Mining International Political Norms from the GDELT Database

Extended Abstract

Rohit Murali
Indian Institute of Science
Bangalore, India
muralirohit@gmail.com

Suravi Patnaik*
Sponsa Limited
Wellington, New Zealand
suravi@sponsa.co

Stephen Cranefield
University of Otago
Dunedin, New Zealand
stephen.cranefield@otago.ac.nz

KEYWORDS

Norm Mining; Bayesian Inference; GDELT; Big Data

ACM Reference Format:
Rohit Murali, Suravi Patnaik, and Stephen Cranefield. 2020. Mining International Political Norms from the GDELT Database. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

1 INTRODUCTION

Researchers have long been interested in the role that norms can play in governing agent actions in multi-agent systems. Norms have been shown to facilitate social order [2] and improve cooperation and coordination among agents [9], and an active research community has investigated many theoretical and practical aspects of normative reasoning in multi-agent systems [1]. Much of this work has focused on formalising normative concepts from human society and adapting them for the government of open software systems, and on the simulation of normative processes in human and artificial societies. However, there has been comparatively little work on applying normative MAS mechanisms to understanding the norms in human society.

This work investigates this issue in the context of international politics. Using the GDELT dataset, containing machine-encoded records of international events extracted from news reports, we extracted bilateral sequences of inter-country events and applied a Bayesian norm mining mechanism to identify norms that best explained the observed behaviour. A statistical evaluation showed that a probabilistic model enhanced with explicit normative reasoning fitted the data significantly better than a baseline probabilistic model.

The Global Database of Events, Language and Tone (GDELT) [6] is a continuously updated geopolitical event database with over half a billion records. It is free and open, and the current version, GDELT 2.0, is updated every 15 minutes. The database includes an events table containing machine-coded data extracted from news reports, with 60 attributes for each event (e.g. the event type and countries involved) [5], and has been used for studies such as predicting future violence levels in Afghanistan [11] and detecting protest events in the world [8].

We considered the following research question: *Can the GDELT data be better explained by a model combining probabilities and norms than by a purely probabilistic model?* Our research comprised five stages: i) collecting and pre-processing data from the GDELT database; ii) fitting a probabilistic model to serve as the baseline model; iii) defining a model to calculate the likelihood of observed sequences of events, given an assumed norm; iv) using this model to mine norms from the GDELT database based on Bayesian learning; and v) comparing the fit of the two models with the data.

*Suravi Patnaik contributed to this work while at the University of Otago.

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

2 DATA PRE-PROCESSING

The types of events and actors in the database are encoded using the Conflict and Mediation Event Observations (CAMEO) coding scheme [4]. CAMEO event codes have a three-level hierarchical structure, with 20 "root codes" that are further specialised into "base codes", with some of these subdivided into a further level of detail. We used root codes only. For example, a news report that "Albania on Friday denounced as an ugly crime Yugoslavia's suppression of ethnic Albanian unrest" [4] is encoded with root event code 11 (*Disapprove*). Other example root codes are 3 (*Express intent to cooperate*), 4 (*Consult*) and 5 (*Engage in diplomatic cooperation*).

We considered events spanning one year, from 19 June 2018 to 20 June 2019, recorded in 35,039 data files. We considered events involving two "country" actors (as opposed to other international organisations) with a primary role code of "government". We extracted bilateral sequences of mutually relevant events, based on the transitive closure of a *co-mention* relation that holds when two events appear in the same news report. As we aimed to learn generic norms that apply to all countries, we did not retain the events' source and target countries when learning norms. However, the relative directions of events in a bilateral event sequence are significant. We therefore represented each event as a combination of a *direction* and an event root code, where by convention the first event in an event sequence is taken to be in the "forwards" direction. We denote a directed event by a pair $\langle direction, code \rangle$.

3 LEARNING NORMS FROM EVENT SEQUENCES

We used the following language to define our norm hypothesis space, where *ec* and *cec* range over root event codes.

