FORECASTQA: A Question Answering Challenge for Event Forecasting with Temporal Text Data

Woojeong Jin[1] **Rahul Khanna**[1] **Suji Kim**[1] **Dong-Ho Lee**[1]
Fred Morstatter[2] **Aram Galstyan**[2] **Xiang Ren**[1 2]
[1]Department of Computer Science, University of Southern California
[2]Information Sciences Institute, University of Southern California
{woojeong.jin, rahulkha, sujikim, donghole, xiangren}@usc.edu, {fredmors, galstyan}@isi.edu

Abstract

Event forecasting is a challenging, yet important task, as humans seek to constantly plan for the future. Existing automated forecasting studies rely mostly on *structured data*, such as time-series or event-based knowledge graphs, to help predict future events. In this work, we aim to formulate a task, construct a dataset, and provide benchmarks for developing methods for event forecasting with large volumes of *unstructured* text data. To simulate the forecasting scenario on temporal news documents, we formulate the problem as a restricted-domain, multiple-choice, question-answering (QA) task. Unlike existing QA tasks, our task limits accessible information, and thus a model has to make a forecasting judgement. To showcase the usefulness of this task formulation, we introduce FORECASTQA, a question-answering dataset consisting of 10,392 event forecasting questions, which have been collected and verified via crowdsourcing efforts. We present our experiments on FORECASTQA using BERT-based models and find that our best model achieves 61.0% accuracy on the dataset, which still lags behind human performance by about 19%. We hope FORECASTQA will support future research efforts in bridging this gap.[1]

1 Introduction

Forecasting globally significant events, such as outcomes of policy decisions, civil unrest, or the economic ramifications of global pandemics, is a consequential but arduous problem. In recent years there have been significant advances in applying machine learning (*e.g.*, time-series prediction methods) to generate forecasts for various types of events including conflict zones (Schutte, 2017), duration of insurgency (Pilster and Böhmelt, 2014), civil unrest (Ramakrishnan et al., 2014a) and terrorist events (Raghavan et al., 2013).

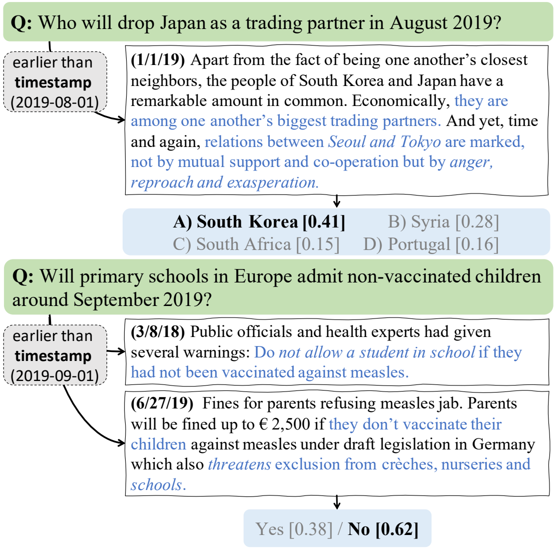

Q: Who will drop Japan as a trading partner in August 2019?
earlier than timestamp (2019-08-01)
(1/1/19) Apart from the fact of being one another's closest neighbors, the people of South Korea and Japan have a remarkable amount in common. Economically, they are among one another's biggest trading partners. And yet, time and again, relations between Seoul and Tokyo are marked, not by mutual support and co-operation but by anger, reproach and exasperation.
A) South Korea [0.41]
B) Syria [0.28]
C) South Africa [0.15]
D) Portugal [0.16]
Q: Will primary schools in Europe admit non-vaccinated children around September 2019?
earlier than timestamp (2019-09-01)
(3/8/18) Public officials and health experts had given several warnings: Do not allow a student in school if they had not been vaccinated against measles.
(6/27/19) Fines for parents refusing measles jab. Parents will be fined up to € 2,500 if they don't vaccinate their children against measles under draft legislation in Germany which also threatens exclusion from crèches, nurseries and schools.
Yes [0.38] / No [0.62]

Figure 1: **Examples from the FORECASTQA dataset.** Models only have access to articles published prior to the *timestamp* associated with each question. Models assign probabilities to each answer choice; bold denotes the correct answer for each question.

Current automated forecasting methods perform well on problems for which there are sufficient *structured* data (*e.g.*, knowledge graphs), but are not well suited for events for which such data may not exist. Humans, though, can often accurately forecast outcomes by leveraging their judgement, domain knowledge, and prior experience (Tetlock and Gardner, 2016), along with the vast amounts of *unstructured* text data available to us (*e.g.*, news articles). We are able to identify and retrieve salient facts from the near-endless pool of unstructured information, synthesize those facts into coherent beliefs, and generate probabilistic forecasts. Unfortunately, the process does not scale well in terms of the amount of information that must be processed and the number of events one has to forecast.

Here we address the above problem by formalizing a forecasting task, creating a dataset, and providing benchmarks to develop methods for the

[1] https://inklab.usc.edu/ForecastQA/

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 4636–4650
August 1–6, 2021. ©2021 Association for Computational Linguistics

task. Specifically, we formulate the forecasting problem as a multiple-choice Question Answering (QA) task, where the input is a news corpus, questions, choices and timestamps associated with each question, and the output is one of the given choices per question. Our approach is rooted in the observation that both forecasting and QA follow a similar process: digesting massive amounts of textual data, identifying supporting pieces of evidence from text, and chaining different pieces to generate answers/forecasts.

Forecast Question Answering (ForecastQA) introduces a novel *timestamp constraint* per question that prohibits the model from accessing new articles published after the *timestamp*. By doing so, ForecastQA simulates a forecasting scenario; each question's timestamp is chosen to ensure that the question is about the outcome of a future event.

To illustrate this, consider the question, "*Will primary schools in Europe admit non-vaccinated children around September 2019?*" in Figure 1, and the fact that models only have access to articles before "2019-09-01." With the addition of this timestamp constraint, our query becomes a question about a future event in "September, 2019" based on articles from the "past"; the model is now being tested for its *forecasting ability*[2]. To answer the question, the model must find pertinent events from "past" information, resolve the temporal and causal relations between them, and finally make a *forecasting judgement* based on its interpretation of past information to answer the question. Our task differs from that of other works that require an understanding of temporal relationships (Ning et al., 2020) and temporal commonsense reasoning (Zhou et al., 2019), as our task forces a model to make a forecasting judgement.

In support of the proposed ForecastQA formulation, we construct a dataset of 10,392 yes-no and multiple-choice questions. This data is collected via crowdsourcing based on news articles, where workers are shown articles and asked to come up with yes-no and multiple-choice questions. We also crowdsourced appropriate timestamps for each question. Finally, we design a method based on pre-trained language models to deal with retrieved articles for our task. In our experiments, the methods using retrieved articles slightly outperform closed-book models, suggesting that our task is still challenging in that finding relevant information for forecasting and making a judgement are not straightforward. Our best attempt achieves 61.0% accuracy on our dataset, a significant performance gap from human performance by 19.3%.

Q: Who will drop Japan as a trading partner in August 2019?
Choices: South Korea (***answer***), South Africa, Syria, Portugal.

Article: *Why Japan and South Korea just can't get along.* (1/1/19) Apart from the fact of being one another's closest neighbours, the people of South Korea and Japan have a remarkable amount in common. Economically, they are among one another's biggest trading partners. And yet, time and again, relations between Seoul and Tokyo are marked, not by mutual support and co-operation but by anger, reproach and exasperation.

Reasoning Process: Seoul is in South Korea, Tokyo is in Japan (**commonsense - world knowledge**). Seoul and Tokyo are big trading partners (**language understanding - lexical variations**). The relations between Seoul and Tokyo are marked by anger, reproach and exasperation and these relations might cause trading relations to cease (**forecasting skills - causal relation** - *we can infer the answer from this part*).

Table 1: **Chain of reasoning.** The question requires the reasoning process to answer.

2 Related Work

Event Forecasting. There are several types of approaches exist to do event forecasting. One approach could learn from highly structured event-coded data such as ICEWS (Boschee et al., 2015) and GDELT (Leetaru and Schrodt, 2013). When these datasets are used for forecasting, they are often represented as a time series (Morstatter et al., 2019; Ramakrishnan et al., 2014b), in which each data point is associated with a timestamp. Another approach is script-learning, in which a model is provided with a chain of events and a subsequent event and is asked to predict the relation between the chain and the "future" event (Hu et al., 2017; Li et al., 2018; Lv et al., 2019). They require to convert text data into event triples and translate the questions and answer choices into their format, which limits the expressiveness of natural text. However, unlike these datasets and approaches, ForecastQA does not provide any structured data to a model. The model must learn how to extract, keep track of, and link pertinent events from unstructured text to solve forecasting questions.

QA and Temporal Reasoning on Text. There are several approaches for QA using unstructured text. Extractive QA approaches rely on finding answer spans from the text that best answer a question (Rajpurkar et al., 2016, 2018; Yang et al., 2018; Kwiatkowski et al., 2019; Huang et al., 2019).

[2] The ability to predict the outcome of future events based on unstructured text describing past events, without access to an extracted sequence of historical event triples, nor provided a fixed set of possible relations between events; as is the case with human forecasters.

