (q, c_v, a) triple. **Static** examples utilize questions that were annotated as **not** temporally-dependent to simulate realistic settings where context dependent and independent questions coexist. For each static question, we uniformly sample a single value of c_v , resulting in one (q, c_v, a) triple per static question.

GEO We construct our evaluation datasets with crowdsourcing. For each stripped question that was annotated as geographically dependent during the identification stage, annotators were presented with the question, the original answer(s), and stripped location. Annotators first validate whether the original location and answer pair is correct for the stripped question before identifying up to two additional location and answer pairs. We then use each validated or identified pair to construct a (q, c_v, a) triple for the stripped question.

We construct our training set via distant supervision. Using our best performing context-dependent question identification model which we introduce in Section 4, we classify unlabeled examples from the NQ-Open training set. For each example that was classified as geographically-dependent, we use its original answer, location, and stripped question

to construct a (q, c_v, a) triple.

3.3 Quality Control

Validation The $\{\text{Context / Answer}\}$ annotations collected above are reviewed by another annotator in a final validation stage. Presented with a question along with all answer timelines or location/answer pairs that have been generated by other workers, annotators are asked to validate or revise each response by marking them as correct or incorrect. We collect 2-way $\{\text{Context / Answer}\}$ annotations and have one validator for each question from the development and test sets. We collect a single $\{\text{Context / Answer}\}$ annotation and skip validation for answer timelines from the training set.

Worker Qualifications / Inter-annotator agreement Table 2 reports Fleiss’s kappa for the context dependent question identification and validation stages. We find moderate to high agreement all tasks, with lower agreement on datasets with highly ambiguous queries (e.g. how much to trust friends) or questions require extensive domain knowledge (e.g., what is the cause of smog in china?). For geographically dependent question answering, questions like “when were electric trains introduced” had split votes from the annotators, as both geographically dependent and independent interpretations are valid. Overall, we observe high quality data with comparable agreement numbers with prior work, yet reaffirming prevalent ambiguity in open retrieval QA (Min et al., 2021).

3.4 Data Statistics / Analysis

Table 2 shows the statistics of our collected dataset. We annotated over 11K questions, and roughly 30-40% of them were identified as context-dependent. Temporally-dependent questions comprised of least

Context Type (c_t)	Train		Identification Data				Agreement Fleiss' κ	{Context / Answer} Data			Agreement Fleiss' κ
	# q	%	Dev # q	Dev %	Test # q	Test %		Train # q, c_v, a	Dev # q, c_v, a	Test # q, c_v, a	
TEMP	4438	36	2572	32	1962	28	0.62	6009	3423	2795	0.57
GEO	1149	46	879	42	367	37	0.56	3548	1398	505	0.56

Table 2: Dataset statistics. For our identification dataset, we report both the total number of questions ($\# q$) and the percent of questions that are context-dependent (%). For our {Context / Answer} dataset, we report the number of unique (question, context value, answer) triples ($\# q, c_v, a$). Each context type and split's (q, c_v, a) triples are collected slightly differently, see Section 3.2 for details. We also report inter-annotator agreement on context dependent question identification and {Context / Answer} validation.

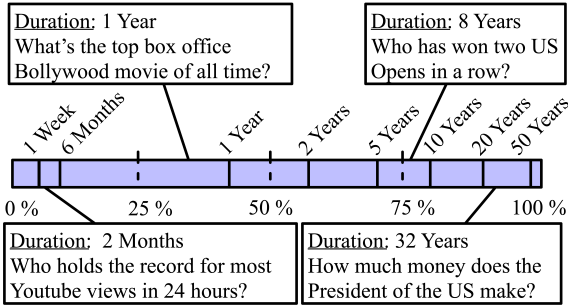


Figure 3: The distribution of temporally dependent questions by the duration its previous answer was true for with examples.

10% of examples in all datasets we looked at, even without any filtering. We also found that examples from Natural Questions where at least one answer span is within a table in its evidence document are more often temporally-dependent. To construct our final dataset, we upsample such questions. We followed the original train, development, and test split for each dataset, which caused slight inconsistencies in the proportion of examples from each dataset across splits, simulating a domain shift.

For 2.8K of those context-dependent questions (2.4K temporal, 0.5K geographical), we collected a total of 5.9K answers from alternate contexts (4.0K temporal, 1.9K geographical). From those alternate temporal context / answer pairs, we construct (q, c_v, a) by sampling valid dates, creating 6K examples. The final TEMP dataset also includes 6.7K examples from temporally-independent questions. More details and statistics on individual datasets can be found in the in Appendix A.

How often do temporally dependent facts change? We investigate how frequently answers change by measuring the distance between the the two most recent answers. Figure 3 depicts how long the previous answer was valid for. We observe a long tailed distribution, with a large proportion of answers changing around the one year mark.

4 Models

In this section, we describe our baseline models for context-dependent question identification and situated QA. Details on our learning settings and hyperparameters are in Appendix D.

4.1 Context Dependent Question Identification

As a lower bound, we provide a naive random-choice baseline, where labels are randomly assigned to match the label distribution in the training data. We also train separate BERT-based classifiers, which encode the input question q using BERT and classify based on each question's encoded [CLS] token representation. We experiment with both BERT-base and BERT-large. We approximate human performance by sampling 50 examples for temporal context and geographical context each, and re-collecting annotations from a new set of crowdworkers.