- $O(ec)$: an unconditional obligation to perform an event with event code *ec*.
- $O(cec, ec, rel\_dir)$: a conditional obligation to perform an event with event code *ec* if a prior condition event with event code *cec* has occurred, and the relative direction of these two events is specified by *rel_dir* (either 'same' or 'different', denoted + and −, respectively).
- $P(ec)$: an unconditional prohibition of events with event code *ec*.

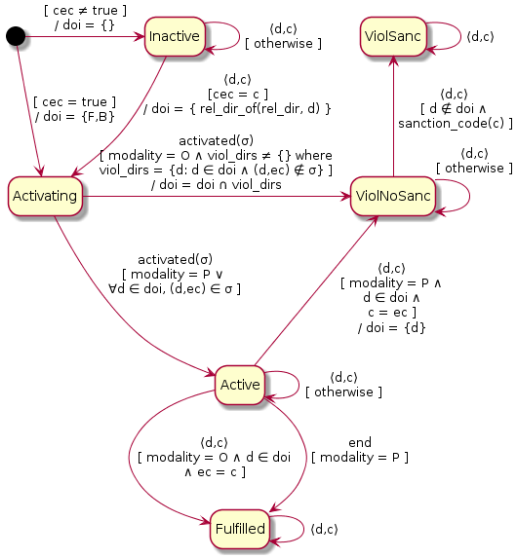

Figure 1: The norm state machine

- $P(cec, ec, rel\_dir)$: a conditional prohibition of events with event code *ec* if a prior condition event with event code *cec* has occurred, and the relative direction of the two events is specified by *rel_dir*.

We followed a Bayesian approach to norm learning [3]. Given a set of norm hypotheses, for every observed event sequence σ in our dataset and each norm hypothesis h we calculated the likelihood of the observation given the hypothesis: $p(\sigma|h)$. Bayes' Rule was then used to update the log odds of a norm hypothesis h versus the null hypothesis (h_0) that there is no norm, given the observation. Let $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_N)$ be our dataset with $N = 513,906$ event sequences. Then, as each sequence is independently observed, we can express the log odds of the hypotheses given the dataset as the sum of the prior log odds and the log likelihood ratios of each sequence under the two hypotheses:

$$log(O(h{:}\, h_0|\sigma)) = log(O(h{:}\, h_0)) + \sum_{i=1}^{N} (log(p(\sigma_i|h)) - log(p(\sigma_i|h_0)))$$

To define the likelihood of an event sequence given the no-norm null hypothesis, we fitted a probabilistic language model to the set of event sequences resulting from our data pre-processing. The trained model gave us the probability of an event given an event history. We chose to use the libPLUMP implementation[1] of the *sequence memoizer* (SM) [10], due to Murphy's description of the SM as "the best-performing language model" [7, p.595].

To define the likelihood of an observation, we defined a state machine for norms (Figure 1), which was instantiated for each observed event sequence and norm hypothesis. We used this to define the likelihood of an event sequence given a (non-null) norm hypotheses using a set of state-specific recursive equations defining $p(\sigma|s, \eta)$ for sequence σ, norm state s and event history η). Through recursive application of the equations, each event (in turn) is fed to the state machine. The norm state can change at each of these steps. Our norm language does not model obligation deadlines. Therefore, in the `Activating` state (once a norm's condition is true), the tail of the sequence is passed to the state machine to be checked for future obligation violation. This allows the likelihood of sanction events to be calculated correctly. The event likelihoods defined by our equations use normative reasoning to modify the base probabilities returned from the SM, e.g. an active obligation makes compliant events more likely, so in this case we compute event probabilities using a norm-specific SM trained on compliant sequences only.