Multiple-Choice QA requires a model to pick the best answer from a set (Talmor et al., 2019; Sap et al., 2019; Zhou et al., 2019), and generative QA prompts the machine to produce its own answer (Khashabi et al., 2020). Our dataset is a type of multiple-choice QA, but it differentiates itself from other QA datasets (all formats) in that the required answer does not exist in the provided text, nor is sufficient evidence provided to be able to answer a question with 100% certainty; a forecast is required. We could convert our questions into alternative query formats such as a text-to-text format, but instead we stick to multiple-choice questions as humans often weigh the benefits of multiple choices when making a forecasting judgement.

QA datasets often exist to test certain types of reasoning. One pertinent example of a reasoning type that QA tasks test is the understanding of temporal and casual relations (Jia et al., 2018a,b; Sun et al., 2018; Ning et al., 2020). However, FORECASTQA requires more than just extraction and understanding of relations; a model must be able to extract and understand the relations present in the text with the goal of making a forecasting judgement about an event whose outcome is *not* found in the text. Another type of reasoning tested in QA tasks is commonsense reasoning (Talmor et al., 2019) and even temporal commonsense reasoning (Zhou et al., 2019). While questions in FORECASTQA often require commonsense to correctly answer, not all do; event outcomes do not always follow common sense. Furthermore, our questions test forecasting abilities, which often includes various types of reasoning in addition to commonsense.

3 The FORECASTQA Task

FORECASTQA is a question answering task whose goal is to test a machine's *forecasting ability*. We consider *forecasting* as the process of anticipating the outcome of future events based on past and present data (Tetlock and Gardner, 2016). We focus on forecasting *outcomes of news-based events* coming from topics such as politics, sports, economics, etc. Training a machine to make forecasting decisions is inherently difficult, as the ground-truth label of event outcome (*e.g.*, whether an event will occur) — so often required for model training — is only obtainable "in the future". To make progress in our goal, we devise a way to *simulate the forecasting scenario* by introducing a novel *time constraint*, allowing us to validate the machine predictions by obtaining desired ground-truth labels.

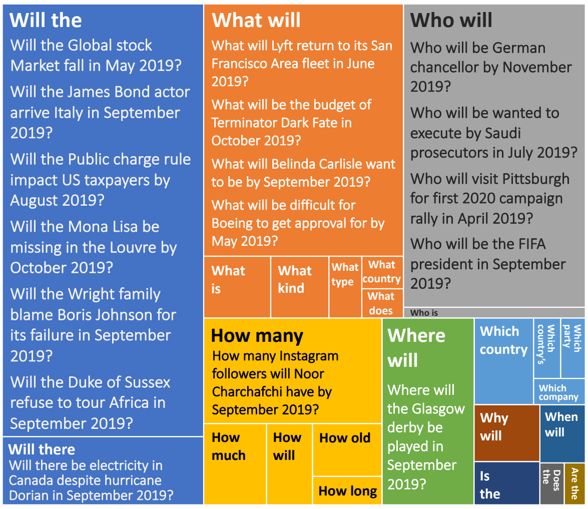


Figure 2: **A treemap visualization of first two words in FORECASTQA questions.** Box area is proportional to number of occurrences.

Statistic	Train	Dev	Test	All
Questions	8,210	1,090	1,092	10,392
Yes-no questions	4,737	582	584	5,903
Multi-choice questions	3,473	508	508	4,489

Table 2: **Size of the FORECASTQA dataset.**

There is also the difficulty of ensuring the quality of question generation via crowdsourcing (necessary when building a dataset of scale), due to possible human errors in question formation (Tetlock et al., 2017). We have taken steps to ensure our questions cannot be answered *with certainty* using "past" data given the time constraint or commonsense knowledge, but the questions are *tractable* to answer with an educated guess (see Sec. 4.1).[3]

Task Definition. Formally, the input of the FORECASTQA task is a forecasting question Q with a corresponding ending timestamp t_Q—the last possible date where Q remains a forecasting question. In addition, we have a set of possible choices, $\mathcal{C}$, and a corpus of news articles, $\mathcal{A}$; the output is a choice $C \in \mathcal{C}$. Our task has a novel *constraint* that any retrieved article $A \in \mathcal{A}$ must satisfy $t_A < t_Q$. In other words, models have access only to articles that are published before t_Q. We have ensured that the information required to solve the question *deterministically* comes out in an article, *gold article*, published after t_Q, i.e., $t_{\text{gold\_article}} \geq t_Q$. Another way to think of our setup is that we are asking Q on the day before t_Q, knowing that the information required to solve Q is not available yet. This for-

[3] This is in contrast to open-domain QA (machine reading comprehension) (Kwiatkowski et al., 2019) where answers can always be found in some given passages.

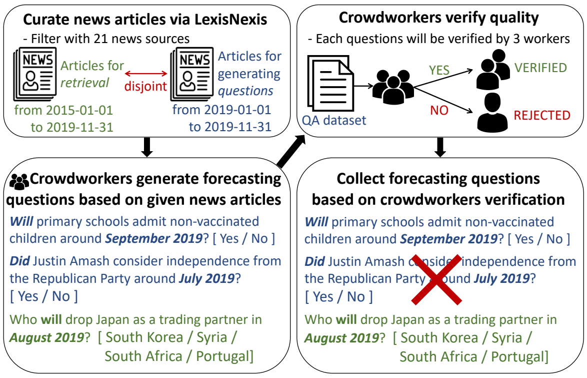

Curate news articles via LexisNexis
- Filter with 21 news sources
NEWS
Articles for retrieval
from 2015-01-01 to 2019-11-31
disjoint
NEWS
Articles for generating questions
from 2019-01-01 to 2019-11-31
Crowdworkers verify quality
- Each questions will be verified by 3 workers
QA dataset
YES
VERIFIED
NO
REJECTED
Crowdworkers generate forecasting questions based on given news articles
Will primary schools admit non-vaccinated children around September 2019? [Yes / No]
Did Justin Amash consider independence from the Republican Party around July 2019? [Yes / No]
Who will drop Japan as a trading partner in August 2019? [South Korea / Syria / South Africa / Portugal]
Collect forecasting questions based on crowdworkers verification
Will primary schools admit non-vaccinated children around September 2019? [Yes / No]
Did Justin Amash consider independence from the Republican Party around July 2019? [Yes / No]
Who will drop Japan as a trading partner in August 2019? [South Korea / Syria / South Africa / Portugal]

Figure 3: **FORECASTQA generation process.** The input of FORECASTQA creation is a news article corpus and the output is yes-no/multiple-choice questions.

mulation makes our task both a constrained open-domain QA and a forecasting problem—distinct from existing QA tasks.

Challenges in FORECASTQA. Due to the constrained open-domain setting and forecasting properties, testing a model's *forecasting ability* encompasses the following challenges: information retrieval (IR) on limited sources, understanding of temporal and causal relations between events, and finally a forecasting judgement. Our time constraint limits the accessible articles and also creates more challenges than in standard open-domain QA; effective IR methods are necessary to anticipate what knowledge will be useful for predictions from past information sources. Once useful articles have been retrieved, models should understand these articles and reason over pertinent facts from them. Finally, these models use the gleaned knowledge to infer the outcome of a future event. Unlike in other reading comprehension tasks, models cannot rely on the existence of an answer within the text, but must make an educated guess as to what will happen in the future. While our task does encompass reasoning abilities tested in other datasets, no other tasks investigate these reasoning abilities in the context of predicting future events. More analysis on reasoning types can be found in Sec. 4.2.

4 Dataset Construction and Analysis

In this section, we describe how we construct our FORECASTQA dataset and analyze it.

4.1 Construction Details

The data collection is broken down into three sections: (1) gathering a news corpus, (2) generating question-answer-timestamp triples with distractor choices, and (3) verifying the triples' quality. The data generation process is summarized in Fig. 3.

News Corpus Collection. We started by gathering English news articles from LexisNexis[4]. We then curated a list of 21 trustful news sources and filtered articles based on their publishers; we also filtered out non-English articles. Finally, we selected the five-year period of 2015-2019 and filtered out articles outside this period, leaving us with 509,776 articles. This corpus is also used for retrieval in our task setting (*i.e.*, constrained open-domain).

