Transforming Textual Information on Events into Event Data within SPEED

June 2011

Cline Center for Democracy
University of Illinois at Urbana-Champaign
ABSTRACT

Creating a valid and reliable body of event data requires meeting a number of challenges (creating a domain ontology, developing reliable sources of information on events of interest, identifying information on relevant events within those information sources, etc.). The fact that most event data projects, including SPEED, use news reports as sources of information generates additional challenges. Some of the most important of these are cognitive challenges involved in transforming textual information contained in news reports into event-centered information. Information in news reports can be convoluted and contain much information that is extraneous to the events of interest. Moreover, some news reports contain information on multiple events, requiring the disambiguation of the multiple streams of information. Dealing properly with the issues involved in transforming news text into event-centered information is crucial to generating sound event data and maximizing the amount of information extracted. This document outlines these issues and how they are addressed within the SPEED project. The first section introduces some of the basic issues involved (non-events, vague references, statements by interested parties, referent events) by focusing on the simplest type of news report: those containing information on a single event. The second section discusses issues posed by reports that contain information on multiple events (related events, recapitulation passages).

CONTRIBUTORS

- Peter F. Nardulli
- Matthew Hayes
Table of Contents

Introduction ........................................................................................................................................ 1
Single-event News Reports ................................................................................................................ 2
  Non-events .................................................................................................................................. 2
  Vague references ......................................................................................................................... 2
  Statements by interested parties ................................................................................................. 3
  Referent events - ....................................................................................................................... 3
Multiple-event News Reports: Related Events and Recapitulation Passages ......................... 5
  Related events .......................................................................................................................... 5
  Recapitulation Passages ........................................................................................................... 7
Creating a valid and reliable body of event data requires meeting a number of challenges. The most basic of these involve clearly and comprehensively defining the events to be studied (i.e., creating a domain ontology that defines the substantive scope of the project) and developing reliable sources of information on those events. The fact that most event data projects, including SPEED, use news reports as the source of information on events generates an additional set of challenges. Among the most salient of these concern issues such as the credibility and comprehensiveness of news coverage, as well as media bias and fatigue. Also relevant are technical issues involved in the digitization of historical news texts (which can lead to extensive transcription error) and the systematic identification of relevant news documents. The use of multiple news sources, optical character recognition (OCR) programs, and powerful search algorithms embedded in automatic text categorization (ATC) routines can be used to address these issues. But quite independent of these matters are cognitive challenges involved in transforming textual information organized within news reports into event-centered information. This document outlines these challenges and how they are addressed within the SPEED project.

As in most event data projects, the unit of analysis within SPEED is the event, which is to say that the event is the focus of the information extraction process. Correspondingly, the most elementary output of SPEED’s operating environment (a suite of software programs labeled EXTRACT) is an event record. An event record is an electronic entry in a database system that contains information on a single, discrete event. To generate that record, textual information in news reports must be transformed into quantitative data on events. This transformation is challenging – even if an effective means is devised for identifying news reports with information on relevant events. Difficulties arise because even “screened” news reports can be convoluted and contain much information that is extraneous to the subset of events delineated in the project’s domain ontology. Moreover, some news reports contain information on multiple events; disambiguating the multiple streams of information embedded in these reports further complicates the transformation process.

Dealing properly with the issues involved in transforming news text into event-centered information is crucial to generating sound event data and maximizing the amount of information extracted from news reports. Thus, the next section introduces some of the basic issues involved by focusing on the simplest type of news report: those containing information on a single event. The second section discusses issues posed by reports that contain information on multiple events.
Single-event News Reports

Within event data projects the primary objective in analyzing news reports is to identify and extract information about discrete events that meet certain definitional criteria. These criteria determine whether textual references describe an “event” (i.e., a happening that occurs at a specific time and place) and, if so, whether that event belongs to the subset of events specified in the domain ontology (e.g., Does it qualify as a destabilizing event? A property rights violation? A civil war battle?). In thinking about how to structure the content analysis of information about this subset of events it is useful to introduce the notion of a “focal event.” A focal event is simply the event that is the focus of the information extraction process. The event record that is generated by the extraction of event-specific information (event type, actors, origins, etc.) should contain only designated information that pertains to the focal event. Identifying this event-specific information properly can be challenging even when there is just one focal event in a news report because it normally contains extraneous information. The most important types of extraneous information include text on non-events, vague references about seemingly pertinent events, self-serving statements, and referent events. These sources of extraneous information are list in Figure 1, which depicts the transformation process for single event reports.

Non-events - Many news reports contain information that is not about an “event,” which SPEED defines as “happenings that occur at a discernible time and place.” Sometimes these spatial and temporal attributes are very specific (an assassination in New York on November 2, 1983 at 4:00); other times they are more opaque (the plane was taken over by terrorists as it crossed the Atlantic late last night; the riot unfolded over three days in northwestern Afghanistan). In contrast, information in news reports that is not about events often takes the form of speculation about future happenings. For example, there may be a news report about a conference in which a speaker reports that if something is not done about global warming, “large segments of Africa will be subject to droughts and famines that will generate massive refugee movements and widespread violence.” The same can be said about a threat issued by a mayor to authorize his security forces to shoot demonstrators if they approach government buildings. The threat could be an event; the shootings are not events because they did not happen.