4.2 Situated Question Answering

There are two types of models for open retrieval QA: retrieval based (open book) and closed book. Retrieval based models (Karpukhin et al., 2020; Lee et al., 2019; Guu et al., 2020) answer questions by retrieving relevant passages from a corpus before producing an answer based on the retrieved passages. Closed book approaches (Roberts et al., 2020; Lewis et al., 2020a; Févry et al., 2020) do not rely on access to a corpus during inference, instead use the knowledge that has been stored in the model's parameters. These models, which typically have a large number of parameters and are trained over a single snapshot of web data, generate answers from each question.

We present two competitive baselines: one retrieval-based model, DPR (Karpukhin et al., 2020), and one closed-book model, BART-large

Context (c_t, c_v)	Augmented Question
(TEMP, 2017)	Who went number 1 in the WNBA draft as of 2017 ?
(GEO, TOKYO)	Where do we rank among the world’s largest cities in Tokyo ?

Table 3: Examples of modified queries. Text shown in bold is concatenated onto the original question. To query for the answer for some context where $c_t = \text{TEMP}$ or $c_t = \text{GEO}$, we append the phrase “as of c_v ” or “in c_v ”, replacing c_v with the context value.

(Lewis et al., 2020a).⁴ Both baselines are trained on NQ-Open (Kwiatkowski et al., 2019; Lee et al., 2019). We use the best performing model configuration from DPR paper, but swap the retrieval corpus with an up-to-date Wikipedia dump (timestamped 2021-Feb-20). In standard settings, DPR achieves an accuracy of 41.5 and BART achieves an accuracy of 24.1 on NQ-Open. We use these baselines in three settings which we introduce below: unmodified, query modified, and query modified with finetuning.

We also approximate human performance on situated QA using 100 randomly sampled examples from each context type which are annotated by the authors of this paper. While our estimated human performance appear somewhat low, they are in line with agreement rates in the original Natural Questions study (Kwiatkowski et al., 2019) and are tied to the challenges of evaluating open-retrieval QA. Discrepancies are often due to ambiguous questions or equivalent answer forms (e.g., July 5, 1945 vs. 5 July 1945, The Patriots vs. New England Patriots). The latter is especially problematic for geographically dependent questions, where answers involving time expression occur much more frequently, comprising over 50% of examples. See Appendix A for more details.

Unmodified Baseline We evaluate open-retrieval QA models on our task without any alterations. Each model is only given the question, ignoring the extra-linguistic context, and thus predicts the same answer no matter the context. For temporally-dependent questions, models are expected to output the most up-to-date answer. For geographically-dependent questions, models must assume some geographical context to produce an answer.

⁴We use the pretrained model and implementation from <https://github.com/shmsw25/bart-closed-book-qa>.

Query Modified One simple method for incorporating extra-linguistic contexts into question answering is to concatenate the context onto the question, separated by some special token. We adopt a slightly different approach that leverages what models have already learned about producing answers from specific contexts: concatenating a phrase that specifies the relevant context onto each question, transforming it into a context-independent question. The concatenation templates for each context type are described in Table 3. We find that these simple modifications mostly generate valid, fluent questions that closely resemble examples found in our models’ training data. We estimate that 10% of questions in NQ-Open contain similar augmentations, inserting or concatenating a phrase that specifies the context (See Appendix B for details).

Query Modified w/ Finetuning We also experiment with finetuning our query modified baselines. For our retrieval based approach (DPR), we finetune separate the reader and retriever models for each context type. We finetune retriever models using gold and hard-negative passages from the retrieval results of our Query Modified DPR baseline. Likewise, we also finetune separate closed book models for each context type.

5 Results

5.1 Context Dependent Question Identification

Table 4 reports our results on context dependent question identification. We find that pretrained language models perform competitively in this binary classification task, matching human agreements. Thus, it may be a useful tool for identifying context dependent questions which closed book systems are poorly suited for or identifying examples in benchmark datasets that require re-annotation.

5.2 Situated Question Answering

Table 5 shows our results on the situated QA task. All models lag behind our estimated human-level performance by a significant margin, especially for temporally dependent questions. Overall, retrieval based models outperform closed book models, as in the original QA datasets (41.5 EM vs. 24.1 EM). We find that applying our query modifications to the question works well for geographically dependent questions, but see no improvement for temporally dependent questions. Finetuning

Model	TEMP				GEO			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Random	45.3 / 44.8	31.9 / 28.3	61.7 / 64.9	42.1 / 39.4	49.7 / 48.0	42.3 / 38.1	54.2 / 62.8	47.5 / 47.4
BERT _{base}	94.5 / 93.5	94.2 / 92.7	88.5 / 83.1	91.2 / 87.6	86.0 / 79.6	79.0 / 67.4	90.8 / 87.6	84.5 / 76.2
BERT _{large}	95.7 / 93.8	94.7 / 91.2	91.7 / 85.9	93.2 / 88.5	86.1 / 80.7	80.3 / 69.9	88.6 / 84.7	84.3 / 76.6
Human*	96.0 / 94.0	88.9 / 83.3	100 / 100	94.1 / 90.9	88.0 / 84.0	76.5 / 72.7	86.7 / 88.9	81.2 / 80.0

Table 4: Results for context-dependent question identification. Each cell report results on development / test set.