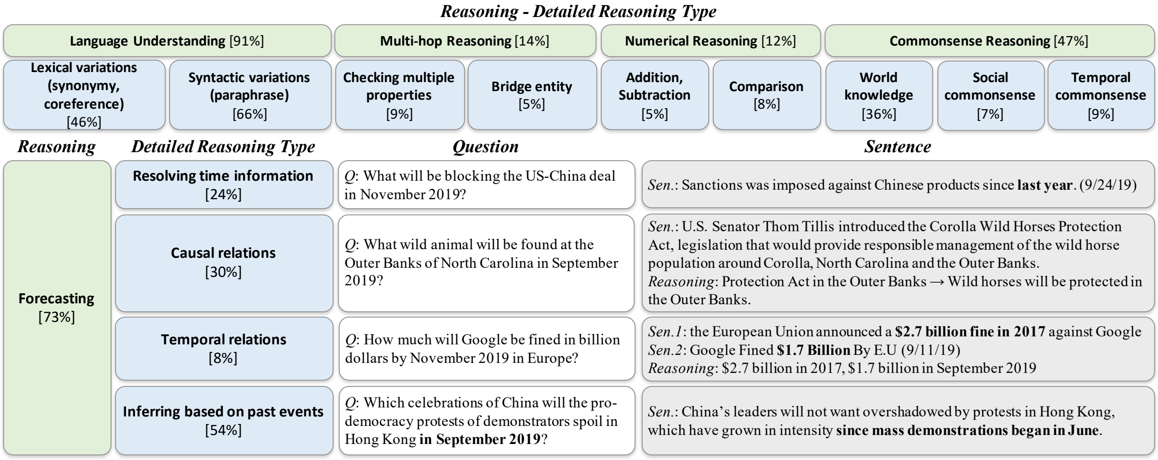
Q-Answer-timestamp Triple Creation.[5] Once we assembled the news corpus, we built (question, answer, timestamp) triples to accompany the new corpus as inputs for our task. To generate the needed triples we looked to crowdsourcing via Amazon Mechanical Turk. Our generation task consists of the following steps: (1) we selected a random news article from 2019 from the collected news corpus (these news articles are *gold articles* and will be hidden for experiments); (2) workers created questions, which *if posed before the respective article's publication date* would be seen as a forecasting question; (3) they indicated the answer, along with supporting evidence that the question consisted of (to ensure the correctness of the true answer); (4) they were asked to make multiple-choice distractors with their own knowledge and/or access to search engines; and (5) we ensured that a temporal phrase is present in the questions, for example: "*After May of 2020...*", "... *in June of 2021?*" to provide a temporal context (constraint) for each question, yielding more precise and well-defined forecasting questions. Completion of this task results in the desired triple of: a forecasting question, an answer to the question (with distractor choices), and a timestamp as our temporal constraint. The timestamp is set as the first day of the month in which the gold article was published.

To diversify questions in the dataset, we created two kinds of questions: binary yes-no questions and multiple-choice questions with *four choices*. Multiple-choice questions start with one of the six Ws (*i.e.*, who, what, when, where, why, and how) and are more challenging as they require determining the correctness of each choice.

Question Quality Verification. We performed a separate crowdsourcing data verification to test and enforce the following criteria: (1) is answering the question a *tractable* problem given (relevant)

[4] https://risk.lexisnexis.com

[5] Due to the limited space, for more details of our triple creation guidelines for human annotators, verification steps, and screenshots of our data collection/verification AMT interfaces, please refer to Sec. A of the appendix.



Reasoning - Detailed Reasoning Type

Language Understanding [91%]		Multi-hop Reasoning [14%]		Numerical Reasoning [12%]		Commonsense Reasoning [47%]		
Lexical variations (synonymy, coreference) [46%]	Syntactic variations (paraphrase) [66%]	Checking multiple properties [9%]	Bridge entity [5%]	Addition, Subtraction [5%]	Comparison [8%]	World knowledge [36%]	Social commonsense [7%]	Temporal commonsense [9%]

Reasoning	*Detailed Reasoning Type*	*Question*	*Sentence*
Forecasting [73%]	Resolving time information [24%]	*Q*: What will be blocking the US-China deal in November 2019?	*Sen.*: Sanctions was imposed against Chinese products since **last year**. (9/24/19)
	Causal relations [30%]	*Q*: What wild animal will be found at the Outer Banks of North Carolina in September 2019?	*Sen.*: U.S. Senator Thom Tillis introduced the Corolla Wild Horses Protection Act, legislation that would provide responsible management of the wild horse population around Corolla, North Carolina and the Outer Banks. *Reasoning*: Protection Act in the Outer Banks → Wild horses will be protected in the Outer Banks.
	Temporal relations [8%]	*Q*: How much will Google be fined in billion dollars by November 2019 in Europe?	*Sen.1*: the European Union announced a **$2.7 billion fine in 2017** against Google *Sen.2*: Google Fined **$1.7 Billion** By E.U (9/11/19) *Reasoning*: $2.7 billion in 2017, $1.7 billion in September 2019
	Inferring based on past events [54%]	*Q*: Which celebrations of China will the pro-democracy protests of demonstrators spoil in Hong Kong **in September 2019**?	*Sen.*: China's leaders will not want overshadowed by protests in Hong Kong, which have grown in intensity **since mass demonstrations began in June**.

Figure 4: **Reasoning skills (types) and their frequency (in %) in the sampled data.** As each question can be labeled with multiple types, the total frequency does not sum to 100%. On average, 3 reasoning skills are required for each question. Examples of other reasoning types can be found in Fig. 11 in the appendix.

"past" articles?, and (2) is the question *deterministically* answerable given any article adhering to the question's temporal constraint? — If a question is too difficult, *i.e.*, an educated guess to the answer (when given relevant, constraint-adhering articles) is not possible, then we filter the question out. On the other hand, if the questions are answerable *with certainty* using "past" articles, or commonsense/world knowledge, then they are *not* considered to be forecasting questions. The desired response (majority vote from 3 annotators) is a "*yes*" for criterion (1) and "*no*" for (2), as that would show that the tuple of question and time constraint simulates the desired forecasting scenario. With the above method, we filtered out 31% of the questions collected in the triple creation step and were left with 5,704 yes-no questions and 4,513 multi-choice questions. More details about the verification step are included in Sec. A of the appendix.

4.2 Dataset Analysis

To better understand the properties of the questions in FORECASTQA, we examine: 1) a few data statistics 2) types of questions asked, and 3) the types of reasoning required to answer our questions.

Summary Statistics. FORECASTQA dataset is composed of 10,392 questions, divided into a 80/10/10 split of train, dev, and test data. Our 10k questions are roughly evenly split between multiple-choice and yes-no binary questions (Table 2). Over 17K distinct words were used to construct our questions and we have 218 unique time constraints associated with them; time constraints range from 2019-01-11 to 2019-11-12. We include additional statistics in Sec. D of appendix.

Types of Questions. To understand the types of questions in FORECASTQA, we examined the popular beginnings of sentences and created a tree-map plot (see Fig. 2). As shown, nearly half the questions start with the word *will* (44%), a result of over half of the questions being yes-no questions.

Reasoning Types. To examine types of reasoning required to answer our questions we sampled 100 questions and manually annotated them with reasoning types. Due to the forecasting nature of our dataset, we are particularly interested in questions containing the forecasting ability and thus spend more time looking into these questions. Our condensed results can be found in Figure 4, and more results from our cataloguing effort can be found in Sec. C of the appendix. Note that most questions contain more than one reasoning type.

5 Methods

To evaluate the forecasting capabilities of recent *multi-choice/binary* QA model architectures on FORECASTQA, we provide a comprehensive benchmarking analysis in this work. We run the experiments in two settings: (1) *closed-book* and (2) *constrained open-domain* setup. In the *closed-book* scenario only Q (question) and C (answer choices) are provided to the model (Q, C), while $\overline{A}$ (news articles) is provided for setting (2), $(Q, C, \overline{A})$[6]. We run these settings to understand the difficulty of both the closed-book and open-domain challenges presented by the questions in FORECASTQA.

[6] t_Q is always applied to $\overline{A}$, we left it out of the notation for simplicity.

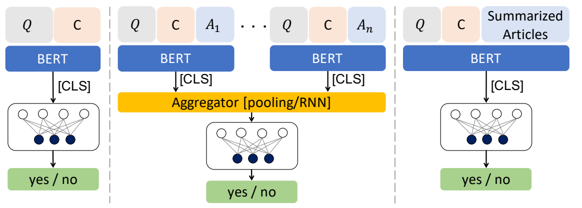


Figure 5: **Our baseline model architectures.** The CLS token is either fed into an MLP for classification or to the aggregator, which collects the information from each article before classifying.

For both settings, we explore several baseline models, but all follows a general architecture of a text encoder f and an optional context aggregation module g to aggregate information from a set of retrieved articles. Fig. 5 shows the architectures used. We model both yes-no and multiple-choice questions as a *binary* classification task; a model's prediction is the class with the largest probability. Below we introduce the details of our baselines.

Text Encoder. We use pre-trained language model, BERT (Devlin et al., 2019), as a text encoder (f from above)[7]. f is designed to deal with (Q, C) and $(Q, C, \overline{A})$ inputs, where $\overline{A}$ is a set of time-stamped articles that are retrieved from $\mathcal{A}$ to answer Q. Each input of f is transformed into $[\texttt{[CLS]}\,Q\,\texttt{[SEP]}\,C\,\texttt{[SEP]}\,A_i]$ (for each $A_i \in \overline{A}$, $C \in \mathcal{C}$), or $[\texttt{[CLS]}\,Q\,\texttt{[SEP]}\,C]$ (for each $C \in \mathcal{C}$) if articles are not supplied. The `[CLS]` token is the same as the one commonly used for fine-tuning PTLMs for a classification task, and `[SEP]` is the special separator token. The embedding of `[CLS]` is then used for predictions with an MLP layer (the leftmost model architecture in Fig. 5), or as input into a context aggregation module (the middle architecture in Fig. 5) subsequently introduced.

