USING GDELT FOR ATROCITY EARLY WARNING
A QUICK START GUIDE V1.0

3/9/2014

http://gdeltproject.org/

UNDERSTANDING GDELT

It is important to begin this guide by noting that GDELT is truly the first of its kind. It is the first multi-decade all-countries city-georeferenced broad-spectrum realtime event dataset ever constructed for open academic research. It is also the largest event dataset ever created for any purpose and is the first to blend within a single project multiple decades of tens of thousands of distinct news sources from across the globe, including realtime updates, of translated print and broadcast news reports, processing the global news media ecosystem.

In this regard, GDELT is truly on the "bleeding-edge" of the limits of current technology. It was designed to demonstrate that we have reached a point where the technologies and methodologies have matured to the point where it is possible to create a database this advanced, detailed, and comprehensive. Yet the limits to which it pushes technology necessitate that it is also rough around the edges and is an active platform for learning how to construct these classes of truly "societal scale" databases to build the first "realtime social sciences earth observatories."

Indeed, many of the core components of GDELT are themselves areas of active research. For example, when typing a location into Google Maps, the entire text can be assumed to be a location, with the geocoding engine's only task being to properly parse the location and interpolate it from its geographic database. GDELT, on the other hand, must read through a potentially pages-long news article written by a non-native-speaker or translated by machine, searching for possible geographic references, determine whether they really are mentions of location rather than similarly-spelled person names, disambiguate them using the surrounding context of the rest of the document, and ultimately output a set of centroid locations for each mention: a process known as "full text geocoding." It must go even further, however, and perform "geographic coreference" in which it must associate each person, relationship, and activity in the text with its affiliated location, and propagate these affiliations throughout the entire text.

Pronoun coreference and temporal extraction, relative offset resolution, action-level affiliation, and semantic network construction are all key components of the GDELT processing pipeline. Generalized dyadic relationship and claim extraction, including in complex grammatical constructs, lie at the very core of GDELT's ability to transform lengthy and often flowery textual descriptions of events into codified numeric records. Even after all of the events have been extracted, new deduplication processes were required that take into account for the first time city-level geographic information and the ambiguity of geographic and actor attribute information, especially in active conflict environments. Each of these is a research area of its own with a large and active literature, which GDELT had to draw together into a single unified pipeline. To make matters even more complex, GDELT must be robust towards an enormous variety of content, from non-native speakers to translated transcribed radio and television broadcasts to OCR'd historical materials to realtime closed captioning, each of which may deviate considerably from the well-formed grammatical constructs that most natural language processing tools today have been trained and tested upon.

In this way, almost every component of GDELT pushes the boundaries of an array of fields of research and is the first project to draw all of these methodologies and approaches together into a single system.

At the same time it is also not a multi-million dollar initiative with a large team of people behind it that have exhaustively explored every possible algorithm, methodology and tool available. Instead, it is a “social good” project, designed to collect a cross-section of tools and approaches available today and demonstrate a vision of what's possible and, most importantly, to inspire scholars across all fields to "think big" and imagine a world populated with this kind of “societal scale” data and to step forward to help lead the charge towards a new era of studying our global world.

HUMAN ACCURACY

When examining the accuracy of GDELT, it is important to note that contrary to intuition, the baseline of having trained human analysts read each article and identify and code each event therein is actually highly inaccurate and extremely variable across coders. As King & Lowe (2003) [1] point out in one of the few detailed comparisons of human and machine coding, machine coding achieves around 93% accuracy at recognizing the existence of an event, while human coding varies between 80-94%. Assigning events to specific detailed event codes, both machine and human perform at around 70% accuracy at the aggregate level, though humans perform as low as 23% accuracy when coding individual event categories, with significant variability, even over the short period of time of the authors’ study. Machines do have a substantially higher false positive rate, but as the authors note “because these data are unrelated to any measured variable, they should not bias any subsequent inferences” and “these extra events are not more likely to appear in some categories than others.” Such error is not limited to small academic test environments. Even when examining the widely-used human-constructed production event database ACLED, Kristine Eck (2012) [2] found that between 25-50% of its records have error substantial enough to affect their use in analysis and that even basic facts such as whether the event took place in a city or other location was wrong for nearly a third of events.

Thus, it is critically important to understand that human-coded event records are actually highly error-prone themselves and that machine coding performs on par with human coding other than a higher false positive rate, though this rate is evenly distributed across categories and so dissipates under analysis as noise. However, even though it is easily filtered as noise, this higher false positive rate means that one cannot simply filter for isolated records mentioning an attack against civilians – one must instead look for trends that shift away from the baseline, as the next section will detail.