	Query Mod.	Fine-tuned	TEMP				GEO		
			Static (400)	Samp. (1472)	Start (923)	Total	Comm. (265)	Rare (240)	Total
Retrieval based			44.2	16.0	14.2	19.4	9.1	2.9	6.1
	✓		28.8	15.9	18.5	18.6	27.5	22.1	25.0
	✓	✓	39.8	17.2	24.9	23.0	27.9	25.0	26.5
Closed Book			27.2	15.3	12.9	16.2	9.4	4.6	7.1
	✓		19.5	12.4	15.7	14.5	19.2	9.2	14.5
	✓	✓	26.0	16.2	18.3	18.3	21.5	11.7	16.8
Human*			-	-	-	57.0	-	-	34.0

Table 5: Results situated question answering, reporting exact match score on the test set. In addition to reporting overall EM for each context type, we also report EM for partitions of the test set. For TEMP, we partition the test set based on how the example’s context value was generated. For GEO, we split on whether the context-value is a location is common (Comm.) or uncommon (Rare), which is determined by whether the location appears at least five times in our dataset as a geographical context.

	Query Mod.	Fine-tuned	TEMP		GEO	
			One	Any	One	Any
Retrieval based			19.4	28.1	6.1	22.6
	✓		18.6	26.4	25.0	30.7
	✓	✓	23.0	29.7	26.5	32.3
Closed Book			16.2	24.0	7.1	25.1
	✓		14.5	19.9	14.5	22.6
	✓	✓	18.3	25.9	16.8	24.4

Table 6: Breakdown of errors from context dependent QA. We report EM accuracy evaluated against answers from the correct context (*One*) and answers from the union of all our annotated contexts (*Any*).

models on answers from multiple extra-linguistic contexts with modified questions improved performance across the board, especially for the TEMP context type.

Adding TEMP context to static questions distracts the model, significantly decreasing performance especially for the closed book baseline. This suggests that models are not robust to semantically equivalent edits, as observed in Ribeiro et al. (2020). Models also perform better when provided with an answer’s start date as context compared to a sampled date, which falls between the transitions between answers. This gap is especially pronounced for our retrieval based model.

# of Examples	US	India	Other	Rare
	42	20	203	240
Retrieval Based	23.8	10.0	5.9	2.9
Closed Book	23.8	10.0	6.4	4.6

Table 7: Results of our unmodified baseline run on GEO. We further break down performance common locations by separating the two most common ones, the “United States” and “India”, the “Other” common locations.

Do systems performance vary based on the frequency of geographic context? Table 5 splits results on GEO questions into those from common and rare locations, many of which unseen during training. Both methods have greater difficulty with rare locations; however, the drop in performance for closed book models is significantly worse. Closed book models perform 10 percentage points worse on questions from rare locations while retrieval based models only perform 3 points worse. These findings show that retrieval based systems generalize better to new locations and are better equipped for answering geographically dependent questions. We also see that the gap in performance between common and rare locations shrinks for both models after finetuning, suggesting that explicitly modeling geographical contexts helps with generalization.

	Query Mod.	Fine-tuned	Current 529	Previous 373	Total 902
Retrieval based			19.5	6.2	14.0
	✓		19.5	6.4	14.1
	✓	✓	22.5	11.3	17.8
Closed Book			18.5	4.0	12.5
	✓		18.5	4.8	12.9
	✓	✓	18.9	8.0	14.4
Human*			-	-	51.0

Table 8: Results on training our baselines on predicting the current and previous answers to a question. Human performance is estimated using 100 randomly sampled examples annotated by the authors of this paper.

What is the assumed context for a geographically dependent question? When presented with a geographically dependent question without the context, open retrieval QA models must assume some geographical context to produce an answer. We hypothesize that models will exhibit bias toward assuming the few geographical co-texts are most frequently asked about. We test this by reporting the results of our unmodified baseline, further breaking down the results from Table 5 by individual location. Each time the unmodified baseline correctly answers a question for some location, it did so by assuming that location as its context. We report our results in In Table 7, which show that models are heavily biased toward assuming the question was asked by someone in the United States or India.

How often do models provide answers from the wrong context? In Table 6, we report error analysis for our situated QA baselines that shows how often models are provide the answer from the specified context versus the union of all annotated contexts. We see that models often fail to incorporate extra-linguistic contexts into the question, producing the answer from another context.

6 Analysis

Can QA systems differentiate and recall previous answers to a question? To answer this, we study a new setting where we query models for the current and previous answer to a given question. Following the same steps as above for the TEMP and GEO context types, we develop a suite of retrieval-based and closed-book baselines for this new setting. Our query-modified baselines in this setting consist of prepending the word “previously” to the beginning of each question to query for the

	Fine-tuned	Knowledge Source Date	Stable 267	Updated 209
Retrieval Based		Wiki. Dec 2018	27.3	7.2
		Wiki. Feb 2021	26.6	11.5
	✓	Wiki. Feb 2021	28.1	17.2
Closed Book		Feb 2019	28.8	5.3
	✓	Feb 2019	28.1	6.7

Table 9: Results on the default open-retrieval QA setting where all questions are assumed to be about the present timestamp (Feb 2021), evaluation split by questions that require updating the world state to *after* the large scale training dataset was collected, *Updated*, and questions where answers didn’t change since the data collection, *Stable*. Our retrieval based systems use the English Wikipedia dumps from 2018-12-20 and 2021-02-20 while our closed book systems are pretrained on data from 2019-02 and earlier.

previous answer. In the existing framing of QA, all questions are looking for the current answer we do no augmentations to query for the current answer. For our finetuned baselines, we train separate reader models and closed book models to query for current and previous answers.