Context Aggregation (AGG). Two architectures are used when aggregating information from multiple, time-stamped articles $\overline{A}$ retrieved for a question. (1) *Temporal Aggregation:* This aggregator utilizes temporal ordering of the retrieved articles. Articles are sorted by their timestamps and their `[CLS]` token representation from f are aggregated by a Gated Recurrent Unit (GRU) (Cho et al., 2014) with a MLP head to make final predictions. (2) *Set Aggregation:* Alternatively, we ignore the temporal ordering of articles and use a *maxpooling operation* on the `[CLS]` token representations of each article. This pooled representation is passed to an MLP layer to make a prediction. Comparison between these aggregations helps understand the effect of modeling temporal order of evidence. These two aggregation modules are denoted by "AGG (GRU)" and "AGG (Maxpool)," respectively.

Multi-document Summarization (MDS). Rather than conducting context aggregation of the retrieved articles, we consider an MMR summarizer (Carbonell and Goldstein, 1998) which performs extractive, multi-document summarization of text to generate a summary A_{summ} (rightmost architecture in Fig. 5). The summary article A_{summ} is treated as if it is an $A_i \in \overline{A}$ and fed into a text encoder along with Q and C which then produce the `[CLS]` embedding for making a prediction. We name this method "MDS."

Integrated Approach. To take the best of both worlds in (Q, C) and $(Q, C, \overline{A})$ settings, we integrate two architectures (the leftmost and middle ones in Fig. 5). We concatenate the last two hidden representations of each architecture before passing the concatenated representation through a shared MLP layer. We use BERT$_{\text{LARGE}}$ as f in both architectures, AGG (GRU) for g and call this model "BERT$_{\text{LARGE}}$ ++ (integrated)" in Table 3.

Other Baselines. We also consider other baselines: ESIM (Chen et al., 2017b), BiDAF++ (Clark and Gardner, 2018), prepending extracted open event triples (Liu et al., 2019a) to BERT input, and a script learning approach, SAM-Net (Lv et al., 2019). We modify the approaches to fit into our setup. Detailed descriptions of each baseline method are included in Sec. E.3 of appendix.

6 Experiments

6.1 Experimental Setup

We adopt two types of settings: the closed-book setting (Q, C) and the constrained open-domain setting $(Q, C, \overline{A})$. In the constrained open-domain setting, we use BM25 (Robertson et al., 1995; Qi et al., 2019) as our IR method[8] to obtain $\overline{A}$, 10 retrieved articles. We also explore other IR methods in the later section. Note that we retrieve articles that do not violate the time constraints. We feed the question Q as a query and limit our access to articles in $\mathcal{A}$ by t_Q. Additionally, we validate the

[7] We did not include more recent pre-trained language models (*e.g.*, RoBERTa (Liu et al., 2019b), ALBERT (Lan et al., 2020), T5 (Raffel et al., 2020)) or pre-trained QA models like UnifiedQA (Khashabi et al., 2020), as these models are trained using text data published *after* the earliest timestamp in our dataset (2019-01-01), meaning information leakage could occur (and violates the forecasting setup). We tested more LMs in Sec. E.5 of appendix.

[8] Details of IR methods are described in appendix Sec. E.2.

Methods / Metrics	Accuracy (%, ↑)			Brier score (↓)		
	yes/no	*multi*	*all*	*yes/no*	*multi*	*all*
Random	48.6	25.3	37.8	0.684	0.827	0.750
ESIM-ELMo (closed-book)	63.3	45.8	54.5	0.515	0.897	0.706
BERT$_{\text{BASE}}$ (closed-book)	66.2	41.5	54.7	0.511	0.715	0.606
BERT$_{\text{LARGE}}$ (closed-book)	67.3	45.4	57.6	0.447	0.653	**0.543**
BIDAF++ (Clark and Gardner, 2018)	51.7	30.1	40.9	0.478	0.898	0.688
BERT$_{\text{BASE}}$, MDS	63.1	39.1	52.0	0.504	0.716	0.603
BERT$_{\text{BASE}}$, AGG (Maxpool)	67.2	39.1	54.2	0.453	0.701	0.568
BERT$_{\text{BASE}}$, AGG (GRU)	67.6	41.5	55.4	0.477	0.705	0.583
SAM-Net (Lv et al., 2019)	64.5	40.9	53.5	0.531	0.719	0.619
BERT$_{\text{LARGE}}$, MDS	67.4	40.1	54.7	0.542	0.738	0.633
BERT$_{\text{LARGE}}$, Event triples	66.7	45.0	56.6	0.589	0.719	0.649
BERT$_{\text{LARGE}}$, AGG (Maxpool)	68.8	46.9	58.6	0.476	**0.648**	0.556
BERT$_{\text{LARGE}}$, AGG (GRU)	69.2	47.5	59.1	0.483	0.655	0.563
BERT$_{\text{LARGE}}$, AGG (Maxpool), DPR	70.2	47.0	59.4	0.554	0.728	0.635
BERT$_{\text{LARGE}}$, AGG (Maxpool), BT	70.0	48.0	59.7	**0.444**	0.662	0.545
BERT$_{\text{LARGE}}$ ++ (integrated)	**70.3**	**48.4**	**60.1**	0.537	0.650	0.589
Human performance$^{(\alpha)}$	74.6	64.9	71.2	-	-	-
Human performance$^{(\beta)}$	81.3	77.4	79.4	-	-	-

Table 3: **Performance of baseline models on FORECASTQA test set.** "yes/no" refers to yes-no questions, and "multi" to multi-choice questions. We test the closed-book setting, and the constrained open-domain setting, where the accessible articles are limited by t_Q, our time constraint. We use BM25 as the article retriever to select top-10 articles, if not particularly specified. "BT" concatenates the binary encoding of date string to an article encoding before aggregation (see Sec. 6.3 "Ablation on Timestamp Modeling"). Human performance is based on the top-10 retrieved articles (α), and Google Search with the question's time constraint (β).

answerability of our questions by providing gold articles instead of retrieved articles (Sec. 6.3).

Evaluation Metrics. Because forecasting is uncertain, a system's prediction probabilities indicate its confidence answering the question. In addition to accuracy, we consider Brier score (Brier, 1950), which measures the mean squared *error* of probabilities assigned to sets of answer choices (outcomes). Formally, Brier $= \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}(p_{ic} - y_{ic})^2$, where p_{ic} is the probability of prediction; y_{ic} is a label indicator for class c of the instance (1 or 0), N is the number of prediction instances, and C is the number of classes (2 or 4). The highest Brier score is 0 (probability 1 for the correct class, probability 0 else), while the worst possible Brier score is 2 (probability 1 for the wrong class, probability 0 else). A confident model gets low Brier scores.

6.2 Human Performance

To benchmark human performance, seven annotators (computer science graduate students) who were not involved in question generation were asked to answer 150 randomly sampled questions from the test set. We consider two scenarios: 1) annotators are provided with retrieved articles, $\overline{A}$; and 2) annotators can access any article published *before the timestamp* via Google Search. Moreover, as annotators live in the "future" with respect to the timestamp of a question, they might already know the actual answer. To avoid the over-estimation of accuracy, we asked the annotators to not use their "future" knowledge. If they felt this is not possible, we asked them to skip the question. On average, 28.3% of questions are skipped. Given this setup, humans achieve 71.2% and 79.4% accuracy respectively, for the two scenarios when taking a majority vote for each question; we also observed good inter-annotator agreement. The two scenarios are referred as "(α)" and "(β)" in Table 3.

Methods	GRU	Maxpool	MDS
BERT$_{\text{BASE}}$, TF-IDF	53.2	53.9	51.6
BERT$_{\text{BASE}}$, DPR	53.7	**54.6**	**54.3**
BERT$_{\text{BASE}}$, BM25	**55.4**	54.2	52.0
BERT$_{\text{LARGE}}$, TF-IDF	56.5	55.4	**55.0**
BERT$_{\text{LARGE}}$, DPR	56.1	**59.4**	54.6
BERT$_{\text{LARGE}}$, BM25	**59.1**	58.6	54.7

Table 4: **Accuracy with different retrievers:** BM25, TF-IDF, and dense passage retrieval (DPR). We test the retrievers with different aggregators: GRU, Maxpool, and MDS.

6.3 Results and Performance Analysis

Results on the Constrained Open-domain Setting. Table 3 shows the results of baseline methods for comparison. We compare pre-trained language models with different context aggregators and other baselines. The integrated model, BERT$_{\text{LARGE}}$ ++ shows the best performance in terms of accuracy, while BERT$_{\text{LARGE}}$ (closed-book) shows the best Brier score. Unlike the accuracy metric, the Brier score penalizes over- and under- confident forecasts (Mellers et al., 2014) — thus the best model under each metric can be different. The marginal differences in performance between the two settings suggest that access to information (text evidence) alone does not solve the forecasting problem. We hypothesize an inability to encode salient relations for forecasting purposes prevents the additional information from proving useful. Among the aggregators in BERT$_{\text{BASE}}$, the GRU aggregator outperforms other aggregators and summarizers. This suggests that utilizing articles' temporal order helps the reasoning. Overall, baselines fall behind human performance by over 10% points given the same retrieved articles.

