asecondmouse

Reflections on social science, politics and education

Seven current challenges in event data

Posted on March 13, 2019

This is the promised follow-up to last week's opus, "Stuff I Tell People About Event Data", herein referenced as SITPAED. It is motivated by four concerns:

- As I have noted on multiple occasions, the odd thing about event data is that it never really takes off, but neither does it ever really go away
- As noted in SITPAED, we presently seem to be languishing with a couple "good enough" approaches—ICEWS on the data side and PETRARCH-2 on the open-source coder side—and not pushing forward, nor is there any apparent interest in doing so
- To further refine the temporal and spatial coverage of instability forecasting models (IFMs)—where there are substantial current developments—we need to deal with near-real-time news input. This may not look exactly like event data, but it is hard to imagine it won't look fairly similar, and confront most of the same issues of near-real-time automation, duplicate resolution, source quality and so forth
- Major technological changes have occurred in recent years but, at least in the open source domain, coding software lags well behind these, and as far as I know, coder development has stopped even in the proprietary domain

I will grant that in current US political circumstances—things are much more positive in Europe—"good enough" may be the best we can hope for, but just as the "IFM winter" of the 2000s saw the maturation of projects which would fuel the current proliferation of IFMs, perhaps this is the point to redouble efforts precisely because so little is going on.

Hey, a guy can dream.

Two years ago I provided something of a road-map for next steps in terms of some open conjectures and additional reflections can be found here and here. This essay is going to be more directed, with an explicit research agenda, along the lines of the proposal for a $5M research program at the conclusion of this entry from four years ago. [1] These involve quite a variety of levels of effort—some could be done as part of a dissertation, or even an ambitious M.A. thesis, others would require a team with substantial funding—but I think all are quite practical. I'll start with seven in detail, then briefly discuss seven more.

1. Produce a fully-functional, well-tested, open-source coder based on universal dependency parsing

As I noted in SITPAED, PETRARCH-2 (PETR-2)—the most recent open source coder in active use, deployed recently to produce three major data sets—was in fact only intended as a prototype. As I also noted in SITPAED, universal dependency parsing provides most of the information required for event data coding in an easily processed form, and as a bonus is by design multi-lingual, so for example, in the proof-of-concept **mudflat** coder, Python code sufficient for most of the functionality required for event coding is about 10% the length of comparable earlier code processing a constituency parse or just doing an internal sparse parse. So, one would think, we've got a nice opportunity here, eh?

Yes, one would think, and for a while it appeared this would be provided by the open-source "UniversalPetrarch" (UP) coder developed over the past four years under NSF funding. Alas, it now looks like UP won't go beyond the prototype/proof-of-concept stage due to an assortment of "made sense at the time"—and frankly, quite a few "what the hell were they thinking???"—decisions, and, critically, severe understaffing. [2] With funding exhausted, the project winding down, and UP's sole beleaguered programmer mercifully reassigned to less Sisyphean tasks, the project has 31 open—that is, unresolved—issues on GitHub, nine of these designated "critical."

UP works for a couple of proofs-of-concepts—the coder as debugged in English will, with appropriate if very finely tuned dictionaries, also code in Arabic, no small feat—but as far as I am following the code, the program essentially extracts from the dependency parse the information found in a constituency parse, this approach consistent with UP using older PETR-1 and PETR-2 dictionaries and being based on the PETR-2 source code. It sort of works, and is of course the classical Pólya method of converting a new problem to something you've already solved, [8] but seems to be going backwards. Furthermore PETR-1/-2 constituency-parse-based dictionaries [10] are all that UP has to work with: no dictionaries based on dependency parses were developed in the project. Because obviously the problem of writing a new event coder was going to be trivial to solve.

Thus putting us essentially back to square one, except that NSF presumably now feels under no obligation to pour additional money down what appears to be a hopeless rathole. [11] So it's more like square zero.

Well, there's an opportunity here, eh? And soon: there is no guarantee either the ICEWS or UT/D-Phoenix near-real-time data sets will continue!!