We report our results for these experiments in Table 8. We find that models are far better at producing the current answer versus past answers to a question. Finetuning, however, greatly increases performance on queries for previous answers.

Can QA systems trained on outdated answers adapt to the present? We investigate whether QA systems that are pretrained on outdated corpora and that are trained on large-scale QA datasets containing outdated answers can adapt to answering questions situated in the present. Lewis et al. (2020b) shows that updating the corpora of retrieval based models can be effective method for updating simple facts, such as current heads of state. We test whether this extends to a broader range of temporally-dependent facts by evaluating our baselines introduced above on their ability to predict the current answer in our dataset, which are up-to-date as of Feb 2021. We split the current answers into two categories, *stable* where answer did not change since 2018, and *updated* where answer changed after 2018. We chose 2018 as a threshold as it is when Natural Questions (Kwiatkowski et al., 2019) was collected and roughly matches the timestamp of the most recent data our closed book model (Lewis et al., 2020a) was pretrained on (Feb 2019).

Table 9 shows that, while retrieval based models

were able to update some world knowledge after swapping the retrieval corpus, significant gains on *updated* questions only come after finetuning with newer data. This suggests that simply updating the corpora that models retrieve passages from is not sufficient to keep models up-to-date.

7 Related Work

Comparison to other QA datasets While a few prior studies mention that answers to a question can change over time (Min et al., 2020), no prior work investigated the performance of existing models on temporally dependent queries and their prevalence.

Addressing ambiguous questions (Min et al., 2020) overlaps with our study, as ambiguity can arise from extra-linguistic contexts. Their study found 13% of examples were ambiguous due to temporal deixis. However, we found that context-dependence can co-occur with inherent ambiguities in the question which cannot be resolved with context. Thus, this figure underestimates the true proportion of temporally dependent questions. Furthermore, humans don't always consider context-dependent questions as ambiguous when there is only one answer given the context, which is normally assumed to be present time and location. Thus, ambiguities that arise due to lack of context should, therefore, be modeled separately from semantic ambiguities of the question.

Recent work has also studied generating temporally-dependent questions with timestamped answers from temporal knowledge-bases (Chen et al., 2021; Saxena et al., 2021). Similarly, Dhingra et al. (2021) generates temporally-dependent cloze-style prompts and present a temporally aware language model to address them. These works, however, synthetically generate examples and are therefore limited in scope and diversity. In contrast, we manually annotate temporally and geographically dependent questions.

Temporal understanding Temporal understanding in NLP has been typically studied in an intra-document setting, i.e., ordering events or finding temporal relationship between two events in the same document (Pustejovsky et al., 2003; Bethard et al., 2007; Cassidy et al., 2014; Llorens et al., 2015; Zhou et al., 2020).

Dynamic evaluation Dynamic evaluation has been studied under the language modeling (Osborne et al., 2014; Yogatama et al., 2014), topic

modeling (Wang et al., 2008) and entity linking (Rijhwani and Preotiuc-Pietro, 2020). Recent work (Lazaridou et al., 2021) studied temporal drift in the context of large scale pretrained language model, hinting reusing models from previous snapshot can cause performance decay, as we observe in our study as well.

Adversarial, phased data collection has been proposed (Paperno et al., 2016; Zellers et al., 2018; Potts et al., 2020) to drive model development, constantly feeding models examples that current models are unable to address. We suggest collecting questions dynamically for slightly different goal, to accurately reflect the changing world by identifying temporally dependent facts from such benchmarks and continuously updating them. Building a benchmark based on a fixed snapshot (Petroni et al., 2021) can be a viable alternative.

Researchers studied keeping knowledge sources up-to-date. Konovalov et al. (2017) looked at the constantly changing facts from the perspective of automatically extracting knowledgebase revisions. Schuster et al. (2021) presents contrastive sentences based on Wikipedia revisions, showing how entailment decision can change over time.

8 Conclusion & Future Work

We present the first study of how extra-linguistic contexts affect open retrieval QA. Our study reveals that current systems fail to adapt to shifts in the temporal or geographical context. We, therefore, propose tasks and create a dataset for training and evaluating QA systems on modeling how facts change across contexts. Our dataset will support ample future work for developing models which can gracefully update its predictions based on new temporal and geographical contexts. Future research may address incorporating in temporally and geographically dependent source documents, such as news articles, or considering other extra-linguistic contexts such as *who* is asking the question, taking individual's preferences into account.

Acknowledgement

We would like to thank Sewon Min, Raymond Mooney, and members of UT NLP group for comments and discussions. The work is partially funded by Google Faculty Awards.

