FORECAST2023: A Forecast and Reasoning Corpus of Argumentation Structures

Kamila Górska, John Lawrence, Chris Reed
University of Dundee
{k.gorska, j.lawrence, c.a.reed}@dundee.ac.uk

Abstract

It is known from large-scale crowd experimentation that some people are innately better at analysing complex situations and making justified predictions – the so-called 'superforecasters'. Surprisingly, however, there has to date been no work exploring the role played by the reasoning in those justifications. Bag-of-words analyses might tell us something, but the real value lies in understanding what features of reasoning and argumentation lead to better forecasts – both in providing an objective measure for argument quality, and even more importantly, in providing guidance on how to improve forecasting performance. The work presented here covers the creation of a unique dataset of such prediction rationales, the structure of which naturally lends itself to partially automated annotation which in turn is used as the basis for subsequent manual enhancement that provides a uniquely fine-grained and close characterisation of the structure of argumentation, with potential impact on forecasting domains from intelligence analysis to investment decision-making.

Keywords: forecasting, argumentation, corpus

1. Introduction

The ability to make accurate predictions is vital when it comes to decision-making across many domains including intelligence analysis, healthcare and finance. This broad-ranging importance has led to the general study of *forecasting*, the process of making predictions based on past and present data, and in particular, variance in forecasting ability, most commonly explored through forecasting competitions. Such competitions consist of questions being posed about the outcome of future events, such as "Will the United Kingdom (UK) leave the European Union (EU) before 1 November 2019?". The forecasters taking part in the competition assign probabilities to the possible outcomes, indicating how likely they believe each possible outcome to be, and providing a supporting rationale. Once the outcome of the question is known, the accuracy of the forecasts can then be calculated using a variety of metrics, such as Brier scoring (Brier, 1950).

The rationale the forecasters provide offers insight into their thought process, and previous work has shown correlations between the content of the provided rationale and eventual forecast accuracy. To date, such work has generally focused on surface-level features looking at, for example: linguistic markers, the use of comparison classes, and overall dialectical complexity (Karvetski et al., 2022). Though intuitions may suggest that quality of reasoning is connected to the quality of forecasting, current Large Language Model approaches to understanding reasoning structures remain poor. In this paper we take a first step towards a deeper study of the reasoning contained in forecast rationales, by providing a dataset through which it is possible to explore the argumentative strategies employed in these texts, and paving the way to future study of the correlation between reasoning structure and forecasting ability.

Described here is the production of FORECAST2023: the FOrecast and REasoning Corpus of Argumentative STructures, consisting of 120 argument maps of forecast rationales. The annotation is provided in Argument Interchange Format (AIF) using Inference Anchoring Theory (IAT) to model the argument structure, which was completed through a two-step process; the first step consists of an automatic generation of the question, hypothesised outcomes, and forecast section of the text, followed by a manual annotation process of the rationale and how it relates to the predictions. [1] An Inter-Annotator Agreement (IAA) score of κ = 0.78 was achieved after this manual annotation step, constituting substantial agreement.

2. Related Work

Geopolitical forecasting competitions test individuals on their accuracy in predicting future geopolitical events, the question's topics cover many areas such as health, finance, climate, politics, international relations and technology (Karvetski et al., 2022). Participants are often non-experts in the topics and assign probabilities to potential outcomes. These competitions have been used to

[1] All code is available at: https://github.com/arg-tech/Forecast2023

LREC-COLING 2024, pages 7395–7405

inform researchers across many domains including intelligence work, where analysts must make decisions involving national security (Chang et al., 2016, 2017), psychology involving judgement making (Atanasov et al., 2020; Mellers et al., 2015; Moore et al., 2017), economics and political science exploring predictability (Baron et al., 2014; Friedman et al., 2018; Tetlock, 2017), and the statistics of drawing the wisdom of the crowd (Satopää et al., 2014).

There have been various forecasting competitions such as the Intelligence Advanced Research Projects Activity (IARPA) funded Aggregative Contingent Estimation (ACE) tournament which was won by Good Judgement Project research team investigating what sets high-skilled and low-skilled forecasters apart (Atanasov et al., 2020), and the Hybrid Forecasting Competition (HFC), aiming to improve forecasting accuracy by combining the strengths of human forecasters with computational techniques (Karvetski et al., 2022). Previous work has found that forecasting is a task that some individuals excel at (Katsagounos et al., 2021), even outperforming experts (Tetlock and Gardner, 2016). Work has focused on improving forecasting accuracy, such as through training forecasters (Chang et al., 2016), or combining human forecasting with computational systems (Beger and Ward, 2019; Karvetski et al., 2022; Shinitzky et al., 2023).

A linguistic analysis of rationales shows that the usage of comparison classes and dialectical complexity correlate to forecaster accuracy. (Karvetski et al., 2022). However, a deeper analysis of how forecasters form their arguments in rationales and how that argument structure plays a role in the accuracy of the predictions is yet to be studied.

Identifying argumentation structures aids in the analysis of texts, and has been applied to for example, analyse the structures of scientific publications (Kirschner et al., 2015), or assess the helpfulness of reviews (Liu et al., 2017). Argumentation is a key component in essay writing and argument structure has been used to evaluate the quality and persuasiveness of student essays (Stab and Gurevych, 2014a,b), as well as argumentation schemes (Song et al., 2014). Considering arguments and their validity has also been applied in schools in order to help students identify fake news and increase their critical understanding of the text (Visser et al., 2020b).