In fact, machine coding is the only tractable method of creating the daily or realtime update streams that reflect the up-to-the-moment state of the world needed to create the types of nowcasting and forecasting dashboards used in operational settings. It would be nearly impossible to assemble a team of human coders large enough, and to maintain their training levels over time, to be able to code the tens of millions of daily news reports published and broadcast each day around the globe, and entirely impossible to code the tens of billions of daily social media items created each day. Indeed, machine coded event datasets now account for a substantial fraction of available event data, and machine coding pipelines are at the heart of nearly all production dashboards today. So, while the results of GDELT are far from perfect, the core approaches to machine event coding that it relies upon are at the heart of most production systems today, only expanded and scaled up a thousand-fold in GDELT.

[1] http://gking.harvard.edu/files/gking/files/infoex.pdf
[2] http://pcr.uu.se/digitalAssets/147/147084_eck.coco.2012.final.full.pdf

HOW TO USE GDELT FOR ATROCITY EARLY WARNING

Using large real-world automatically-constructed datasets like GDELT require a fundamentally different approach to monitoring, modeling, and filtering than most researchers are used to. At first glance, it might seem that a trivial approach to identifying new atrocities would be to simply search for events that include the "Civilian" code in the victim Actor Role and/or one of the Mass Atrocity event type codes. However, think for a moment about how one would organize this search if one was conducting it entirely by hand through a manual review of media coverage.

Using Syria as an example, imagine what would happen if one came across a single news article this afternoon that reports that a chemical attack occurred today with hundreds of civilian casualties, yet by this evening not a single additional news outlet in the world has reported on this attack. As a human analyst, one would likely be highly suspicious of this account, even if it was reported in a major publication that is normally trustworthy. If, on the other hand, hundreds of papers around the world reported on the event within a few hours, one would still have to be suspicious about whether chemical weapons were actually used and the specifics of the attack, but one could at least be more confident in the likelihood that some form of attack probably occurred. In other words, as a human analyst, one would never sound the alarm of a mass atrocity against civilians based on a single news report – significant triangulation would be required first. Similarly, when the Associated Press Twitter account was hacked in April 2013 and reported a bombing at the White House that injured President Obama, the absence of other media coverage of the attack within a short duration was a clear indication that the report was likely false.

Similarly, if a single event record appears in the GDELT daily update stream reporting on a mass killing, but that killing is mentioned only in a single news report that day (NumMentions and NumSources are both set to "1"), this is often a sign that the record could possibly be a false positive (perhaps a past killing being mentioned on its anniversary and using language that it was unable to properly process to date-shift the event), or false or speculative information conveyed by a single news outlet (such as the hacked AP Twitter account). A true mass killing will attract substantial media attention from a diversity of outlets, even at the earliest "rumor" stage. Yet, this is complicated by the fact that GDELT uniquely monitors small local domestic media outlets, reaching even into rural vernacular-language broadcast outlets throughout the globe. This allows it to pick up the earliest mentions of attacks long before they reach larger outlets elsewhere.

Indeed, when searching for emerging atrocities, the goal is to catch their earliest traces. By the time the international press is blaring a collective headline about thousands of civilians killed due to their religious beliefs, that is no longer early warning, it is simply assembling post-hoc notification. The atrocity-related event categories in CAMEO are therefore less useful for atrocity early warning, since at best they merely record what is already a widely-reported assessment. Instead, one should focus on geographically-centered bursts of events that share common attributes of perpetrators, victims, religion, ethnic, or actor roles. For example, a surge in events near a refugee camp in Africa over a period of days, or a rise in attacks towards civilians by gunmen in a specific region are all potential indicators of an impending atrocity, even before the world's major newspapers run a headline announcing it weeks or months later when it has been formally documented as such.

Atrocity early warning projects should therefore focus on pattern detection, especially measurable increases in violent events within a specific geographic area and time period or involving specific actors or attributes. This is similar to how sentiment (tonal) measures are used to assess changes in views towards a product or organization, by looking not for an isolated negative tweet, but rather by looking for sudden shifts away from the recent baseline. At the most basic level, one could liken using GDELT for atrocity early warning to using the Google Books NGram collection to understand our literary history – their greatest value lies in allowing one to see broad trends that simply would never otherwise be visible. Ngrams don't replace intensive reading of a single book if one is trying to deconstruct that book's views onto a topic, but rather let one place that book in the context of millions of other books on that topic and others published through time and space to understand its broader significance and overarching themes and patterns. In both cases the goal is not to focus or dissect a single incident or book in considerable detail, but rather to look across the data for macro-level patterns, such as spatial diffusion of conflict or to examine the broader conditions under which certain types of behavior occur more or less often.