2. Learn dictionaries and/or classifiers from the millions of existing, if crappy, text-event pairs

But the solution to that opportunity might look completely different from any existing coder, being based on machine-learning classifiers—for example some sort of largely unsupervised indicator extraction based on the texts alone, without an intervening ontology (I've seen several experiments along these lines, as well as doing a couple myself)—rather than dictionaries. Or maybe it will still be based on dictionaries. Or maybe it will be a hybrid, for example doing actor assignment from dictionaries—there are an assortment of large open-access actor dictionaries available, both from the PETRARCH coders and ICEWS, and these should be relatively easy to update —and event assignment (or, for PLOVER, event, mode, and context assignment) from classifiers. Let a thousand—actually, I'd be happy with one or ideally at least two—flowers bloom.

But unless someone has a lot of time [12]—no…—or a whole lot of money—also no…—this new approach will require largely automated extraction of phrases or training cases from existing data: the old style of human development won't scale to contemporary requirements.

On the very positive side, compared to when these efforts started three decades ago, we now have millions of coded cases, particularly for projects such as TERRIER and Cline-Phoenix (or for anyone with access to the LDC

Gigaword corpus and the various open-source coding programs) which have both the source texts and corresponding events. [13] Existing coding, however, is very noisy—if it wasn't, there would be no need for a new coder—so the challenge is extracting meaningful information (dictionaries, training cases, or both) for a new system, either in a fully-automated or largely automated fashion. I don't have any suggestions for how to do this—or I would have done it already—but I think the problem is sufficiently well defined as to be solvable.

3. ABC: Anything but CAMEO

As I pointed out in detail in SITPAED, and which is further elaborated in the PLOVER manual and various earlier entries in this blog, despite being used by all current event data sets, CAMEO was never intended as a general-purpose event ontology! I have a bias towards replacing it with PLOVER—presumably with some additional refinements—and in particular I think PLOVER's proposed *event-mode-context* format is a huge improvement, both from a coding, interpretation, and analytical perspective, over the hierarchical format embedded in earlier schemes, starting with WEIS but maintained, for example, in BCOW as well as CAMEO.

But, alas, zero progress on this, despite the great deal of enthusiasm following the original meeting at NSF where we brought together people from a number of academic and government research projects. Recent initiatives on automated coding have, if anything, gone further away, focusing exclusively on coding limited sets of *dependent* variables, notably protests. Just getting the dependent variable is not enough: you need the precursors.

Note, by the way, that precursors do not need to be triggers: they can be short-term structural changes that can only be detected via event data because they are unavailable in the tradition structural indicators reported only on an annual basis and/or national level. For at least some IFMs, it has been demonstrated that at the nation-year level, event measures can be *substituted* for structural measures and provide roughly the same level of forecasting accuracy (sometimes a bit more, sometimes a bit less, always more or less in the ballpark). While this has meant there is little gained from adding events to models with nation-year resolution, at the monthly and sub-state geographical levels, events (or something very similar to events) are almost certainly going to be the only indicators available.

4. Native coders vs machine translation

At various points in the past couple of years, I've conjectured that the likelihood that native-language event coders—a very small niche application—would progress more rapidly than machine translation (MT)—an extremely large and potentially very lucrative application—is pretty close to zero. But that is only a conjecture, and both fields are changing rapidly. Multi-language capability is certainly *possible* with universal dependency parsing—that is much of the point of the approach—and in combination with largely automated dictionary development (or, skipping the dictionaries all together, classifiers), it is possible that specialized programs would be better than simply coding translated text, particularly for highly-resourced languages like Spanish, Portuguese, French, Arabic, and Chinese, and possibly in specialized niches such as protests, terrorism, and/or drug-related violence.

Again, I'm much more pessimistic about the future of language-specific event coders than I was five years ago, before the dramatic advances in the quality of MT using deep-learning methods, but this is an empirical question. [14]

5. Assessing the marginal contribution of additional news sources

As I noted in SITPAED, over the course of the past 50 years, event data coding has gone from depending on a small number of news sources—not uncommonly, a single source such as the *New York Times* or Reuters [15]—to using hundreds or even thousands of sources, this transition occurring during the period from roughly 2005 to 2015 when essentially every news source on the planet established a readily-scraped web presence, often at least partially in English and if not, accessible, at least to those with sufficient resources, using MT. Implicit to this model, as with so many things in data science, was the assumption that "bigger is better."