Arguments have been analysed in various domains ranging from legal texts (Walker et al., 2014; Weber et al., 2023) and medical data (Fox et al., 2007), to political discussions such as presidential elections (Visser et al., 2020a) and televised debates (Hautli-Janisz et al., 2022). Despite this wide-ranging application, no other work has focused on looking at argument structure in geopolitical forecasting data. Irwin et al. (2022) approach this problem by creating an argumentation framework which supports forecasters while making predictions, by empowering agents to argue over time about the probability of outcomes.

3. Inference Anchoring Theory

Inference Anchoring Theory (IAT) ((Budzynska et al., 2014, 2016) is an Argument Interchange Format (AIF) compatible model, which models argument structures and relations, representing the argumentation which occurs in a discourse. There are five main elements to IAT: (i) units of information more or less corresponding to propositions; (ii) a special subclass of propositions known as locutions that refer specifically to discourse events; (iii) relations of inference, conflict, and rephrase that hold between propositions; (iv) relations of protocol-defined transition that hold between locutions; and (v) relations of illocution that hold between locutions and transitions on the one hand, and propositional contents and relations on the other.

Locutions correspond to argumentative discourse units (ADUs), the minimal unit into which text is segmented (Peldszus and Stede, 2013). For convenience we associate discourse participants with discourse material using a convention of textual styling by which ADU content is preceded by a string of the form, 'firstname lastname: '.

Propositions are derived from locutions and are grammatically complete instantiations of the content of the locution. Ideally, each proposition should be interpretable without further context, which often means the content must be reconstructed to resolve any elliptical and anaphoric expressions and other forms of deixis and underspecification.

Propositional Relations are divided into three different classes used to model argumentative relations between propositions: inference, conflict and rephrase. Transition Relations capture relevance between contributions in dialogue, and are governed by the rules of dialogical context captured by a game or protocol (Wells and Reed, 2012). Illocutionary Relations use a Speech Act Theory foundation to describe the intentional structure in discourse, capturing assertives, question types, argumentative moves and so on.

An entire IAT structure therefore simultaneously captures the surface discourse activity (directly connecting to transcripts, for example), and the intentional structure of dialogical interaction, and the informational structure that is co-created and navigated by the dialogue. One interesting feature of IAT, given its focus on argumentation and debate, is that the propositional relation of inference constitutes the content of the illocutionary force of arguing, which in turn is typically anchored in a transition between locutions that constitute, for example, a

challenge-response pair. The reason we choose IAT as the annotation scheme is twofold: firstly, to distinguish the illocution of hypothesising from other types such as asserting as they interact in reasoning structures in different ways, and secondly, for the ability to distinguish different speaker's views on given conclusions which IAT provides.

4. Data Processing

The HFC data is a publicly available dataset[2] which was published as part of the Hybrid Forecasting Competition (Benjamin et al., 2023). The aim of the HFC was to combine the strengths of both human and machine forecasting, into a hybrid model. To annotate forecast and reasoning argument structures, data which was collected as part of the IARPA *Geopolitical Forecasting Challenge 2*[3] is used. IARPA posted questions about future geopolitical events, such as "Will France's President Emmanuel Macron experience a significant leadership disruption between 3 April 2019 and 29 November 2019?". In total, 537 individuals took part in this competition, answering a subset of the questions, assigning a probability to each possible outcome, and a rationale explaining their reasoning behind it. The participants ("forecasters"), answered questions set in the future, having no way of finding out the correct answer, making their best guess and detailing their reasoning. The forecasters could answer each question as many times as they wished, meaning part of the forecasts have updates to their prediction, which the forecaster has made over time. The full dataset is made up of 629,874 entries, which constitute 75,479 individual rationales.

4.1. Data Filtering

In order to ensure all forecasts which had been made before the outcome of the question was known, any forecasts which were flagged as such, were removed from the data. Any forecasts which had a rationale which was deemed too short (less than 10 words) were also removed, discarding rationales such as 'because I think so' or 'comment removed'.

Each individual row of data is a forecast made by an individual as a response to a question. The forecast evaluates one of the hypotheses provided by the competition, which creates between two and five entries with the same rationale but different scores, one for each hypothesised outcome. In order to restructure the data to have a unique rationale and all of the corresponding forecasts made by the forecaster as a single row of data, the rows are aggregated, grouping the forecasts by the question ID, participant ID, comment ID, time of forecast and rationale.

Consensus scores were calculated for each forecast as part of the HFC, as a measure of standardised accuracy ranging from 0 (best) to 1 (worst). Any rationale was removed from the data if there was not more than one of the forecasts' consensus scores which indicated that an answer was not virtually certain, leaving at least two not certain options. A forecast is almost certain when the forecast is submitted near the end of a question's timeline, which following Karvetski et al. (2022) is defined as a consensus score which is less than or equal to 0.0025^6. This data-cleaning process left 75,115 unique rationales remaining.