References

- Jonathan Berant, Andrew K. Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Steven Bethard, James H. Martin, and Sara Klingenstein. 2007. Timelines from text: Identification of syntactic temporal relations. *International Conference on Semantic Computing (ICSC)*, pages 11–18.
- Daniel Fernando Campos, T. Nguyen, M. Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, L. Deng, and Bhaskar Mitra. 2016. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, abs/1611.09268.
- T. Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. *ArXiv*, abs/2108.06314.
- Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac : Question answering in context. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, abs/1808.07036.
- J. Clark, Eunsol Choi, M. Collins, Dan Garrette, T. Kwiatkowski, V. Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics (TACL)*, 8:454–470.
- J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.
- Bhuvan Dhingra, Jeremy R. Cole, Julian Martin Eisenchlos, D. Gillick, Jacob Eisenstein, and William W. Cohen. 2021. Time-aware language models as temporal knowledge bases. *ArXiv*, abs/2106.15110.
- Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. *ArXiv*, abs/1611.01734.
- 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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.
- Jacob Eisenstein, Brendan T. O’Connor, Noah A. Smith, and E. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and T. Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pappas, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. *Proceedings of the International Conference on Machine Learning (ICML)*.
- Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61.
- Xiaolei Huang and Michael J. Paul. 2018. Examining temporality in document classification. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *ArXiv*, abs/2010.05953.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Nora Kassner and H. Schutze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and D. Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

- Meeting of the Association for Computational Linguistics (ACL)*.
- A. Konovalov, Benjamin Strauss, Alan Ritter, and Brendan T. O'Connor. 2017. Learning to extract events from knowledge base revisions. *Proceedings of the 26th International Conference on World Wide Web*.
- T. Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, C. Alberti, D. Epstein, Illia Polosukhin, J. 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 (TACL)*, 7:453–466.
- A. Lazaridou, Adhiguna Kuncoro, E. Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Sebastian Ruder, Dani Yogatama, Kris Cao, Tomás Kociský, Susannah Young, and P. Blunsom. 2021. Pitfalls of static language modelling. *ArXiv*, abs/2102.01951.
- Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880, Online. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandara Piktus, F. Petroni, V. Karpukhin, Naman Goyal, Heinrich Kuttler, M. Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.
- Xin Li and D. Roth. 2002. Learning question classifiers. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.
- Xiao Ling and Daniel S. Weld. 2010. Temporal information extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Hector Llorens, Nathanael Chambers, N. UzZaman, N. Mostafazadeh, James F. Allen, and J. Pustejovsky. 2015. Semeval-2015 task 5: Qa tempeval - evaluating temporal information understanding with question answering. In *SemEval@NAACL-HLT*.
- Sewon Min, Jordan L. Boyd-Graber, C. Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, Colin Raffel, Adam Roberts, T. Kwiatkowski, Patrick Lewis, Yuxiang Wu, Heinrich Kuttler, L. Liu, Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel, Sohee Yang, Minjoon Seo, Gautier Izacard, F. Petroni, Lucas Hosseini, Nicola De Cao, E. Grave, Ikuya Yamada, Sonse Shimaoka, Masatoshi Suzuki, Shumpei Miyawaki, S. Sato, Ryo Takahashi, J. Suzuki, Martin Fajcik, Martin Docekal, Karel Ondrej, P. Smrz, Hao Cheng, Y. Shen, X. Liu, Pengcheng He, W. Chen, Jianfeng Gao, Barlas Oğuz, Xilun Chen, V. Karpukhin, Stanislav Peshterliev, Dmytro Okhonko, M. Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Wen tau Yih. 2021. Neurips 2020 efficientqa competition: Systems, analyses and lessons learned. *ArXiv*, abs/2101.00133.
- Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Miles Osborne, Ashwin Lall, and Benjamin Van Durme. 2014. Exponential reservoir sampling for streaming language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and R. Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks.
- Fabio Petroni, Tim Rocktäschel, A. H. Miller, P. Lewis, A. Bakhtin, Y. Wu, and S. Riedel. 2019. Language models as knowledge bases? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Karl Pichotta and R. Mooney. 2014. Statistical script learning with multi-argument events. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

- Christopher Potts, Zhengxuan Wu, A. Geiger, and Douwe Kiela. 2020. Dynasent: A dynamic benchmark for sentiment analysis. *ArXiv*, abs/2012.15349.
- J. Pustejovsky, J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, G. Katz, and Dragomir R. Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don’t know: Unanswerable questions for squad. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Shruti Rijhwani and Daniel Preotiu-Pietro. 2020. Temporally-informed analysis of named entity recognition. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Apoorv Saxena, Soumen Chakrabarti, and Partha P. Talukdar. 2021. Question answering over temporal knowledge graphs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Tal Schuster, A. Fisch, and R. Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, abs/2103.08541.
- S. Shankar, Yoni Halpern, Eric Breck, J. Atwood, J. Wilson, and D. Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv: Machine Learning*.
- Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hananeh Hajishirzi, and Jonathan Berant. 2021. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations (ICLR)*.
- Siddharth Vashishtha, Benjamin Van Durme, and A. White. 2019. Fine-grained temporal relation extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- C. Wang, D. Blei, and D. Heckerman. 2008. Continuous time dynamic topic models. In *UAI*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Dani Yogatama, C. Wang, Bryan R. Routledge, Noah A. Smith, and E. Xing. 2014. Dynamic language models for streaming text. *Transactions of the Association for Computational Linguistics (TACL)*, 2:181–192.
- Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Ben Zhou, Qiang Ning, Daniel Khashabi, and D. Roth. 2020. Temporal common sense acquisition with minimal supervision. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, abs/2005.04304.
- Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and D. Roth. 2021. Temporal reasoning on implicit events from distant supervision. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Appendix