There are, however, two serious problems to this. The first—always present—was the possibility that all of the event signal relevant to the common applications of event data—currently mostly IFMs and related academic research—is already captured by a few—I'm guessing the number is about a dozen—major news sources, specifically the half-dozen or so major international sources (Reuters, Agence France Presse, BBC Monitoring, Associated Press and probably Xinhua) and another small number of regional sources or aggregators (for example, All Africa). The rest is, at best, redundant because anything useful will have been picked up by the international sources. [16] and/or noise. Unfortunately, as processing pipelines become more computationally intensive (notably with external rather than internal parsing, and with geolocation) those additional sources consume a huge amount of resources, in some cases to supercomputer levels, and limit the possible sponsors of near-real-time data.

That's the *best* scenario: the worst is that with the "inversion"—more information on the web is fake than real—these other sources, unless constantly and carefully vetted, are introducing systematic noise and bias.

Fortunately it would be very easy to study this with ICEWS (which includes the news source for each coded event, though not the URL) by taking a few existing applications—ideally, something where replication code is already available—and seeing how much the results change by eliminating various news sources (starting with the extremely long tail of sources which generate coded events very infrequently). It is also possible that there are some information-theoretic measures that could do this in the abstract, independent of any specific application. Okay, it's not that it *might* be possible, there are definitely measures available, but I've no idea whether they will produce results meaningful in the context of common applications of event data.

6. Analyze the TERRIER and Cline Center long time series

The University of Oklahoma and University of Illinois/Urbana Champaign have both recently released historical data sets—TERRIER and yet-another-data-set-called Phoenix [17] respectively—which vary significantly from ICEWS: TERRIER is "only" about 50% longer (to 1980) but [legally] includes every news source available on LexisNexis, and the single-sourced Cline Center sets are *much* longer, back to 1945.

As I noted in SITPAED, the downsides of both are they were coded using the largely untested PETR-2 coder and with ca. 2011 actor dictionaries, which themselves are largely based on ca. 2005 TABARI dictionaries, so both recent and historical actors will be missing. That said, as I also showed in SITPAED, at higher levels of aggregation the *overall* picture provided by PETR-2 may not differ much from other coders (but it might: another open but readily researched question), and because lede sentences almost always refer to actors in the context of their nation-states, simply using dictionaries with nation-states may be sufficient. [18] But most importantly, these are both very rich new sources for event data that are far more extensive than anything available to date, and need to be studied.

7. Find an open, non-trivial true prediction

This one is not suitable for dissertation research.

For decades—and most recently, well, about two months ago—whenever I talked with the media (back in the days when we had things like local newspapers) about event data and forecasting, they would inevitably—and quite reasonably—ask "Can you give us an example of a forecast?" And I would mumble something about rare events, and think "Yeah, like you want me to tell you the Islamic Republic has like six months to go, max!" and then more recently, with respect to PITF, do a variant on "I could tell you but then I'd have to kill you." [19]

For reasons I outlined in considerable detail here, this absence of unambiguous contemporary success stories is not going to change, probably ever, with respect to forecasts by governments and IGOs, even as these become more frequent, and since these same groups probably don't want to tip their hand as to the capabilities of the models they are using, we will probably only get the retrospective assessments by accident (which will, in fact, occur, particularly as these models proliferate [20]) and—decades from now—when material is declassified.

Leaving the task of providing accessible examples of the utility of CRMs instead to academics (and maybe some specialized NGOs) though for reasons discussed earlier, doing so obscurely would not bother me. Actually, we need two things: retrospective assessments using the likes of ICEWS, TERRIER, and Cline-Phoenix on what *could* have been predicted (no over-fitting the models, please…) based on data available at the time, and then at some point, a documentable—hey, use a blockchain!—true prediction of something important and unexpected. Two or three of these, and we can take everything back undercover.