In order to aid in the automated annotation algorithm (described in Section 5.2), the forecaster's unique ID, their forecast scores, and which hypothesised outcome the forecasts correspond to, are added to the start of the rationale. Forecasters were encouraged to update their forecasts throughout the period that the question was open, while making the updates forecasters would often refer back to their previous forecast and rationale. In order to not lose any context which is required to understand the updated rationale, the update made by the forecaster is appended to the initial forecast rationale along with any previous updates, resulting in a chain of forecasts and rationale, making the rationale field contain, the forecast scores, followed by a rationale, for each time the forecaster updated their prediction

4.2. Data Sampling

To choose a sample of forecasts to be annotated, forecasts were chosen at random, while ensuring that the sample was not weighed towards a particular question or participant. The proportion of how many rationales answer the same question on average, and how many rationales were written by one forecaster on average, were calculated. When a rationale was chosen at random, the following checks were conducted: (i) the rationale has not already been selected to form the sample, (ii) the question the rationale answers has not already been selected more times or equal to the proportion cut-off, and (iii) the forecaster who's rationale was chosen, has not already been chosen more or equal to the participant cut-off. If any of the checks did not pass, another rationale would be chosen at random and the same checks would be performed until a rationale satisfied the requirements. A sample of 120 forecasts was chosen, which including updates, totals 205 unique rationales.

[2]Available at https://dataverse.harvard.edu/dataverse/hfc

[3]https://www.herox.com/IARPAGFChallenge2

5. Annotation

The following example (Example 1) is an example of a forecast from the corpus:

(1) Part ID: 1999

a. HFC: Will France's President Emmanuel Macron experience a significant leadership disruption between 3 April 2019 and 29 November 2019?

b. HFC: Yes or No

c. Forecaster: Yes, 0.05 probability

d. Forecaster: No, 0.95 probability

e. Forecaster: I believe with the yellow-vest protests put on by the working class who have been hit with low wages and heightened gasoline taxes are a sign of the people's abilities, but I do not believe this is something that will take Macron out of office. Macron has attempted to meet voters demands and is trying to rectify the situation by being more involved and less out of touch. Macron is hosting or involved in political discussions, debates and meetings to address the unrest in his country. I do think he's made plenty of mistakes that caused a lot of damage (albeit some being done by the protestors), but I believe he is regaining 'popularity' enough to maintain office.

In the example, part 1a is the **question** set by the competition, part 1b are the provided possible outcomes, referred to as the **hypotheses**, parts 1c and 1d are the probabilities assigned to each outcome, referred to as the **forecasts**, and part 1e is the **rationale** provided by the forecaster in support of their forecasts.

5.1. Automated Annotation Algorithm

In order to streamline the process, part of the IAT annotation is completed automatically. The maps are formed in an extended version of the Argument Interchange Format (AIF), in order to be used in the annotation tool, OVA[4] (Janier et al., 2014). Table 1 shows an example of the required fields to build the automatically generated portion of the argument map, which is shown in Figure 1.

The 'Unique ID' field of Table 1 forms the Part ID in Example 1, which is used to match the argument map with the data table. The 'Forecaster ID' is used to uniquely identify each speaker, this becomes the speaker name of the forecaster in each analysis. The 'Question' field is used to form the Pure Question asked, the speaker name is set to 'HFC' to represent the competition, and the text of the question excluding the question mark is used to form the content of the text in both the locution and proposition (blue) nodes. A YA (relation of illocution, yellow) 'Pure Questioning' node is added to represent the illocutionary force of questioning, an edge represented on the map by an arrow, is added from the locution to the YA node, and from the YA node to the proposition, as shown in Figure 1. The timestamp provided in the 'Date' field of the table is added to each locution into a 'timestamp' field which is not visible in the visualisation of the map but is available to be accessed in the raw data.

To form the nodes which represent the hypotheses that the forecasters are presented with, the 'Hypotheses' field of Table 1 is used. The list of hypotheses is combined, separated by the word 'or', and is used as the text for the next locution. A TA (transition, purple) node is added from the previous locution to the current locution, along with the required edges. There can be a range of 2 to 5 hypotheses available, a proposition node is created for each one, which is anchored in YA 'Hypothesising', with edges which go from the TA node to each YA node, and from every YA node to one of the propositions. Each hypothesis also performs the function of answering the question, which means that for each proposition an MA (rephrase, orange) 'Default Rephrase' node is added, and a YA 'Default Illocuting' along with the required edges as shown in Figure 1. To model the conflict between each proposition, we iterate through the list of hypotheses, adding CA (conflict, red) nodes and edges from each hypothesis proposition to each other hypothesis proposition, all anchored in a YA 'Alternative Giving'.

The forecaster ID and, and forecast for each hypothesis are already part of the rationale (see Section 4.1 above for details). To add each set of forecast nodes, the list of 'Forecasts' from Table 1 is iterated through to ensure the correct number of forecast nodes are added, the text is split up and a locution and proposition pair are created for each forecast. A TA node is added between the new locution node and the locution node directly preceding it. Each of the forecast proposition nodes is anchored with a YA 'Asserting' node, and an MA 'Evaluation' node is also created which links from the forecast proposition to the corresponding hypothesis proposition above. This 'Evaluation' node is anchored with YA 'Evaluating' from the locution.