[1] http://www.gatsby.ucl.ac.uk/~ucabjga/libplump.html

Table 1: Top six norms

	log odds	Prob. compliance	Prob. sanction
$O(4)$	75355.1	0.225	0.011
$O(4, 4, -)$	65428.0	0.497	0.019
$O(3)$	21550.6	0.032	0.004
$O(3, 3, -)$	14569.4	0.288	0.025
$O(5, 5, -)$	13814.1	0.189	0.032
$O(4, 3, +)$	12092.1	0.020	0.009

4 RESULTS

Applying the Bayesian inference procedure to our event sequence dataset and hypothesis set resulted in 173 norms with posterior log odds greater than 0 (i.e. odds greater than 1), and which were therefore found to be more likely hypotheses than the null hypothesis that there are no norms. Among these 173 norms, there were 154 conditional obligation norms, 16 conditional prohibition norms and 3 unconditional obligation norms. The top six norms are shown in Table 1, along with their posterior log odds, and the inferred probability of compliance and sanctioning. We provide a description for the top two norms:

- $O(4)$ **:** CAMEO root code 4 stands for 'Consult'. This unconditional obligation to consult the other country is complied with 22.5% of the time, which accounts for its high log odds. Violations of this norm were rarely sanctioned.
- $O(4, 4, -)$ **:** This obligation is triggered by a consultation, and obliges the other party to consult. More than 50% of the time it is violated, but some violations (1.8%) are sanctioned. A 49.7% fulfilment rate and the sanctions account for this norm's very high log odds.

We used a likelihood-ratio test to compare our norm-aware probabilistic model (given the top norm above) with the purely probabilistic model embodied by the SM. We used the posterior distribution of sequences, i.e. the trained SM, to generate synthetic datasets under the null hypothesis. This allowed us to generate an empirical distribution of the likelihood ratio test statistic LRT, and calculate a p value for the observed value LRT(D), given our set of observed event sequences, D. We found that $LRT(D)$ exceeds $LRT(d)$ for all synthetic datasets d, and the p-value was less than 0.0173. Thus we could reject the null hypothesis at the 0.1 significance level and answer our research question in the affirmative: our model including the top norm explains the GDELT dataset better than when using the sequence memoizer alone.

REFERENCES

[1] Giulia Andrighetto, Guido Governatori, Pablo Noriega, and Leendert W. N. van der Torre (Eds.). 2013. *Normative Multi-Agent Systems.* Dagstuhl Follow-Ups, Vol. 4. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

[2] Rosaria Conte and Chrysanthos Dellarocas (Eds.). 2001. *Social order in multiagent systems.* Kluwer Academic Publishers.

[3] Stephen Cranefield, Felipe Meneguzzi, Nir Oren, and Bastin Tony Roy Savarimuthu. 2016. A Bayesian Approach to Norm Identification. In *22nd European Conference on Artificial Intelligence (Frontiers in Artificial Intelligence and Applications)*, Vol. 285. IOS Press, 622–629. https://doi.org/10.3233/978-1-61499-672-9-622

[4] Event Data Project 2012. *CAMEO: Conflict and Mediation Event Observations Event and Actor Codebook.* http://eventdata.parusanalytics.com/cameo.dir/CAMEO.Manual.1.1b3.pdf

[5] GDELT Project 2015. *The GDELT Event Database: Data Format Codebook v2.0.* http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf

[6] Kalev Leetaru and Philip A. Schrodt. 2013. GDELT: Global data on events, location, and tone, 1979–2012. In *Proceedings of the International Studies Association Annual Convention.* http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf

[7] Kevin P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective.* The MIT Press.

[8] Fengcai Qiao, Pei Li, Jingsheng Deng, Zhaoyun Ding, and Hui Wang. 2015. Graph-Based Method for Detecting Occupy Protest Events Using GDELT Dataset. In *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC).* IEEE, 164–168.

[9] Yoav Shoham and Moshe Tennenholtz. 1997. On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence* 94, 1 (1997), 139–166.

[10] Frank Wood, Jan Gasthaus, Cédric Archambeau, Lancelot James, and Yee Whye Teh. 2011. The Sequence Memoizer. *Commun. ACM* 54, 2 (2011), 91–98.

[11] James E Yonamine. 2013. *A nuanced study of political conflict using the Global Datasets of Events Location and Tone (GDELT) dataset.* Ph.D. Dissertation. The Pennsylvania State University.