Vague references - In contrast to information about non-events, vague references may contain information that pertains to relevant events – but not enough to provide the basis for creating an event record. To create an event record three crucial pieces of information are needed: what, where, and when. Vague references have at least one of these pieces of information missing, and so they cannot provide the basis for generating an event record. For example, a news report might refer to the "destruction of the temple at Ayodhya," but give no clues as to when it occurred. Similarly, a news report may omit spatial information, such as "the killing of the prime minister while traveling abroad," which does not provide even basic information (i.e., country-level information) on where the event occurred. These vague references should be differentiated from recapitulation passages, which are crucial sources of information on events. In contrast to vague references, recapitulation passages contain information on where an event occurred (at
least at the country-level), when it occurred (at least a date range), and what occurred. What differentiates recapitulation passages from descriptions of discrete events is that they bundle information on a number of relevant events; they are discussed more fully in the next section.

**Statements by interested parties** - SPEED aims to extract information on events from *disinterested* accounts of events contained in news reports. Correspondingly, SPEED’s global news archive is constructed using only sources from reputable news providers. Thus, the vast majority of news accounts stored in SPEED’s archive provide disinterested, reasonably objective accounts of events. Problems emerge, however, when text in a news report does not focus on an event but on a self-serving statement by an interested party about an event (i.e., propaganda). Consider, for example, a news report on a statement issued by the government that repressed demonstrations were instigated by communist agents planning an armed insurrection. While a news provider may consider this statement to be newsworthy, the information on the event’s origins should not be included in an event coding unless the assertion was corroborated independently in the news report. In contrast, if the same assertion was made by an independent and fair-minded governmental commission, then the information should be included in the event coding. Statements by insurgents who claimed that they were the object of human rights violations perpetrated by government forces would not provide a “disinterested and objective” basis for initiating an event coding; reports of such violations by a reputable NGO in the field would justify the creation of a coding.

Referent events - Referent events are highly visible “landmark” events that generate a large number of oblique references in myriad news accounts – even though they are only distally related to the focus of a particular story. Examples of landmark events are highly visible happenings such as the 9-11 attacks in the U.S., the Tiananmen Square massacre in Beijing, the assassination of Anwar Sadat in Egypt, the bombing of the U.S. Marine base in Lebanon, etc. For example, in a news report on the student protests against tuition increases, it may be noted that “The protests were the first mass demonstrations since the massacres at Tiananmen Square.” The Tiananmen Square incident is mentioned in the article and, like the student protests, it is a civil unrest event. But it is not related to the protests over tuition. Nor is it an independent event that is described in the news report. It is just a distal occurrence offered as background.

Dealing with information on referent events is challenging because, while they may meet the definitional criteria for a particular domain ontology, coding every reference to them can generate multiple event records for the same event. This, of course, will bias the event database and the analyses based on it. This notwithstanding, it is difficult to inform coders of the broad array of happenings that qualify as referent events. Moreover, news reports on referent events often provide updated information (number killed or injured, identities of the initiators, number of security officials injured, motivations of the initiators, etc.) on referent events. Thus, while less efficient, all but the most salient referent events (e.g., 9/11, Tiananmen Square, the Kennedy assassination) are dealt with in a post hoc screening procedure that uses clustering techniques to identify multiple event records for referent events.
Figure 1
Transforming News Text Into Event Data:
Single Event Reports

News Report

Uncodeable Text
- Non-events
- Vague, Incomplete Event References
- Self-serving Statements
- Referent Events

Codeable Text
- Information on Focal Event

Event Database
- Event Record
- Event Record
- Event Record
- Event Record
Multiple-event News Reports: Related Events and Recapitulation Passages

Single-event news reports provide the simplest illustration of the challenges involved in transforming textual information into quantitative event data. But many, if not most, news reports contain information on several events that are relevant to a particular domain ontology; information on each of these events should be reflected in separate event records. Some news reports also provide background information that, while insufficient to generate an event record, is highly relevant to the goals of event-based research. To maximize the amount of information extracted from news reports and better capture the historical record, this background information should be captured. Disambiguating the multiple streams of information embedded in multiple-event reports complicates the transformation process, but it also presents opportunities that enhance the utility of event data. The most important of these are the creation of electronic links among related events and the integration of information from recapitulation passages.