For each node which is added to the map, the information stored includes a unique node ID, the node's type and content, whether the node should be visible and displayed on the argument map, x and y coordinates, and a timestamp. For each locu-

[4] Available at ova.arg.tech

Unique ID	Forecaster ID	Question	Date	Rationale	Forecasts	Hypotheses
1999	c70e139	Will France's President Emmanuel Macron experience a significant ...?	2019-04-08 17:35:12	c70e139: Yes, 0.05 probability, c70e139: No, 0.95 probability, c70e139: I believe with the yellow-vest protests ...	0.05,0.95	'Yes', 'No'

Table 1: Example of processed data, (as described in Section 4.1), which is used to generate the automated annotation (described in Section 5.2)

tion additionally, a speaker is stored with a unique ID, and a span tag is added to the text stored in the file, surrounding the text which makes up the content of the locution. The span tag's ID matches the locution node's unique ID, in order to link each locution to the original text the content came from.

5.2. Automatically Generated Annotation Example

To annotate the data, IAT is used as it allows for a fine-grained analysis of the text, including illocutionary connections, which allow the distinction between, for example, the act of asserting and the act of hypothesising. Given that the forecasts follow a set structure, the IAT analysis of the first part of the text which includes the question, hypotheses and forecasts is completed automatically, without a human annotator, which results in the partial analysis shown in Figure 1. The automatic process allows for the annotation to be completed faster, allowing the annotators to complete more analysis in the same time frame.

The analysis begins with the question *'HFC: Will France's President Emmanuel Macron experience a significant leadership disruption between 3 April 2019 and 29 November 2019?'* (Example 1a), which is annotated as a 'Pure Question' asked by the HFC. The question is followed by *'HFC: Yes or No'* (Example 1b), where the HFC lists each hypothesis for the forecaster to consider, outlining what the possible outcomes of the event in question are. This is captured through a locution on the right which creates two propositions on the left, *Yes* and *No*, both of which are anchored with the illocutionary relation of 'Hypothesising'. Both of these propositions are anchored in the Default Transition and not in the locution, as the content of the locution above is required in order to reconstruct and understand what 'Yes' and 'No' refer to.

The intention of 'Yes' or 'No' to be possible answers to the question is captured through the 'Default Rephrase' relation between the proposition 'Yes' (or 'No') and the proposition of the question. The 'Default Rephrase' relation is anchored with the 'Default Illocuting' relation as set out in IAT guidelines. Each hypothesis is an alternative to each other hypothesis as only one can come true, which is captured through the 'Default Conflict' relation, anchored with the 'Alternative Giving' illocutionary relation.

Example 1 parts 1c and 1d are forecasts, evaluating each hypothesis. They are captured as a locution and proposition pair, in which the forecaster is the speaker and the content which the forecaster asserts is the probability assigned for each hypothesised outcome. The evaluation of each hypothesis through the forecast is captured with an MA 'Evaluation' propositional relation, to the proposition of the hypothesis which is being evaluated, this is anchored from the forecast locution with 'Evaluating'. Capturing the intention of the turn as asserting a forecast, which evaluates the probability of each hypothesis as an answer to the question. A simplified view is shown in Figure 2.

5.3. Manual Annotation Process

The annotation process of FORECAST2023 follows approximately the manual annotation process described in (Lawrence and Reed, 2020). The analysts segment the text into ADUs, which form locutions and propositions, in the same step the propositions are reconstructed by the analyst. The analyst also reconstructs any of the existing propositional nodes which were created through the automatic generation process. The analyst identifies whether a node has an argumentative function or not, and adds the propositional and illocutionary relations. Once one analyst completes a first pass of the annotation, it is peer-reviewed by a second analyst, who checks for any mistakes and discusses annotation choices made and any complex elements. The final version of the annotation is completed by the original analyst to reflect any changes which came out of the discussion, which aims to improve the quality of the annotation.

The annotation was carried out by 11 analysts. All of the analysts completed an IAT training course which included a test at the end to assess the qual-