The many downsides to this task involve the combination of rare events, with the unexpected cases being even rarer [21], and long time horizons, these typically being two years at the moment. So if I had a model which, say—and I'm completely making this up!—predicted a civil war in Ghana [22] during a twelve month period after two years, a minimum of 24 months, and a maximum of 36 months, will pass before that prediction can be assessed. Even then we are still looking at probabilities: a country may be at a high *relative* risk, for example in the top quintile, but still have a probability of experiencing instability well below 100%. And 36 months from now we'll probably have newer, groovier models so the old forecast still won't demonstrate state of the art methods.

All of those caveats notwithstanding, things will get easier as one moves to shorter time frames and sub-national geographical regions: for example Nigeria has at least three more or less independent loci of conflict: Boko Haram in the northeast, escalating (and possibly climate-change-induced) farmer-herder violence in the middle of the country, and somewhat organized violence which may or may not be political in the oil-rich areas in the Delta, as well as potential Christian-Muslim, and/or Sunni-Shia religiously-motivated violence in several areas, and at least a couple of still-simmering independence movements. So going to the sub-state level both increases the population of non-obvious rare events, and of course going to a shorter time horizon decreases the time it will take to assess this. Consequently a prospective—and completely open—system such as ViEWS, which is doing monthly forecasts for instability in Africa at a 36-month horizon with a geographical resolution of 0.5 x 0.5 decimal degrees (PRIO-GRID; roughly 50 x 50 km) is likely to provide these sorts of forecasts in the relatively near future, though getting a longer time frame retrospective assessment would still be useful.

A few other things that might go into this list

- **Trigger models:** As I noted in my discussion of IFMs , I'm very skeptical about trigger models (particularly in the post-inversion news environment), having spent considerable time over three decades trying to find them in various data sets, but I don't regard the issue as closed.
- **Optimal geolocation:** MORDECAI seems to be the best open-source program out there at the moment (ICEWS does geolocation but the code is proprietary and, shall we say, seems a bit flakey), but it turns out this is a really hard problem and probably also isn't well defined: not every event has a meaningful location.
- **More inter-coder and inter-dataset comparison:** as noted in SITPAED, I believe the Cline Center has a research project underway on this, but more would be useful, particularly since there are almost endless different metrics for doing the comparison.
- **How important are dictionaries containing individual actors?:** The massive dictionaries available from ICEWS contain large compendia of individual actors, but how much is actually gained by this, particularly if one could develop robust cross-sentence co-referencing? E.g. if "British Prime Minister Theresa May" is mentioned in the first sentence, a reference to "May" in the fourth sentence—assuming the parser has managed to correctly resolve "May" to a proper noun rather than a modal verb or a date—will also resolve to "GBRGOV".
- **Lede vs full-story coding:** the current norm is coding the first four or six sentences of articles, but to my knowledge no one has systematically explored the implications of this. Same for whether or not direct quotations should be coded.
- **Gold standard records:** also on the older list. These are fabulously expensive, unfortunately, though a suitably designed protocol using the "radically efficient" prodigy approach might make this practical. By definition this is not a one-person project.
- **A couple more near-real-time data generation projects:** As noted in SITPAED, I've consistently under-estimated the attention these need to guarantee 24/7/365 coverage, but as we transition from maintaining servers in isolated rooms cooled to meat-locker temperatures and with fans so noisy as to risk damage to the hearing of their operators except server operators tend to frequent heavy metal concerts…I digress…to cloud-based servers based in Oregon and Northern Virginia, this *should* get easier, and not terribly expensive.

Finally, if you do any of these, please quickly provide the research in an open access venue rather than providing it five years from now somewhere paywalled.

Footnotes

1. You will be shocked, shocked to learn that these suggestions have gone absolutely nowhere in terms of funding, though some erratic progress has been made, e.g. on at least outlining a CAMEO alternative. One of the suggestions—comparison of native-language vs MT approaches—even remains on this list.

2. Severely understaffed because the entire project was predicated on the supposition that political scientists—as well as the professional programming team at BBN/Raytheon who had devoted years to writing and calibrating an event coder—were just too frigging stupid to realize the event coding problem had already been solved by academic computer scientists and a fully functioning system could be knocked out in a couple months or so by a single student working half time. Two months turned into two years turned into three years—still no additional resources added—and eventually the clock just ran out. Maybe next time.