Study of Different IR Methods. We further test several retrieval methods: BM25 (Robertson et al., 1995; Qi et al., 2019), TF-IDF (Chen et al., 2017a), and a pre-trained dense passage retriever (DPR) (Karpukhin et al., 2020). As in Table 4, BERT$_{\text{LARGE}}$ with DPR retriever and the Maxpool aggregator shows the best performance than other combinations. However, DPR does not achieve the best accuracy for all methods. This implies that 1)

Methods / Metrics	GRU		Maxpool	
	ACC (↑)	Brier (↓)	ACC (↑)	Brier (↓)
w/o timestamps	**55.4**	**0.583**	54.2	**0.568**
Pre-pend timestamps	54.2	0.634	54.8	0.599
Binary timestamp encoding	51.1	0.623	**55.6**	0.624
Char-RNN timestamp encoding	54.0	0.640	54.3	0.620

Table 5: **Study on modeling article timestamps (publication dates) in the constrained open-domain setting.** We test several methods for temporal modeling and use BERT$_{\text{BASE}}$ with two different aggregators: GRU and Maxpool.

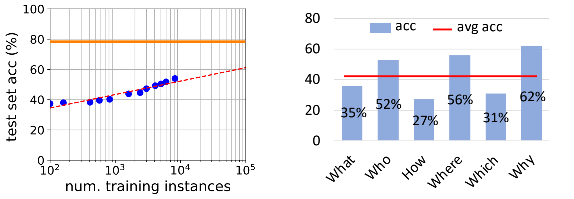
Methods / Metrics	Accuracy (↑)			Brier score (↓)		
	yes/no	*multi*	*all*	*yes/no*	*multi*	*all*
Random	48.6	25.3	37.8	0.684	0.827	0.750
Question	66.2	41.5	54.7	0.511	0.715	0.606
Article	73.6	80.7	76.9	0.428	0.263	0.351
Evidence sentence	79.9	89.5	84.4	0.355	0.171	0.269

Table 6: **Answerability study on test set.** Instead of retrieved articles, we provide BERT$_{\text{BASE}}$ with ground-truth context: a gold article or evidence sentence. We thus convert FORECASTQA to a reading comprehension task and examine the answerability of the questions.

stronger retrieval methods are required to identify useful evidence; 2) complex forecasting abilities may be a bottleneck of current systems.

Ablation on Timestamp Modeling. We conduct an ablation study on modeling time information (publication date) of the retrieved articles, as seen in Table 5. We test: a) pre-pending date string as BERT input, b) using binary encodings of dates[9] and concatenate with article encoding before aggregation, and c) using char-RNN (Goyal and Durrett, 2019) for encoding date string before aggregation[10]. We find that using binary encodings of dates improves the accuracy for the maxpool aggregator. However, the GRU aggregator's accuracy decreases when given date information. We conjecture that our modeling for the time information of each article is not strong enough to help forecasting. We leave more sophisticated modeling for future work.

Answerability of Questions. To validate that the questions in FORECASTQA are indeed answerable, we convert our setup into a machine reading comprehension (MRC) task — find an answer given an assumed appropriate context. We provide the model with a gold article or the evidence sentence (Sec. 4.1). Since pre-trained models have achieved high performance on MRC tasks (Rajpurkar et al., 2016), we expect adequate performance when provided the correct context. As seen in Table 6, we observe that in closed-book setting, BERT is able to beat out a random baseline, but it still does not perform well; implying our questions are not trivial for BERT, and context is required to answer them correctly. When given the gold article, BERT achieves 76.9% (+22%) and it even performs better (84.4%) given the evidence sentence. This all implies that given the right information, our forecasting questions can be answered correctly.



(a) Varying amounts of data. (b) Different question types.

Figure 6: (a) Test accuracy of BERT$_{\text{BASE}}$ trained with varying amounts of training data, with human performance (79.1%) shown in orange, and (b) development accuracy breakdown by different types of multichoice questions.

Study of Data Efficiency. To examine how models might perform with less/more training data, we evaluate BERT$_{\text{BASE}}$ (closed-book) on the test set, by training it with varying amounts of labeled data. Fig. 6a shows the the resulting "learning curve." We observe the accuracy of the model is "expected" to reach 70%, assuming 100k examples — which is still 9% point lower than human performance.

Results on Different Question Types. We test BERT$_{\text{BASE}}$ (closed-book) on different question types of multi-choice questions from our development set (Fig. 6b). We find that the accuracy of the model varies across different question types: "*how*" questions are the most difficult to predict while higher accuracy is achieved on "*why*" questions. Also for yes-no questions, the method achieves 69.5% on "*yes*" questions and 62.9% "*no*" questions, indicating that there is no significant bias towards certain type of binary questions.

Error Analysis. We observe 4 main categories of errors produced by the methods in our analysis: (1) retrieving irrelevant articles, (2) incorrect reasoning on relevant evidence, (3) lacking (temporal) common sense, and (4) lacking numerical knowledge. Please refer to Sec. E.7 of appendix for examples and in-depth discussions of these errors.

7 Conclusion

Forecasting is a difficult task that requires every possible advantage to do well. It would be wise to harness this pool of unstructured data for training automatic event forecasting agents. To utilize this form of data for forecasting, we proposed a

[9] `https://temporenc.org`

[10] Details are described in appendix Sec. E.4

question-answering task that requires forecasting skills to solve FORECASTQA, and provided the accompanying dataset. Various baseline methods did not perform well, but this is not surprising given the inherent difficulty of forecasting. Our benchmark dataset can benefit future research beyond natural language understanding and hope forecasting performance will be significantly improved.

References

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. 2015. Icews coded event data. *Harvard Dataverse*, 12.

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017a. 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.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017b. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. 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.

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.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963.

Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine.* " O'Reilly Media, Inc.".

Tanya Goyal and Greg Durrett. 2019. Embedding time expressions for deep temporal ordering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4400–4406, Florence, Italy. Association for Computational Linguistics.

Linmei Hu, Juanzi Li, Liqiang Nie, Xiaoli Li, and Chao Shao. 2017. What happens next? future subevent prediction using contextual hierarchical LSTM. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3450–3456. AAAI Press.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018a. Tempquestions: A benchmark for temporal question answering. In *WWW*.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018b. TEQUILA: temporal question answering over knowledge bases. In *Proceedings of the 27th ACM International Conference on Information and*

Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, pages 1807–1810. ACM.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

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*, 7:452–466.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Constructing narrative event evolutionary graph for script event prediction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4201–4207. ijcai.org.

Xiao Liu, Heyan Huang, and Yue Zhang. 2019a. Open domain event extraction using neural latent variable models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2860–2871, Florence, Italy. Association for Computational Linguistics.

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. *ArXiv*, abs/1907.11692.

Shangwen Lv, Wanhui Qian, Longtao Huang, Jizhong Han, and Songlin Hu. 2019. Sam-net: Integrating event-level and chain-level attentions to predict what happens next. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6802–6809. AAAI Press.

Barbara Mellers, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E Scott, Don Moore, Pavel Atanasov, Samuel A Swift, et al. 2014. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science*, 25(5):1106–1115.

Fred Morstatter, Aram Galstyan, Gleb Satyukov, Daniel Benjamin, Andrés Abeliuk, Mehrnoosh Mirtaheri, KSM Tozammel Hossain, Pedro A. Szekely, Emilio Ferrara, Akira Matsui, Mark Steyvers, Stephen Bennett, David V. Budescu, Mark Himmelstein, Michael D. Ward, Andreas Beger, Michele Catasta, Rok Sosic, Jure Leskovec, Pavel Atanasov, Regina Joseph, Rajiv Sethi, and Ali E. Abbas. 2019. SAGE: A hybrid geopolitical event forecasting system. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6557–6559. ijcai.org.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

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.

Ulrich Pilster and Tobias Böhmelt. 2014. Predicting the duration of the syrian insurgency. *Research & Politics*, 1(2):2053168014544586.

Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D. Manning. 2019. Answering complex open-domain questions through iterative query generation. 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 2590–2602, Hong Kong, China. Association for Computational Linguistics.

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

Vasanthan Raghavan, Aram Galstyan, and Alexander G. Tartakovsky. 2013. Hidden markov models for the activity profile of terrorist groups. *Ann. Appl. Stat.*, 7(4):2402–2430.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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.

Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Paul Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, Chris J. Kuhlman, Achla Marathe, Liang Zhao, Ting Hua, Feng Chen, Chang-Tien Lu, Bert Huang, Aravind Srinivasan, Khoa Trinh, Lise Getoor, Graham Katz, Andy Doyle, Chris Ackermann, Ilya Zavorin, Jim Ford, Kristen Maria Summers, Youssef Fayed, Jaime Arredondo, Dipak Gupta, and David Mares. 2014a. 'beating the news' with EMBERS: forecasting civil unrest using open source indicators. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 1799–1808. ACM.

Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Paul Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, Chris J. Kuhlman, Achla Marathe, Liang Zhao, Ting Hua, Feng Chen, Chang-Tien Lu, Bert Huang, Aravind Srinivasan, Khoa Trinh, Lise Getoor, Graham Katz, Andy Doyle, Chris Ackermann, Ilya Zavorin, Jim Ford, Kristen Maria Summers, Youssef Fayed, Jaime Arredondo, Dipak Gupta, and David Mares. 2014b. 'beating the news' with EMBERS: forecasting civil unrest using open source indicators. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 1799–1808. ACM.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Sebastian Schutte. 2017. Regions at risk: Predicting conflict zones in african insurgencies. *Political Science Research and Methods*, 5(3):447–465.

Yawei Sun, Gong Cheng, and Yuzhong Qu. 2018. Reading comprehension with graph-based temporal-casual reasoning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 806–817, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

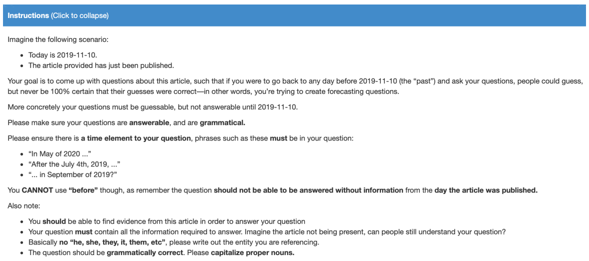
Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Philip Tetlock, Barbara A. Mellers, and J. Peter Scoblic. 2017. Bringing probability judgments into policy debates via forecasting tournaments. *Science*, 355:481–483.

Philip E Tetlock and Dan Gardner. 2016. *Superforecasting: The art and science of prediction*. Random House.

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.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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 3363–3369, Hong Kong, China. Association for Computational Linguistics.



Instructions (Click to collapse)

Imagine the following scenario:

- Today is 2019-11-10.
- The article provided has just been published.

Your goal is to come up with questions about this article, such that if you were to go back to any day before 2019-11-10 (the "past") and ask your questions, people could guess, but never be 100% certain that their guesses were correct—in other words, you're trying to create forecasting questions.

More concretely your questions must be guessable, but not answerable until 2019-11-10.

Please make sure your questions are **answerable**, and are **grammatical.**

Please ensure there is **a time element to your question**, phrases such as these **must** be in your question:

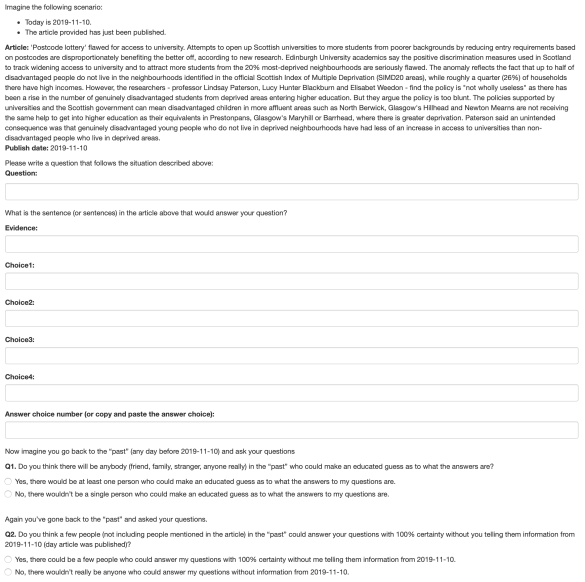
- "In May of 2020 ..."
- "After the July 4th, 2019, ..."
- "... in September of 2019?"

You **CANNOT** use **"before"** though, as remember the question **should not be able to be answered without information** from the **day the article was published.**

Also note:

- You **should** be able to find evidence from this article in order to answer your question
- Your question **must** contain all the information required to answer. Imagine the article not being present, can people still understand your question?
- Basically **no "he, she, they, it, them, etc"**, please write out the entity you are referencing.
- The question should be **grammatically correct**. Please **capitalize proper nouns.**

Figure 7: Instruction of creating multiple-choice questions.



Imagine the following scenario:

- Today is 2019-11-10.
- The article provided has just been published.

Article: 'Postcode lottery' flawed for access to university. Attempts to open up Scottish universities to more students from poorer backgrounds by reducing entry requirements based on postcodes are disproportionately benefiting the better off, according to new research. Edinburgh University academics say the positive discrimination measures used in Scotland to track widening access to university and to attract more students from the 20% most-deprived neighbourhoods are seriously flawed. The anomaly reflects the fact that up to half of disadvantaged people do not live in the neighbourhoods identified in the official Scottish Index of Multiple Deprivation (SIMD20 areas), while roughly a quarter (26%) of households there have high incomes. However, the researchers - professor Lindsay Paterson, Lucy Hunter Blackburn and Elisabet Weedon - find the policy is "not wholly useless" as there has been a rise in the number of genuinely disadvantaged students from deprived areas entering higher education. But they argue the policy is too blunt. The policies supported by universities and the Scottish government can mean disadvantaged children in more affluent areas such as North Berwick, Glasgow's Hillhead and Newton Mearns are not receiving the same help to get into higher education as their equivalents in Prestonpans, Glasgow's Maryhill or Barrhead, where there is greater deprivation. Paterson said an unintended consequence was that genuinely disadvantaged young people who do not live in deprived neighbourhoods have had less of an increase in access to universities than non-disadvantaged people who live in deprived areas.
Publish date: 2019-11-10

Please write a question that follows the situation described above:
Question:

What is the sentence (or sentences) in the article above that would answer your question?

Evidence:

Choice1:

Choice2:

Choice3:

Choice4:

Answer choice number (or copy and paste the answer choice):

Now imagine you go back to the "past" (any day before 2019-11-10) and ask your questions

Q1. Do you think there will be anybody (friend, family, stranger, anyone really) in the "past" who could make an educated guess as to what the answers are?

○ Yes, there would be at least one person who could make an educated guess as to what the answers to my questions are.
○ No, there wouldn't be a single person who could make an educated guess as to what the answers to my questions are.

Again you've gone back to the "past" and asked your questions.

Q2. Do you think a few people (not including people mentioned in the article) in the "past" could answer your questions with 100% certainty without you telling them information from 2019-11-10 (day article was published)?

○ Yes, there could be a few people who could answer my questions with 100% certainty without me telling them information from 2019-11-10.
○ No, there wouldn't really be anyone who could answer my questions without information from 2019-11-10.

Figure 8: Interface of creating multiple-choice questions.

A Detailed Dataset Creation

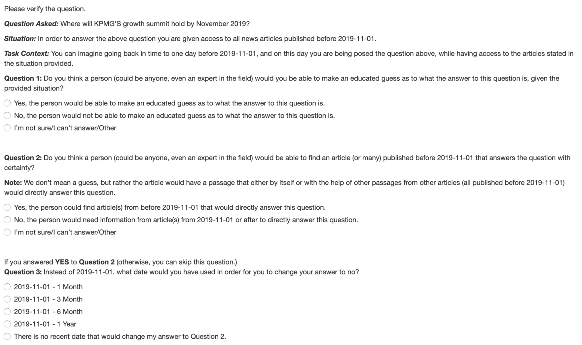
In this section, we present detailed explanations of dataset creation. We first selected news sources as in the following section.

A.1 List of News Sources

The New York Post, The New York Times, New York Magazine, Daily News (New York), The Washington Post, NPR All Things Considered, NPR Weekend Edition Saturday, NPR Morning Edition, CNN Wire, CNN.com, CNNMoney.com, CNN INTERNATIONAL, Fox News Network, York Guardian, Washingtonpost.com, The Washington Post Magazine, thetimes.co.uk, Guardian Weekly, Russia & CIS General Newswire, US Official News, The Times (London).

A.2 Dataset Creation

Turking Guidelines. Figs 7 and 8 show the instructions and interface for creating our multiple-choice questions. Workers made multiple-choice distractors with their own knowledge, but they were encouraged to find good distractors using search engines. To ensure the answerability of the created questions, we ask them to indicate the answer along with the supporting evidence that the question is made from. We omit the interfaces due to the space limit.



Please verify the question.

Question Asked: Where will KPMG'S growth summit hold by November 2019?

Situation: In order to answer the above question you are given access to all news articles published before 2019-11-01.

Task Context: You can imagine going back in time to one day before 2019-11-01, and on this day you are being posed the question above, while having access to the articles stated in the situation provided.

Question 1: Do you think a person (could be anyone, even an expert in the field) would you be able to make an educated guess as to what the answer to this question is, given the provided situation?

○ Yes, the person would be able to make an educated guess as to what the answer to this question is.
○ No, the person would not be able to make an educated guess as to what the answer to this question is.
○ I'm not sure/I can't answer/Other

Question 2: Do you think a person (could be anyone, even an expert in the field) would be able to find an article (or many) published before 2019-11-01 that answers the question with certainty?

Note: We don't mean a guess, but rather the article would have a passage that either by itself or with the help of other passages from other articles (all published before 2019-11-01) would directly answer this question.

○ Yes, the person could find article(s) from before 2019-11-01 that would directly answer this question.
○ No, the person would need information from article(s) from 2019-11-01 or after to directly answer this question.
○ I'm not sure/I can't answer/Other

If you answered **YES** to **Question 2** (otherwise, you can skip this question.)
Question 3: Instead of 2019-11-01, what date would you have used in order for you to change your answer to no?