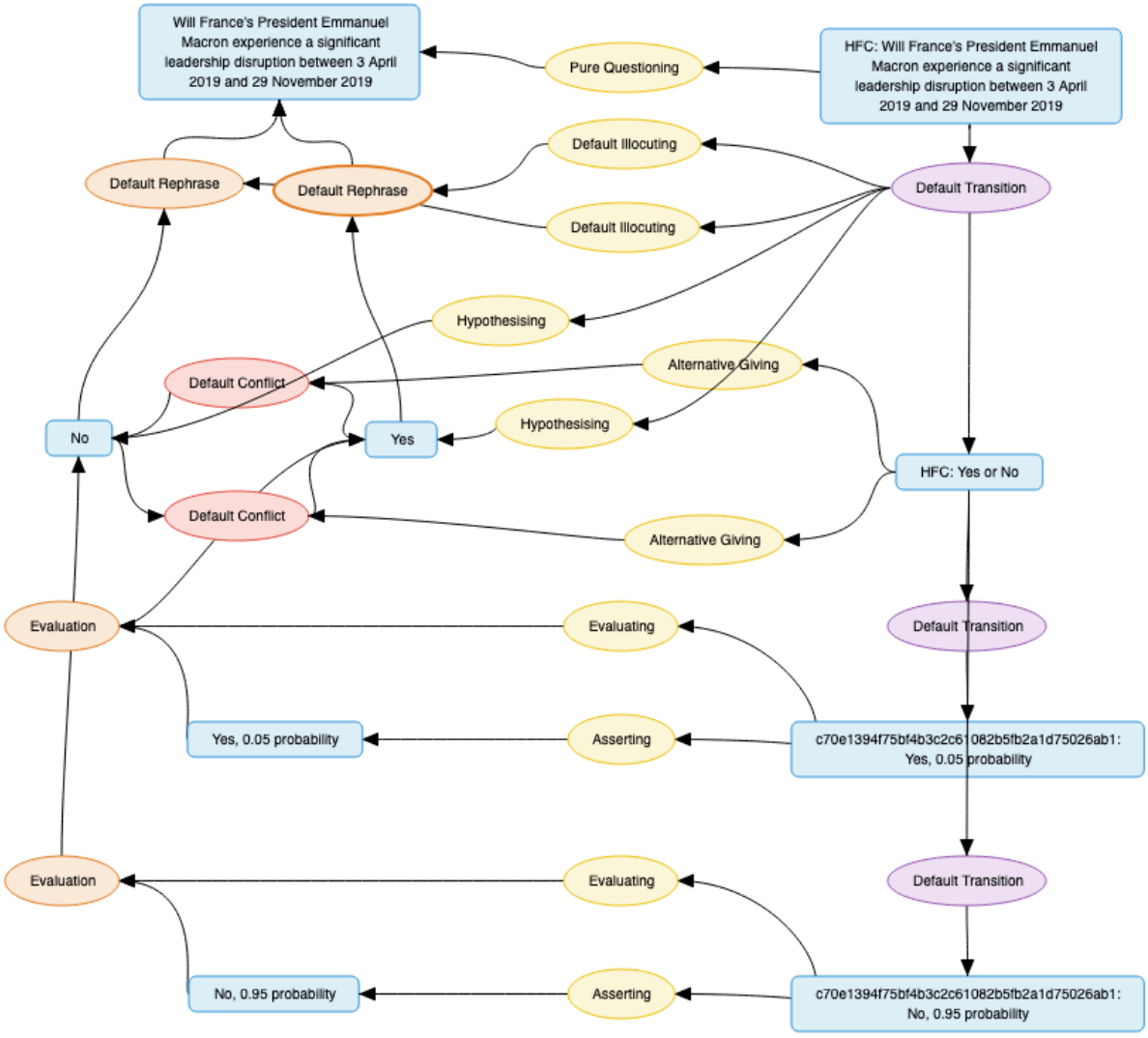

Figure 1: Autmatically-generated IAT Analysis of Example 1, showing the complete structure including propositions, locutions, propositional relations and illocutionary relations.

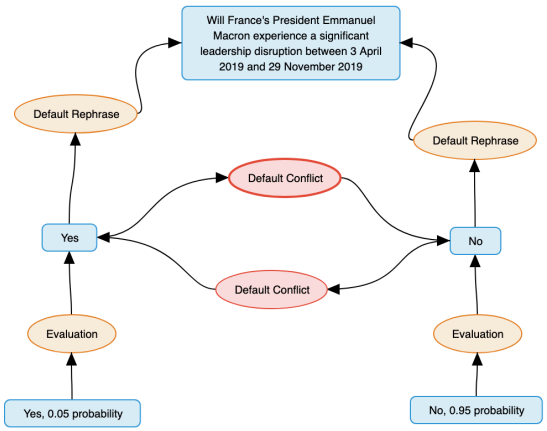

Figure 2: Analysis of Example 1 showing a simplified view of the propositions and propositional relations, not displaying the locutions and illocutionary relations, which are present in corpus

ity of annotation, and also attended a briefing session on how to annotate aspects unique to forecasts. The analysts who completed the annotation task have between six months and five years of experience in IAT annotation. The IAT annotation guidelines[5] and further forecasting-specific guidelines are available online[6].

5.4. Manually Annotated Forecast Example

The annotator is provided with the partially annotated map to complete the analysis as shown in Figure 2. Once the annotation steps have been completed, the resulting annotation is shown in Fig-

[5]Available at: https://www.arg.tech/f/IAT_guidelines_and_tutorials-2023-10.pdf

[6]Available At: https://www.arg.tech/~kamila/ForecastingAnnotationGuidelines.pdf

ure 3. The propositional content of the question (Example 1a) is reconstructed, as is the propositional content of the hypotheses (Example 1b), and the forecasts (Examples 1c and 1d), as each proposition should be interpretable without context. The remaining rationale (Example 1e) is split into ADUs and annotated according to IAT guidelines. The first unit "the yellow-vest protests put on by the working class who have been hit with low wages and heightened gasoline taxes are a sign of the people's abilities" is captured as a locution and proposition pair. This unit creates a conflict relation with the forecasts, as the forecaster is providing a counterargument, as to why President Macron may experience a disruption, which they have predicted is unlikely.

Next, "I do not believe this is something that will take Macron out of office" is reconstructed on the left to "the protests are not something that will take Macron out of office". This assertion supports all of the forecasts, both that there is a 5% probability that President Macron will experience disruption, and that there is a 95% probability that President Macron will not experience disruption, as both forecasts reflect the view that a disruption is very unlikely. This proposition is the conclusion to a serial argument, in which the forecaster supports this statement through a chain of reasoning, shown in Figure 3 through the 'Default Inference' relations.

