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

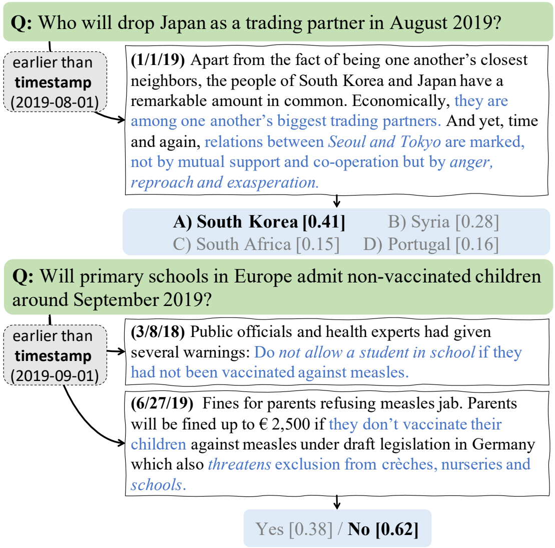
Woojeong Jin[1] **Rahul Khanna**[1] **Suji Kim**[1] **Dong-Ho Lee**[1]
Fred Morstatter[2] **Aram Galstyan**[2] **Xiang Ren**[1 2]
[1]Department of Computer Science, University of Southern California
[2]Information Sciences Institute, University of Southern California
{woojeong.jin, rahulkha, sujikim, 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?

(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.*

Q: Will primary schools in Europe admit non-vaccinated children around September 2019?

(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.*

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.

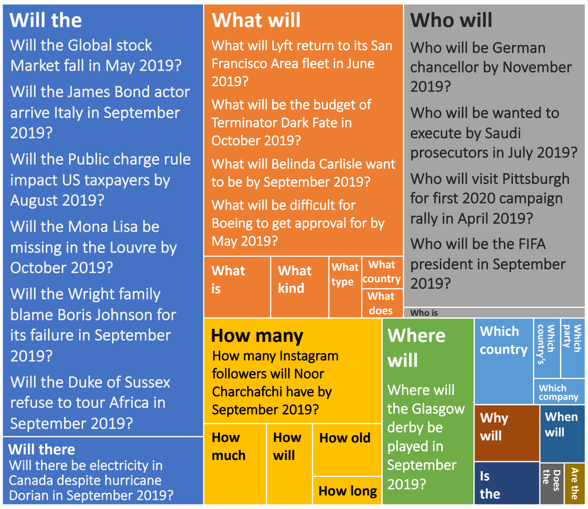


Figure 2: **A treemap visualization of first two words in ForecastQA questions.** Box area is proportional to number of occurrences.

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.