○ 2019-11-01 - 1 Month
○ 2019-11-01 - 3 Month
○ 2019-11-01 - 6 Month
○ 2019-11-01 - 1 Year
○ There is no recent date that would change my answer to Question 2.

Figure 9: Interface of verifying questions.

Initial Screening. The ideal result of our crowdsourcing task are forecasting questions that are tractable but not trivial, and by definition not answerable with certitude using information currently available. Thus to avoid undesirable questions, we asked two additional questions to help screen poorly constructed questions. As shown in Fig 8, we try to determine the difficulty of the question and whether it is answerable using "current" or "past" information. Question 1 attempts to establish whether the question is indeed tractable and asks whether there exists some qualified group of people who could reason and make an educated guess at the answer to the question. On the other hand, question 2 tries to determine if the question is either too easy or is definitively answerable given "current" and "past" information. Thus, the desired response is "yes" and "no" for Questions 1 and 2, respectively; we filtered out created questions that do not satisfy the desired condition.

A.3 Additional Question Quality Checks

We asked the same two questions from our initial quality screening and an additional question to help adjust the timestamp associated with the question if needed. Per question, we got 3 crowd workers to answer the three questions and took the majority vote for question 1 and 2, while selecting the earliest selected timestamp for question 3. We dropped the question, if the majority vote was "no" for question 1 or "yes" for question 2. Moreover, if at least one worker selected "e" in the question 3 (There is no appropriate recent time stamp), then we filtered out the question. Additionally, if the created ques-

Q: What wild animal will be found at the Outer banks of North Carolina in September 2019?
Choices: Horses (*answer*), Cows, Turtles, Donkeys.

Article: *Tillis Introduces Legislation to Protect Corolla Wild Horses Washington: Office of the Senator Thom Tillis has issued the following news release:* (1/29/19)
U.S. Senator Thom Tillis (R-NC) introduced the Corolla Wild Horses Protection Act, legislation that would provide responsible management of the wild horse population around Corolla, North Carolina and the Outer Banks. Representative Walter Jones (R-NC) introduced companion legislation in the House of Representatives in previous Congresses and has been a long time champion of protecting the Corolla wild horse population.

Reasoning Process: The Corolla Wild Horses Protection Act will make people to protect the wild horses (**forecasting skills - causal relations**). If people start to protect the wild horses from January, the wild horses will be found in September (**forecasting skills - inferring based on past events** - *we can find the answer from this part*). Horse is an animal (**commonsense - world knowledge**). The Outer banks of North Carolina = North Carolina and the Outer Banks (**language understanding - paraphrase**).

Table 7: Detailed example to show how to solve a question.

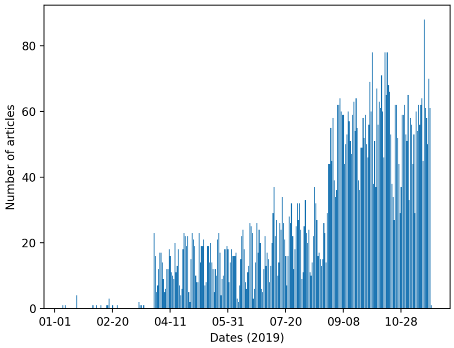
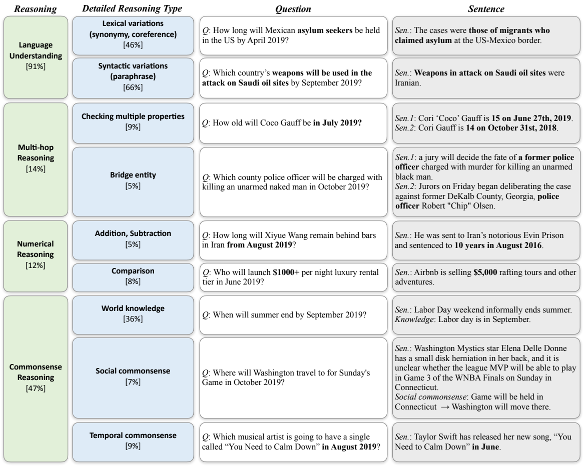


Figure 10: Date distribution of gold articles for questions. Each question is made from gold articles. The dates denote release dates of news articles and they range from 01-01-2019 to 11-31-2019.



Reasoning	*Detailed Reasoning Type*	*Question*	*Sentence*
Language Understanding [91%]	**Lexical variations (synonymy, coreference)** [46%]	*Q*: How long will Mexican **asylum seekers** be held in the US by April 2019?	*Sen.*: The cases were **those of migrants who claimed asylum** at the US-Mexico border.
	Syntactic variations (paraphrase) [66%]	*Q*: Which country's **weapons will be used in the attack on Saudi oil sites** by September 2019?	*Sen.*: **Weapons in attack on Saudi oil sites** were Iranian.
Multi-hop Reasoning [14%]	**Checking multiple properties** [9%]	*Q*: How old will Coco Gauff be **in July 2019?**	*Sen.1*: Cori 'Coco' Gauff is **15 on June 27th, 2019**. *Sen.2*: Cori Gauff is **14 on October 31st, 2018**.
	Bridge entity [5%]	*Q*: Which county police officer will be charged with killing an unarmed naked man in October 2019?	*Sen.1*: a jury will decide the fate of **a former police officer** charged with murder for killing an unarmed black man. *Sen.2*: Jurors on Friday began deliberating the case against former DeKalb County, Georgia, **police officer** Robert "Chip" Olsen.
Numerical Reasoning [12%]	**Addition, Subtraction** [5%]	*Q*: How long will Xiyue Wang remain behind bars in Iran **from August 2019**?	*Sen.*: He was sent to Iran's notorious Evin Prison and sentenced to **10 years in August 2016**.
	Comparison [8%]	*Q*: Who will launch **$1000+** per night luxury rental tier in June 2019?	*Sen.*: Airbnb is selling **$5,000** rafting tours and other adventures.
Commonsense Reasoning [47%]	**World knowledge** [36%]	*Q*: When will summer end by September 2019?	*Sen.*: Labor Day weekend informally ends summer. *Knowledge*: Labor day is in September.
	Social commonsense [7%]	*Q*: Where will Washington travel to for Sunday's Game in October 2019?	*Sen.*: Washington Mystics star Elena Delle Donne has a small disk herniation in her back, and it is unclear whether the league MVP will be able to play in Game 3 of the WNBA Finals on Sunday in Connecticut. *Social commonsense*: Game will be held in Connecticut → Washington will move there.
	Temporal commonsense [9%]	*Q*: Which musical artist is going to have a single called "You Need to Calm Down" **in August 2019**?	*Sen.*: Taylor Swift has released her new song, "You Need to Calm Down" **in June**.

Figure 11: Examples of each type of reasoning in FORECASTQA. Words relevant to the corresponding reasoning type are bolded. Also, [%] represents the percentage of questions that requires the reasoning type.

tion does not have a temporal phrase, then we filter out the question.

B Example of Reasoning

Table 7 shows an example of reasoning process to solve a question.

Measurement	Value
Average question length (tokens)	13.85
Average answer length (tokens)	2.46
# of distinct words in questions	17,521
# of distinct words in choices	5,187
# of distinct time stamps associated w. questions	218
Average gold article length (# tokens)	720.21
Maximum question time stamp	2019-11-22
Minimum question time stamp	2019-01-01

Table 8: Statistics of FORECASTQA.

C Additional Reasoning Types

Figure 11 shows additional reasoning types.

Language Understanding. We introduce lexical variations and syntactic variations following Rajpurkar et al. (2016, 2018). Lexical variations represent synonyms or coreferences between the question and the evidence sentence. When the question is paraphrased into another syntactic form and the evidence sentence is matched to the form, we call it syntactic variation. We find that many questions require language understanding; lexical variations account for 46% and syntactic variations do for 66%.

Multi-hop Reasoning. Some questions require multi-hop reasoning (Yang et al., 2018), such as checking multiple properties (9%) and bridge entities (5%) . The former one requires finding multiple properties from an article to find an answer. The latter one works as a bridge between two entities, where one must identify a bridge entity, and find the answer in the second hop.

Numerical Reasoning. To answer our questions, one needs numerical reasoning (Dua et al., 2019). The answer is found by adding or subtracting two numbers (5%), or comparing two numbers (8%) in the given articles.

Commonsense Reasoning. The questions require world knowledge (Talmor et al., 2019), social commonsense (Sap et al., 2019), and temporal commonsense (Zhou et al., 2019). To solve these questions, an AI agent must leverage assumed common knowledge in addition to what it finds in the news corpus. We find that 36% questions need world knowledge and 7% questions require social commonsense. The other type of commonsense reasoning is temporal commonsense, which is related to temporal knowledge (Zhou et al., 2019). 9% questions are related to temporal commonsense.

D Statistics

Tables 8 and 9 show the statistics and answer types in FORECASTQA.

Answer Type	%	Examples
Yes/No	56.8%	-
Person	8.1%	Boris Johnson, Mark Zuckerberg
Group/Org	5.8%	BBC, United Nations, EU
Location	8.0%	Canada, Iran, U.S.
Date/Time	1.6%	January, July
Number	6.7%	530, Thirty eight
Other Entity	1.1%	Boeing 737
Common Noun Phrase	5.8%	A hurricane, Asteroid dust
Verb Phrase	3.1%	Defend his innocence
Adjective Phrase	1.4%	Cruel and Misguided, Due to the bad weather
Sentence	1.6%	Liverpool will become the first English team to play their 400th international game.