The forecaster provides another counterargument when asserting "Macron has made plenty of mistakes that caused a lot of damage", which they then provide a rebuttal to, "Macron is regaining 'popularity' enough to maintain office". The disagreements are captured through 'Default Conflict' relations.

6. The FORECAST2023 Dataset

In total, FORECAST2023 consists of 205 unique rationales. The rationales span 100 questions, each having between 2 and 5 possible outcomes which were evaluated. Out of the 205 rationales, 85 were updates to previous forecasts making up 41.5% of the corpus.

Each of the responses to the questions (counting an update together with the original rationale) has 220 words on average, across 120 maps in total. The quality and length of the writing do not differ drastically between the rationales, however, the rationales do differ in the objectivity of the forecaster and focus on arguing for one hypothesis as opposed to having arguments to support or attack each possible option.

Some of the forecasters tend to focus on past data and the evidence available at the time to draw their conclusions, such as *"The index has been decreasing for several months. It seems to be heading toward 11.0 and possibly lower. The last few months or so it has been hovering around 12.7. It is likely to continue average around 12.5."* The forecaster is not referencing personal opinions in contrast to forecasters who are less objective and use anecdotal evidence. Another strategy forecasters take is including much more personal opinion on the facts available, using phrases like "I feel" and "I believe", such as *"With Russia already having a love of keeping a tight grip on its citizens I feel there is no doubt in my mind that this bill will pass. With the newest Christchurch shooting have been live streamed I feel like that may influence the decision to try and police the internet ever further. Despite the protests support of the bill by the large money backers grows."*

	In total	**%**
Propositional Relations		
Default Inference	968	31%
Default Conflict	946	31%
Default Rephrase	1170	38%
Total	3084	100%
Illocutionary Relations		
Asserting	1760	33%
Arguing	964	18%
Agreeing	2	0%
Alternative Giving	872	16%
Challenging	1	0%
Disagreeing	73	1%
Evaluating	629	12%
Hypothesising	353	7%
Restating	189	4%
Pure Questioning	121	2%
Default Illocuting	353	7%
Total	5317	100%

Table 2: Distribution of Propositional and Illocutionary relations in FORECAST2023, showing the total number of occurrences for each type of relation and the percentage of the total relations each type makes up

Table 2 outlines the proportion of the dataset in terms of propositional and illocutionary relations. In total, the corpus has 3084 propositional relations, which are rather evenly split between support, conflict and rephrase relations. However, the conflict and rephrase relations mostly pertain to the question, hypotheses and forecasts, as opposed to the rationale and how the rationale supports the forecasts made. The corpus has 5317 illocutionary relations in total, the most common being asserting which makes up 33% of the illocutionary relations. The next most common is arguing, followed by alternative giving and asserting.

Table 3 shows the distribution of propositional and illocutionary relations in only the rationale