A Data Collection Details

Data Composition For NaturalQuestions and WebQuestions, we use the open-domain splits established in (Lee et al., 2019) and we only consider questions from TyDi-QA that are written in English. We randomly sampled the questions from all datasets except for NQ, where the ratio of temporally dependent questions was low (around 15%). Initial pilot tests showed that questions where the answer span is within a table in its evidence document (we refer these as NQ-Table) are more often temporally dependent compared to questions where answers are found only in paragraphs (we refer these as NQ-Passage). To better target temporally dependent questions, we sampled heavily from NQ-Table. The source dataset statistics can be found in Table 10, and proportion of temporally dependent questions in each dataset can be found in Table 11.

Geographically Dependent Question Generation To generate geographically dependent questions, we first run the named entity recognition tagger from Peters et al. (2017) and the dependency parser from Dozat and Manning (2017) over each example and filter for those with a geopolitical entity in the question. We then delete the entity based on its syntactic role. If the entity’s syntactic role is either *nn* or *amod*, we delete the entire entity and all of its descendants. If the the entity’s role is *pobj*, the entity’s parent preposition and all its descendants. Finally, if the entity’s role is *root* or *nsubj*, we replace the entity with the pronoun *we*, deleting all determiners and conjugating any auxiliary verbs accordingly. We ignore instances where the dependency is not in one of these categories, there are multiple GPE entities, the stripped questions is has 3 tokens or less, or there is disagreement between our parser and tagger. We use the implementations of Peters et al. (2017) and Dozat and Manning (2017) from AllenNLP (Gardner et al., 2017).

Crowdsourcing Details We pay workers 0.15 USD per identification HIT and 0.40 USD per {Context / Answer} collection and validation HIT. During {Context / Answer} collection, if an annotator is unable to find the previous answer or either transition date after visiting 3 articles, they may leave out that information. In validation stage, in

Source	Train	Dev	Test	Total
NQ-Table	2911	500	839	2647
NQ-Passage	408	500	500	1500
WebQuestions	367	361	500	861
TyDi-QA (en)	412	413	N/A	1000
MS-MARCO	340	314	348	1500
Total	4438	2572	1962	8972

Table 10: Number of identification examples by split and source dataset.

Source	Temporally Dependent		
	No	Maybe	Yes
NQ-Passage	73.3	15.6	11.1
NQ-Table	54.6	3.7	41.6
WebQuestions	59.9	10.4	29.9
TyDi-QA (en)	62.3	14.1	23.5
MS-Marco	60.0	26.4	13.6

Table 11: Temporal volatility distribution of raw annotation for various datasets (numbers on the training split).

addition to allowing annotators to search over English Wikipedia, we also provide a list of articles that were used by workers in the prior stage. In open retrieval setting, the same question may have multiple interpretations, and therefore we ask annotators to mark answers as correct if they consistent with one plausible interpretation. Figure 4, 5, 6, 7, 8 shows our annotation interface.

Generating TEMP (q, c, a) examples We sample values of c_v to match the granularity of the annotated dates (year-month-day or year) from our collected answer timelines. For each static question, we uniformly sample a single value of c_v between the original dataset’s creation (2018) and the present date (2021-03-01), matching the granularity

Source	Y/N Data	Y/N/M Data
TEMP_ABS / TEMP_REL	0.64	0.62
- NQ-Passage	0.54	0.61
- NQ-Table	0.69	0.68
- WebQuestions	0.46	0.49
- TyDi-QA (en)	0.64	0.58
- MS-Marco	0.43	0.45
GEO	0.56	0.56

Table 12: Interannotator Agreement of Identification examples. We compute Fleiss’s kappa all examples where the majority of labels (2 out of 3) are either “Yes” or “No” (first column) and over all examples (second column) “Yes”, “No”, or “Maybe/Unsure”.

	Query Mod.	Fine-tuned	TEMP					GEO			
			Static	Samp.	Start	Total	Any	Comm.	Rare	Total	Any
Retrieval (DPR)			77.2	70.2	67.9	70.4	80.1	53.8	45.8	50.0	85.0
	✓		71.2	72.9	72.2	72.4	79.8	76.7	79.6	78.1	87.9
	✓	✓	76.5	74.7	72.9	74.3	81.6	76.7	79.6	78.1	87.9

Table 13: Retrieval performance on situated question answering, measured by answer recall within the top 50 retrieved documents. We also report performance for partitions of the test set, as well as recall measured against the union of all answers from all annotated contexts (*Any*).

	Fine-tuned	Knowledge Source Date	Stable 267	Updated 209
Retrieval (DPR)		Wiki. Dec 2018	74.9	35.4
		Wiki. Feb 2021	78.3	60.3
	✓	Wiki. Feb 2021	79.4	66.0

Table 14: Retriever results on the default open-retrieval QA setting where all questions are assumed to be about the present timestamp (Feb 2021), given as answer recall at top 50 on our test set. We split examples based on whether the current answer started to be true *after* the large scale training dataset was collected, *Updated*, or whether questions where answers didn't change since the data collection, *Stable*. Retrieval is run over English Wikipedia dumps from 2018-12-20 and 2021-02-20.

distribution we find in our training set annotations.