Table 9: Types of answers in FORECASTQA.

E Experiments

E.1 Details on a Text Encoder

We use Huggingface's codes[11]. We chose the best learning rate among $\{3e-5, 1e-5, 5e-6\}$ and the number of epochs is 3. We set the max sequence length to 512.

E.2 Details on IR methods

We index the English news articles with Elasticsearch (Gormley and Tong, 2015). We followed the setups in Qi et al. (2019). We use Elasticsearch's simple analyzer which performs basic tokenization and lowercasing for the title. We use the standard analyzer which allows for removal of punctuation and stop words from the body of articles. At retrieval time, we use a `multi_match` query in the Elasticsearch against all fields with the same query, which performs a full-text query employing the BM25 ranking function (Robertson et al., 1995) on all fields, and returns the score of the best field for ranking. To promote documents whose title matches the search query, we boost the search score of any result whose title matches the search query by 1.25, which results in a better recall for entities with common names.

E.3 Details on Baselines.

We consider following baselines: (1) **Event-based approaches**: We test event-based approach, BERT with event triples (two entities and a relation between them) and BERT based on SAM-Net (Lv et al., 2019) for our setup. It is non-trivial to apply the event-based approaches to our setup. Thus, we preprocess the retrieved news articles into event triples (subject, relation, object) using Liu et al. (2019a). We simply regard them as text, we concatenate the triples, and feed them into BERT and call it **BERT with event triples**. In addition, we apply a script learning approach (SAM-Net (Lv et al., 2019)) to our setup. A question and choices are not used in their original method; thus we encode them using BERT and concatenate the encodings with the approach's final representation. This representation is fed into a linear layer and the linear layer predicts whether the choice is correct or not. We used $\text{BERT}_{\text{LARGE}}$ for the former one and $\text{BERT}_{\text{BASE}}$ for the latter one. (2) **ESIM** (Chen et al., 2017b). An NLI model, where we change their output layer so that the model outputs probabilities for each answer choice with a softmax layer. We use ELMo (Peters et al., 2018) for word embeddings. (3) **BIDAF++** (Clark and Gardner, 2018). The model requires context, and thus we use a top-1 article by an IR method. We augment it with a self-attention layer and ELMo representations. To adapt to the multiple-choice setting, we choose the answer with the highest probability. The input to ESIM is a question and a set of choices (Q, C), while that of BIDAF++'s is a question, a set of choices, and retrieved articles $(Q, C, \overline{A})$.[12]

E.4 Time Modeling

We conduct an ablation study on modeling time information of the retrieved articles. We test the following models: a) pre-pending date string as BERT input [[CLS] Q [SEP] C [SEP] *Date* [SEP] A_i], where the date format is "YYYY-MM-DD", b) using binary encodings of dates: we first encode the time into a binary encoding using "Temporenc[13]" and concatenate the encoding with an article encoding before aggregation, c) using char-RNN (Goyal and Durrett, 2019) for encoding date string before aggregation.

E.5 Experiments with Recent LMs.

As mentioned in Sec 5, we did not report more recent pre-trained language models (e.g., RoBERTa (Liu et al., 2019b), ALBERT (Lan et al., 2020)) because they are trained using text data published after the earliest timestamp in our dataset

[11] https://github.com/huggingface/transformers

[12] We did not include existing event forecasting methods since they are designed for modeling structured event data (Fawaz et al., 2019) and thus are not directly applicable to FORECASTQA which requires modeling of unstructured text.

[13] https://temporenc.org

Methods / Metrics	Accuracy		
	yes/no	*multi*	*all*
BERT$_{\text{BASE}}$, AGG (GRU)	67.6	41.5	55.4
RoBERTa$_{\text{BASE}}$, AGG (GRU)	69.3	44.8	57.9
ALBERT$_{\text{BASE}}$, AGG (GRU)	67.4	23.4	46.9
BERT$_{\text{LARGE}}$, AGG (GRU)	69.2	47.5	59.1
RoBERTa$_{\text{LARGE}}$, AGG (GRU)	70.1	51.3	61.3
ALBERT$_{\text{LARGE}}$, AGG (GRU)	68.4	30.2	50.6
Human performance	81.3	77.4	79.4

Table 10: Results on different pre-trained language models, BERT, RoBERTa, ALBERT).

Methods / Metrics	Accuracy (%, ↑)			Brier score (↓)		
	yes/no	*multi*	*all*	*yes/no*	*multi*	*all*
BERT$_{\text{LARGE}}$, AGG (GRU)	69.2	47.5	59.1	0.483	0.655	0.563
BERT$_{\text{LARGE}}$, GRU(A), QC	67.8	42.5	56.0	0.583	0.758	0.665

Table 11: Performance of baseline models on FORECASTQA test set.

Methods / Metrics	Accuracy (%)			Brier score		
	yes/no	*multi*	*all*	*yes/no*	*multi*	*all*
BERT$_{\text{BASE}}$						
– Question	65.6	43.7	55.4	0.506	0.698	0.596
– Article	78.1	84.8	81.2	0.351	0.210	0.285
– Evidence sentence	81.4	90.5	85.6	0.324	0.147	0.241

Table 12: Results on gold articles on the dev set. We give different inputs to the BERT to find out which part is important for the questions.

(2019-01-01). We are worried that these models in theory would have access to information that was published after the associated timestamp of a question.

As a reference, we show the results of RoBERTa and ALBERT in Table 10. Even though these two models may violate our forecasting scenario, they still struggle when compared to human performance, suggesting that our task is still challenging.

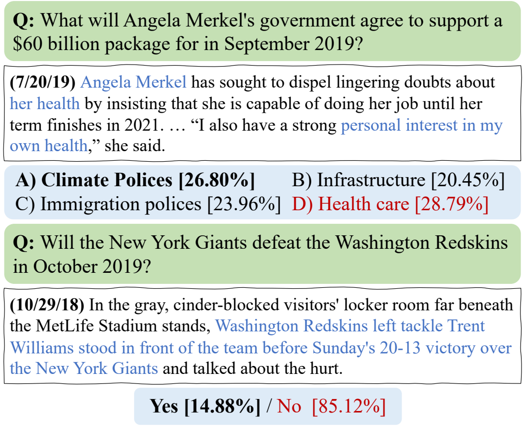
E.6 Experiments with different GRU architectures.

We investigate GRU modeling for the input. BERT$_{\text{LARGE}}$ GRU(A), QC refers to a model that encodes each article with a text encoder, these encodings are fed into GRU, and concatenate the last hidden representation of GRU and Q,C (question and choice) encoding from the text encoder. Table 11 shows comparison between the two architectures. Separating the articles with the question and choice leads to the worse performance.

E.7 Error Analysis

We randomly select 50 errors made by the best baseline method from the test set and identify 4 phenomena:

Retrieving Wrong Articles. 28% of the errors are from the retrieval of irrelevant articles. The baseline approach relies on information retrieval methods such as BM25. Retrieved articles might not be relevant or contain facts that can confuse the model, thus causing incorrect predictions. For example, consider the first question in Fig. 12, the model has retrieved an irrelevant article and conflated Ms. Merkel's health with policy decisions. This results in the model incorrectly choosing Health Care as the appropriate answer.



Q: What will Angela Merkel's government agree to support a $60 billion package for in September 2019?

(7/20/19) Angela Merkel has sought to dispel lingering doubts about her health by insisting that she is capable of doing her job until her term finishes in 2021. … "I also have a strong personal interest in my own health," she said.

A) Climate Polices [26.80%] B) Infrastructure [20.45%]
C) Immigration polices [23.96%] D) Health care [28.79%]

Q: Will the New York Giants defeat the Washington Redskins in October 2019?

(10/29/18) In the gray, cinder-blocked visitors' locker room far beneath the MetLife Stadium stands, Washington Redskins left tackle Trent Williams stood in front of the team before Sunday's 20-13 victory over the New York Giants and talked about the hurt.

Yes [14.88%] / No [85.12%]

Figure 12: Examples of erroneous model predictions. Bold choices are actual answers and red choices are model predictions.

Incorrect Use of Relevant Evidence. 24% of the errors are (partially) caused by incorrect usage of relevant evidence. Even though useful articles are retrieved, the model incorrectly reasons over the evidence. Take the second question in Fig. 12, where the model incorrectly predicts *No*. The model may depend on a relevant, but outdated fact from 2018 (one year before the event in question) to answer the question, and failed to incorporate more recent information.

Lacking Human Common Sense. 32% of the errors are from the model's lack of common sense or world knowledge. An example question is, *"Who will host 2020 Olympics by July 2019?,"* where the answer is Japan, but the model predicts Hong Kong. To answer this question, a model must know the cities of each country, as without this knowledge the model does not know that "Tokyo is in Japan," and thus the model predicts the wrong answer.

Numerical Questions. 8% of the errors are from numerical questions. Numerical questions ask about numbers such as a person's age. For example, *"What will be Roger Federer's age by August 2019."* The model must know his birth month and age and know how to increment on one's birthday.