I've got a 3,000-word screed written on the misalignment of the interests of academic computer scientists and, well, the entire remainder of the universe, but the single most important take-away is to never, ever, ever forget that no computer scientist ever gained an iota of professional merit writing software for social scientists. Computer scientists gain merit by having teams of inexperienced graduate students [3]—fodder for the insatiable global demand by technology companies, where, just as with law schools, some will eventually learn to write software on the job, not in school [4]—randomly permute the hyper-parameters of long-studied algorithms until they can change the third decimal point of a standardized metric or two in some pointless—irises, anyone?—but standardized data set, with these results published immediately in some ephemeral conference proceeding. That's what academic computer scientists do: they don't exist to write software for you. Nor have they the slightest interest in your messy real-world data. Nor in co-authoring an article which will appear in a paywalled venue after four years and three revise-and-resubmits thanks to Reviewer #2. [6] Never, ever, ever forget this fact: if you want software written, train your own students—some, at least in political methodology programs, will be surprisingly good at the task [7]—or hire professionals (remotely) on short-term contracts.

Again, I have written 3,000 words on this topic but, for now, will consign it to the category of "therapy."

3. These rants do not apply to the tiny number of elite programs—clearly MIT, Stanford, and Carnegie Mellon, plus a few more like USC, Cornell and, I've been pleased to discover, Virginia Tech, which are less conspicuous—which consistently attract students who are capable of learning, and at times even developing, advanced new methods and at those institutions may be able to experiment with fancier equipment than they could in the private sector, though this advantage is rapidly fading. Of course, the students at those top programs will have zero interest in working on social science projects: they are totally involved with one or more start-ups.

4. And just as in the profession of law, the incompetent ones presumably are gradually either weeded out, or self-select out: I can imagine no more miserable existence than trying to write code when you have no aptitude for the task, except if you are also surrounded, in a dysfunctional open-plan office setting [5], by people for whom the task is not only very easy, but often fun.

5. The references on this are coming too quickly now: just Google "open plan offices are terrible" to get the latest.

6. I will never forget the reaction of some computer scientists, sharing a shuttle to O'Hare with some political scientists, on learning of the publication delays in social science journals: it felt like we were out of the Paleolithic and trying to explain to some Edo Period swordsmiths that really, honest, we're the smartest kids on the block, just look at the quality of these stone handaxes!

7. Given the well-documented systemic flaws in the current rigged system for recruiting programming talent—see this and this and this and this and this—your best opportunities are to recruit, train, and retain women, Blacks and Hispanics: just do the math. [8]

8. If you are a libertarian snowflake upset with this suggestion, it's an exercise in pure self-interest: again, do the math. You should be happy.

9. I was originally going to call this the "Pólya trap" after George Pólya's *How to Solve It* —once required reading in many graduate programs but now largely forgotten—and Pólya does, in fact, suggest several versions of solving problems by converting them to something you already know how to solve, but his repertoire goes far beyond this.

10. They are also radically different: as I noted in SITPAED, in their event coding PETR-1, PETR-2, and UP are almost completely different programs with only their actor dictionaries in common.

11. Mind you, these sorts of disappointing outcomes are hardly unique to event data, or the social sciences—the National Ecological Observatory Network (NEON), a half-billion-dollar NSF-funded facility has spent the last five years careening from one management disaster to another like some out-of-control car on the black ice of Satan's billiard table. Ironically, the generally unmanaged non-academic open source community—both pure open source and hybrid models—with projects like Linux and the vast ecosystem of Python and R libraries, has far more efficiently generated effective (that is, debugged, documented, and, through StackOverflow, reliably supported) software than the academic community, even with the latter's extensive public funding.

12. Keep in mind the input to the eventual CAMEO dictionaries was developed at the University of Kansas over a period of more than 15 years, and focused primarily on the well-edited Reuters and later Agence France Presse coverage of just six countries (and a few sub-state actors) in the Middle East, with a couple subsets dealing with the Balkans and West Africa.