Inter Annotator Agreement Inter annotator agreement during identification phase can be found in Table 12. At the validation phase, 70% of temporal question-context-answer pairs were annotated as correct and 85% of geographical question-context-answer pairs were annotated as correct, similar to validation phase (76%) in AmigQA (Min et al., 2020).

Agreement with Natural Questions The answer span exact match between our collected answers and the original answers from NQ where the annotated start and end transition dates overlap with NQ's creation (2018) was around 40%, similar to the average agreement rate on NQ-Open test data is 49.2% in the original study (Kwiatkowski et al., 2019). Our manual analysis on randomly sampled 50 errors and find that only 12% of the errors can be attributed to annotation errors. 70% of other errors are due to ambiguities in the question resulting in multiple possible answers (e.g. "Three movies made from Agatha Christie's novels"), or the same answer being given in different forms (e.g. "Chamberlain, Wilt" vs. "Wilt Chamberlain"). The remaining errors are the result of inconsistencies in

NQ annotations (18%).

B Data Analysis

Naturally occurring query modifications In open retrieval QA datasets, questions often ask about the answer in some specific temporal or geographical context. For instance, when people want to ask about the previous answer to a questions, people often add words like "previously" or "last" in their question to specify it. We find that such questions comprise about 4.1% of NQ-Open (Kwiatkowski et al., 2019; Lee et al., 2019). We also find that 5.1% of questions specify a specific point in time, as determined by whether there is a year expression in the question. Finally, we estimate the number of questions that specify a geographical context by counting number of stripped questions that were specified as geographically dependent. These questions, which have a phrase specifying the geographical context, comprise about 4.4% of NQ-Open. Our modifications closely resemble these naturally occurring questions, but can sometimes produce ungrammatical sentences, usually due verb tense agreement.

C Retriever Analysis

Are errors in retrieval based systems due to poor retrieval? In Table 13, we report the retrieval performance from the settings in Section 5. We measure comparing against the results from Table 5 and Table 6, we see similar trends in end-to-end performance reflected in our retriever performance.

We also explore whether the retriever model is able to adapt to updated corpora, reporting retrieval performance in Table 14. Here, we see that while retriever performance also suffers on queries with updated answers, suggesting that retriever systems are perhaps implicitly learning to situate questions within the time period of their large-scale training datasets.

D Implementation Details

Context Dependent Question Identification Baselines We use the `pytorch-transformers` (Wolf et al., 2020) library to implement our classification models. The training batch size is set to 8 and 64 for BERT-base and BERT-large, respectively. We train for 10 epochs using a learning rate of $5e-5$ and 500 warmup steps and select the best performing model measured on dev after each epoch.

Context Dependent Question Answering Baselines We finetune our closed book baselines for 10 epochs with a batch size of 256, using the AdamW optimizer with a learning rate of $1e-5$. We select the best performing model evaluated after each training epoch. We keep all other hyperparameters the same from the original implementation.

For our retrieval based baselines, we finetune both the retriever and reader components. The retriever models are trained using in-batch negatives plus one hard-negative passage per question. We use hard-negative and gold passages from the query modified model’s predictions before finetuning. We train the retriever model for 8 epochs, using a learning rate of $1e-5$, 100 warmup steps, and a batch size of 16. We then finetune our reader models for 16 epochs and a batch size of 16. We select the best model evaluated after each training epoch for both reader and retriever models, and select the best top- k retrieval results to use for in $k \in \{10, 20, 50\}$ evaluated on the development set. All other hyperparameters from the DPR reader and retriever models are kept the same from the original work.

Instructions:

- Help us make search systems (i.e. Google) smarter by teaching them to identify questions whose answers may change over time!
- In this task, you will be have to **identify questions whose answers changed over time**.
- Read through these instructions and examples to get started!

Steps:

- For each question on the right, **identify whether or not the answer has changed in the past**:
 - Yes:** The answer to this question **has changed** in the past.
 - No:** The answer to this question **has not changed**.
 - Maybe:** You should select this option if **you understand the question, but cannot determine if its answer has changed**.
 - You may also select this option if **the answer to this question changes based on opinion** (see Examples below)
 - Before selecting this option, you should perform a **brief Google search** to see if the answer has changed. If you still can't tell, then select this option.
 - Malformed Question:** Respond with this option if you **cannot make sense of the question**.
- Submit** the HIT by clicking the button at the bottom form to the right after you have responded for each question.

Before you begin, read through these examples

Example 1

Question: What city is the next winter olympics in? **Explanation:** The olympics are held in a different city every 4 years. Therefore, the answer changes ever four years.

Yes No Maybe Malformed Question

Question: Where are the adirondack mountains on a map
 Yes No Maybe Malformed Question

Question: Who won third place in fifa world cup 2018
 Yes No Maybe Malformed Question

Question: When was the last time india won olympic gold in hockey
 Yes No Maybe Malformed Question

Question: What is the population of santa cruz bolivia
 Yes No Maybe Malformed Question

Question: What was the first episode of the waltons called
 Yes No Maybe Malformed Question

Figure 4: Interface for identifying temporally and geographically dependent questions.