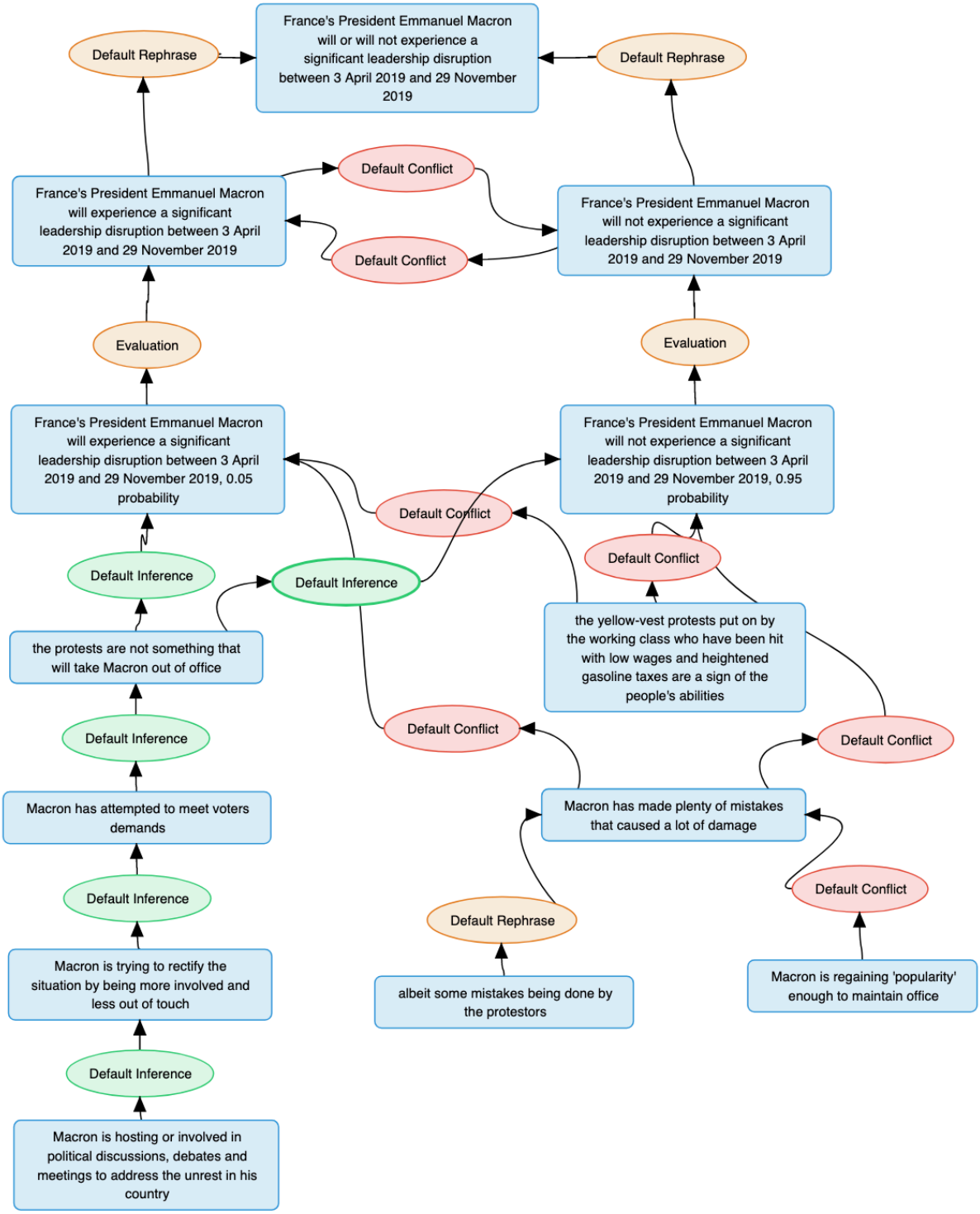

Figure 3: Annotated example after being completed by an annotator, only the propositions and propositional relations are shown in the figure for simplicity, however the full argument map including locutions and illocutionary relations is available as part of the corpus

and relations between rationale and forecasts, not counting the rephrase or conflict relations between the question, hypotheses or rationales themselves. The majority of propositional relations are default inference (61%), with forecasters providing more support for their opinion than rephrasing or providing counterarguments. The majority of the illocutionary relations are made up of asserting (59%), followed by arguing (32%), indicating that forecasters tend towards giving support as opposed to evidence against a hypothesis.

	In total	%
Propositional Relations		
Default Inference	968	61%
Default Conflict	74	5%
Default Rephrase	541	34%
Total	1583	100%
Illocutionary Relations		
Asserting	1760	59%
Arguing	964	32%
Agreeing	2	0%
Challenging	1	0%
Disagreeing	73	2%
Restating	189	6%
Pure Questioning	1	0%
Total	2990	100%

Table 3: Distribution of Propositional and Illocutionary relations in only the rationale, in FORECAST2023

6.1. Inter-Annotator Agreement

The Inter-Annotator agreement is calculated as CASS-κ = 0.78, capturing agreement between annotators in terms of segmentation, argumentation structures and illocutionary forces, and combining it into a single score, using the Combined Argument Similarity Score (CASS) (Duthie et al., 2016). CASS calculates a segmentation score using three techniques: the P_k statistic (Beeferman et al., 1999), the WindowDiff statistic (Pevzner and Hearst, 2002), and the segmentation similarity statistic (Fournier and Inkpen, 2012). Propositional relations and dialogical relations are calculated individually, using Cohen's κ, taking into account that segmentation may not match, but not penalising on this basis, instead matching different segmentation using Levenshtein distance (Levenshtein et al., 1966). The segmentation, propositional and dialogical scores are incorporated, using the CASS technique, into a single score.

At random, 15 argument maps in the corpus were chosen (a 12.5% sample) to be re-annotated by a different annotator. The annotation process for the maps chosen for calculating agreement between annotators followed the same process as the original annotation, including being reviewed by a second analyst, who has not annotated or reviewed that part originally. The result of CASS-κ = 0.78 constitutes a substantial agreement between annotators in carrying out the annotation task.

7. Conclusion

This work facilitates the study of argumentation structures in forecasting rationales by presenting a corpus which consists of 100 questions about the outcomes of future geopolitical events, answered by 205 unique rationales which provide arguments for and against predictions made by forecasters, making up 120 argument maps. The fine-grained analysis carried out in Inference Anchoring Theory (IAT), includes relations of inference, conflict and rephrase that hold between propositions, but also illocutionary relations describing the intention structure. While the study of forecasters has identified 'superforecasters' who excel at the task, exactly how they structure their reasoning and the argumentative strategies they employ are unknown. The argument structure extracted from the text provides a novel perspective to begin studying the connections between argumentative strategies and forecasting accuracy, based on the way in which forecasters reason. Knowing the ways in which the best forecasters structure their reasoning opens the door to building tools and training to aid in accurate prediction-making across forecasting competitions but also more widely, in understanding financial markets, performing intelligence analysis, and responding to disease outbreaks.

8. Ethics

All annotators were paid above minimum wage, and also above the UK's national living wage. The corpus is released under CC-BY-SA.

9. Acknowledgements

This research is supported in part: by the 'AI for Citizen Intelligence Coaching against Disinformation (TITAN)' project, funded by the EU Horizon 2020 research and innovation programme under grant agreement 101070658; by UK Research and innovation under the UK governments Horizon funding guarantee grant numbers 10040483 and 10055990; and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

10. Bibliography

Pavel Atanasov, Jens Witkowski, Lyle Ungar, Barbara Mellers, and Philip Tetlock. 2020. Small steps to accuracy: Incremental belief updaters are better forecasters. In *Proceedings of the 21st*

ACM Conference on Economics and Computation, pages 873–874.