Related events - In thinking about how to handle multiple-event news reports it is useful to introduce a simple typology that describes the types of events embedded in them and their relationship to one another. This typology is depicted in Figure 2. Because of norms that have evolved in news reporting most reports highlight one principal event; it is labeled the primary event (see Figure 2 (a)). Details on the primary event are often provided at the beginning of the report, though some details are often spread throughout the report. All other relevant events in the report are termed secondary events. There are two basic types of secondary events that are important for enhancing the power of event data: related secondary events (Figure 2 (a)) and independent secondary events (Figure 2 (b), (c)). The difference between an independent and a related event is that a related event is thought to be associated in some way with another event in the report. Consider, for example, a report about an arson of a Sunni mosque perpetrated by a group of radical Shiites. If the report documents that the arson was in retaliation for a violent attack committed by Sunni against a group of elderly Shiites, then the Sunni attack would be a related secondary event. If, however, the attack is simply mentioned as part of an accounting of a series of unrelated violent events, then it is an independent secondary event – simply an “add-on” to the story about the primary event. An independent event may be similar to the primary event (i.e., they are both attacks), or it may be a very different type of event that happened in the same locale and day as the primary event. A related event can be associated with either the primary event or an independent secondary event, as noted in Figure 2 (a) and (b).
Figure 2
Illustration of Event Typology for Multiple Event News Reports

Figure 2 also makes it clear that there are several important types of related events. One important distinction is between antecedent and subsequent events. An example of an antecedent event is the abovementioned Sunni attack that preceded the arson of the Sunni mosque. An example of a subsequent event would be a suicide attack by Sunnis on a café frequented by Shiites that occurred in retaliation for the Shiite arson. It is also useful to distinguish subsequent events that have direct implications for the initiators of the prior event and those that do not. The example of the Sunni attack on the Shiite café is an example of a subsequent attack that did not have direct consequences for the Shiite arsonists who destroyed the mosque. However, if a group of Sunnis’s attacked the Shiite arsonists as they were fleeing from the mosque, then it would be an example of a post-hoc reaction that affected the initiators. Another example of a post-hoc reaction would be if police officers arrested the arsonists as they fled from the mosque.
Grasping the distinctions among the different types of events depicted in Figure 2, and the information streams associated with them, is important because they are essential to developing links among related events. The creation of operator links within SPEED is facilitated by EXTRACT’s LINK module, which makes it possible to “chain” related event records to one another. These operator links can be augmented by chaining techniques that involve the post-processing of event records for a given spatial/temporal unit or for related sequences of events. Linking related events is important because it transforms a body of individual records into an integrated database that greatly enhances the power and utility of event data.

Recapitulation Passages - News reports frequently make references to relevant events that do not provide the type of event-specific information required to generate an event record. These references are termed recapitulation passages. An example of a recapitulation passage is the statement that “hundreds of villagers, including women and children have been killed in a series of attacks by heavily armed guerrillas in the regions outside Nairobi over the past month.” While this information is relevant to an event analysis pertaining to civil unrest, the reference to the killings bundles information from a set of actions. Unfortunately, it does not provide event-specific information on the individual events, which would be needed to create an event record. Another illustration would be a reference to the fact that “hundreds of thousands of protesters have stormed the streets of Paris in almost daily demonstrations since the beginning of July.” Such information provides important context for an analysis of government reactions to the unrest or for our understanding of a major rally that followed the earlier protests. But because the reference to hundreds of thousands of protesters constitutes a summary of a series of actions, it cannot be used to create an event record.

Information contained in recapitulation passages is important to capture because it can compensate for important deficiencies in media coverage. Due to resource constraints, it is impossible for media organizations to provide coverage of all relevant happenings in all countries at all times. However, conscientious media organizations can fill gaps in coverage by backtracking information leads and recounting the information in recapitulation passages. Discarding such information because it does conform to the mode used to present descriptions of discrete events is costly and fails to maximize the resources invested in the information extraction process. Thus, SPEED captures this information in what is termed recapitulation codings. Distinguishing recapitulation codings from event records is important because recapitulations bundle information from a number of events. Failing to acknowledge this distinction can lead to the creation of observations that appear to be extreme outliers, when they are simply the aggregation of a number of routine happenings. When handled properly, however, data captured in recapitulation codings can be used to provide a more accurate basis for aggregating relevant events for a given locale or time frame. Moreover, in many instances, the distribution of recapitulation codings look very similar to the distribution of discrete events; in these cases they can be viewed as quasi-events and treated like any other event record.

Figure 3 provides an outline of the transformation process for multiple event news reports. It makes it clear that database entries can be either event records or recapitulation codings. Independent events (either primary or secondary events) can be linked with related events, either antecedent or subsequent; these links are indicated by the arrows between event records in Figure 3. Recapitulation codings can also be linked to discrete event records. Figure 3 also highlights the distinction made between related subsequent events and those with direct consequences for the initiators of the prior event.
Figure 3
Transforming News Text Into Event Data:
Multiple Event Reports

News Report

Uncodeable Text
- Non-events
- Vague, Incomplete Event References
- Self-serving Statements
- Referent Events

Codeable Text
- Antecedent Event
- Primary Event
- Subsequent Event
- Secondary Event
- Recapitulation Passage

Event Database
- Event Record
- Event Record
- Event Record
- Event Record
- Event Record

Cline Center for Democracy, University of Illinois