Search

Instructions:

- Help us make search systems (i.e. Google) smarter by teaching them to properly answer questions whose answer changes based on the asker's location.
- For example, the question "Who is the president?" has a different answer if the asker is in the US or if they're in Brazil since both countries have presidents.
- You will be presented with a question. Your task is to identify **locations** where the **answer** to the question is different, and what the answers are.
- Read through these steps and examples below to get started!

Steps:

- Read the question on the right.**
- Verify the location + answer(s) pair** by marking it as correct or incorrect
 - Some questions will have multiple answers, mark yes/no based on whether **all answers** are correct.
 - Use the search bar above** to find an article.
 - Click the **New Search** button to enter your search.
 - We will then show you Google's top suggested Wikipedia articles for your query.
 - If you response **yes**, then provide the title of the article where you found the answer.
- Provide 2 additional location + answer pairs.**
 - All locations and answers should be unique. This includes the location and answer we provide if you marked it as correct.
 - For each location/answer pair, you'll also be asked to **provide the title of the wikipedia article** where you found the answer.
 - If you cannot find another location + answer pair, mark it by selecting the checkbox and move on. **This should be a rare occurrence.**
- Submit** the HIT by clicking the button at the bottom form to the right.

Example 1:

Question: *When did we gain independence?*

Is This Location + Answer Pair Correct?

- Location:** the united arab emirates
- Answer(s):**
 - 1966

Yes No

- Copy the **title of the article** where you found this answer

Article Title from Provided Answer

New Location #1

There is no other location that has a different answer to the question.

- Enter a possible **location #1** here

Location #1

- Copy the **answer from location #1** here

Answer from Location #1

- Copy the **title of the article** where you found this answer

Article Title from Answer #1

Figure 5: Interface for collecting geographical {Context / Answer} pairs. Annotators are also asked to verify the original location and answer.

mlb record for most homeruns

Search Results:

Showing Results For: mlb record for most homeruns

[List of Major League Baseball career home run leaders - Wikipedia](#)
 Barry Bonds holds the **record for most career home runs**, hitting 762 over his 22-year career. This is a list of the top 300 **Major League Baseball** leaders in **home** ...

[List of Major League Baseball annual home run leaders - Wikipedia](#)
 Barry Bonds, who also has the **most career home runs**, set the current single season **record** of 73 in 2001. The 1998 and 2001 seasons each had 4 players hit ...

[List of Major League Baseball single-game home run leaders ...](#)
 Writers of Sporting News described hitting four **home runs** in a single **Major League Baseball (MLB)** game as "baseball's greatest single-game accomplishment". Eighteen players have accomplished the feat to date, the **most** recent being ... No player has ever hit four **home runs** in a postseason game; that **record** is three, first ...

[List of Major League Baseball home run records - Wikipedia](#)
Most seasons with 40 home runs[edit]. Player, Seasons, Seasons and teams.
 Babe Ruth, 11, 1920–21, ...

[Barry Bonds - Wikipedia](#)
 Barry Lamar Bonds (born July 24, 1964) is an American former professional

Question: *Who holds the mlb record for most homeruns*

Select an article.

The answer to this question is a list that changes over time.

Copy the **current** answer to this question from the selected article

Current Answer...

Copy when the **current** answer **started** to apply from the selected article

Current Answer Start Date...

Copy the **previous** answer to this question from the selected article

Previous Answer...

Copy when the **previous** answer **started** to apply from the selected article

Previous Answer Start Date...

Figure 6: Interface for searching for articles during temporal {Context / Answer} collection. The search interface is shared between temporal and geographical {Context / Answer} collection and verification stages.

The screenshot shows a Wikipedia article titled "List of Major League Baseball career home run leaders". The article text includes information about Barry Bonds holding the record with 762 home runs. A sidebar on the right contains a "Question:" section with the text "Who holds the mlb record for most homeruns". Below the question, there are several input fields for "Current Answer...", "Current Answer Start Date...", "Previous Answer...", and "Previous Answer Start Date...", each with a "Missing" button. A "Submit HIT" button is at the bottom of the sidebar.

Figure 7: Interface for viewing Wikipedia articles during the {Context / Answer} collection and verification stages. Workers select articles from the search results depicted above that they want to view.

The screenshot displays an "Instructions:" section with a list of guidelines for workers, such as "Help us make search systems (i.e. Google) smarter by teaching them to identify past and present answers to your questions!". Below the instructions are three examples: "Example 1, Response 1:", "Example 1, Response 2:", and "Example 1, Explanation:". To the right, a sidebar contains a "Question:" section with the text "where was the final game of the world series played". It lists "Response 1:" with "Current answer: Arlington, TX" and "Date when the current answer started to be true: 2020". Below this is a "Is this response correct?" section with "Yes" and "No" buttons. A second "Question:" section follows with "Response 2:" and "Current answer: Globe Life Field", and another "Is this response correct?" section with "Yes" and "No" buttons.

Figure 8: Annotation interface for {Context / Answer} verification stages. Workers may search for the answers themselves, or view the articles used by workers in the prior stage.