Jonathan Baron, Barbara A Mellers, Philip E Tetlock, Eric Stone, and Lyle H Ungar. 2014. Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 11(2):133–145.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine learning*, 34:177–210.

Andreas Beger and Michael D Ward. 2019. Assessing amazon turker and automated machine forecasts in the hybrid forecasting competition. In *7th Annual Asian Political Methodology Conference*, pages 5–6.

Daniel M Benjamin, Fred Morstatter, Ali E Abbas, Andres Abeliuk, Pavel Atanasov, Stephen Bennett, Andreas Beger, Saurabh Birari, David V Budescu, Michele Catasta, et al. 2023. Hybrid forecasting of geopolitical events. *AI Magazine*.

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

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. Towards argument mining from dialogue. In *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014)*, pages 185–196. IOS Press.

Katarzyna Budzynska, Mathilde Janier, Chris Reed, and Patrick Saint Dizier. 2016. Theoretical foundations for illocutionary structure parsing. *Argument & Computation*, 7(1):91–108.

Welton Chang, Pavel Atanasov, Shefali Patil, Barbara A Mellers, and Philip E Tetlock. 2017. Accountability and adaptive performance under uncertainty: A long-term view. *Judgment and Decision Making*, 12(6):610–626.

Welton Chang, Eva Chen, Barbara Mellers, and Philip Tetlock. 2016. Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision making*, 11(5):509–526.

Rory Duthie, John Lawrence, Katarzyna Budzynska, and Chris Reed. 2016. The cass technique for evaluating the performance of argument mining. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 40–49.

Chris Fournier and Diana Inkpen. 2012. Segmentation similarity and agreement. *arXiv preprint arXiv:1204.2847*.

John Fox, David Glasspool, Dan Grecu, Sanjay Modgil, Matthew South, and Vivek Patkar. 2007. Argumentation-based inference and decision making–a medical perspective. *IEEE Intelligent Systems*, 22(6):34–41.

Jeffrey A Friedman, Joshua D Baker, Barbara A Mellers, Philip E Tetlock, and Richard Zeckhauser. 2018. The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly*, 62(2):410–422.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. Qt30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 3291–3300. European Language Resources Association (ELRA).

Benjamin Irwin, Antonio Rago, and Francesca Toni. 2022. Forecasting argumentation frameworks. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 19, pages 533–543.

M. Janier, J. Lawrence, and C Reed. 2014. OVA+: An argument analysis interface. In *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014)*, pages 463–464, Pitlochry. IOS Press.

Christopher W Karvetski, Carolyn Meinel, Daniel T Maxwell, Yunzi Lu, Barbara A Mellers, and Philip E Tetlock. 2022. What do forecasting rationales reveal about thinking patterns of top geopolitical forecasters? *International Journal of Forecasting*, 38(2):688–704.

Ilias Katsagounos, Dimitrios D Thomakos, Konstantia Litsiou, and Konstantinos Nikolopoulos. 2021. Superforecasting reality check: Evidence from a small pool of experts and expedited identification. *European journal of operational research*, 289(1):107–117.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Haijing Liu, Yang Gao, Pin Lv, Mengxue Li, Shiqiang Geng, Minglan Li, and Hao Wang. 2017. Using argument-based features to predict and analyse review helpfulness. *arXiv preprint arXiv:1707.07279*.

Barbara Mellers, Eric Stone, Pavel Atanasov, Nick Rohrbaugh, S Emlen Metz, Lyle Ungar, Michael M Bishop, Michael Horowitz, Ed Merkle, and Philip Tetlock. 2015. The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of experimental psychology: applied*, 21(1):1.

Don A Moore, Samuel A Swift, Angela Minster, Barbara Mellers, Lyle Ungar, Philip Tetlock, Heather HJ Yang, and Elizabeth R Tenney. 2017. Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, 63(11):3552–3565.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Lev Pevzner and Marti A Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Ville A Satopää, Jonathan Baron, Dean P Foster, Barbara A Mellers, Philip E Tetlock, and Lyle H Ungar. 2014. Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2):344–356.

Hilla Shinitzky, Yhonatan Shemesh, David Leiser, and Michael Gilead. 2023. Improving geopolitical forecasts with 100 brains and one computer. *International Journal of Forecasting*.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the first workshop on argumentation mining*, pages 69–78.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 46–56.

Philip E Tetlock. 2017. Expert political judgment. In *Expert Political Judgment*. Princeton University Press.

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

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020a. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.

Jacky Visser, John Lawrence, and Chris Reed. 2020b. Reason-checking fake news. *Communications of the ACM*, 63(11):38–40.

Vern Walker, Karina Vazirova, and Cass Sanford. 2014. Annotating patterns of reasoning about medical theories of causation in vaccine cases: Toward a type system for arguments. In *Proceedings of the first workshop on argumentation mining*, pages 1–10.

Florian Weber, Thiemo Wambsganss, Seyed Parsa Neshaei, and Matthias Soellner. 2023. Structured persuasive writing support in legal education: A model and tool for german legal case solutions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2296–2313.

Simon Wells and Chris A Reed. 2012. A domain specific language for describing diverse systems of dialogue. *Journal of Applied Logic*, 10(4):309–329.