13. With a bit more work, one can use scrapping of major news sites and the fact that ICEWS, while not providing URLs, does provide the *source* of its coded events, and in most cases the article an event was coded from is quite unambiguous by looking at the actors involved (again, actor dictionaries are open and easy to update). Using this method, over time a substantial set of current article-event pairs could be accumulated. Just saying…

14. This, alas, is a *very* expensive empirical question since it would require a large set of human-curated test cases, ideally with the non-English cases coded by native speakers, to evaluate the two systems, even if one had a credibly-functioning system working in one or more of the non-English languages. Also, of course, even if the language-specific system worked better than MT on one language, that would not necessarily be true on others due to differences on either the event coder, the current state of MT for that language (again, this may differ dramatically between languages), or the types of events common to the region where the language is used (some events are easier to code, and/or the English dictionaries for coding them are better developed, than others). So unless you've got a *lot* of money—and some organizations with access to lots of non-English text and bureaucratic incentives to process these do indeed have a lot of money—I'd stay away from this one.

15. For example for a few years, when we had pretty good funding, the KEDS project at Kansas had its own subscription to Reuters. And when we didn't, we were ably assisted by some friendly librarians who were generous with passwords.

The COPDAB data set, an earlier, if now largely forgotten, competitor to WEIS, claimed to be multi-source (in those days of coding from paper sources, just a couple dozen newspapers), but its event density relative to the single-sourced WEIS came nowhere close to supporting that contention, and the events themselves never indicated the sources: What probably happened is that multiple sourcing was attempted, but the human coders could not keep up and the approach was abandoned.

16. Keep in mind that precisely because these are international and in many instances, their reporters are anonymous, they have a *greater* capacity to provide useful information than do local sources which are subject to the whims/threats/media-ownership of local political elites and/or criminals. Usually overlapping sets.

17. Along with "PETRARCH," let's abandon that one, eh: I'm pretty good with acronyms—along with self-righteous indignation, it's my secret superpower!—so just send me a general idea of what you are looking for and I'll get back to you with a couple of suggestions. Seriously.

Back in the heady days of decolonization, there was some guy who liked to design flags—I think this was just a hobby, and probably a better hobby than writing event coders—who sent some suggestions to various new micro-states and was surprised to learn later that a couple of these flags had been adopted. This is the model I have in mind.

Or do it yourself—Scrabble™-oriented web sites are your best tool!

18. Militarized non-state actors, of course, will be missing and/or misidentified—"Irish Republican Army" might be misclassified as IRLMIL—though these tend to be less important prior to 1990. Managing the period of decolonization covered by the Cline data is also potentially quite problematic: I've not looked at the data so I'm not sure how well this has been handled. But it's a start.

19. PITF, strictly speaking, doesn't provide much information on how the IFM models have been used for policy purposes but—flip side of the rare events—there have been a few occasions where they've seemed be quite appreciative of the insights provided by the IFMs, and it didn't take a whole lot of creativity to figure out what they must have been appreciative about.

That said, I think this issue of finding a few policy-relevant unexpected events is what has distinguished the generally successful PITF from the largely abandoned ICEWS: PITF (and its direct predecessor, the State Failures Project) had a global scope from the beginning and survived long enough—it's now been around more than a quarter century—that the utility of its IFMs became evident. ICEWS had only three years (and barely that: this included development and deployment times) under DARPA funding, and focused on only 27 countries in Asia, some of these (China, North Korea) with difficult news environments and some (Fiji, Solomon Islands) of limited strategic interest. So compared to PITF, the simple likelihood that an unexpected but policy-relevant rare event would occur was quite low, and, as it happened, didn't happen. So to speak.

20. In fact I think I may have picked up such an instance—the release may or may not have been accidental—at a recent workshop, though I'll hold it back for now.

21. In a properly calibrated model, most of the predictions will be "obvious" to most experts: only the unexpected cases, and due to cognitive negativity bias, here largely the unexpected *positive* cases, will generate any interest. So one is left with a really, really small set of potential cases of interest.

22. In an internet cafe in some remote crossroads in Ghana, a group of disgruntled young men are saying "Damn, we're busted! How'd he ever figure this out?"