DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation	King, Gary, and Will Lowe. 2003. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. International Organization 57(3): 617-642.
Published Version	doi:10.1017/S0020818303573064
Citable link	http://nrs.harvard.edu/urn-3:HUL.InstRepos:3965112
Terms of Use	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design

Gary King and Will Lowe

With too few exceptions, quantitative international relations is the study of annual, quarterly, or sometimes monthly observations of the international system. Interactions among nations, in contrast, take place on a day-to-day basis: When the Palestinians launch a mortar attack into Israel, the Israeli army does not wait until the end of the calendar year to react. Although scholars have produced a lot of interesting research from these aggregated observations, we cannot avoid missing a good deal of the structure of the international system unless we examine international events as they occur.

Adding realism to quantitative international relations has been the hope of the events data movement, where quantitative summaries of individual events between or within nations are coded from written accounts by journalists. Newspapers, wire reports, and other journalistic accounts constitute an imperfect summary of the events in international relations: coverage is not uniform, and it varies according to the needs of the reporters rather than the scholarly need for representativeness. However, a large fraction of information available to political scientists for all types of analyses—events data, annual data, other quantitative summaries, qualitative accounts, historical studies, etc.—passes through the hands of reporters at some point. It is imperfect, and much additional research could and should be done to identify and correct the biases, but journalism is the source of most information that academics have about the international community outside of official government sources. And there should be no controversy over the claim that the immense volume of reportage on international relations constitutes an enor-

Thanks to Marianne Abbott, Doug Bond, Joe Bond, Gerard Bradford, Carl Cobb, John Freeman, Joshua Goldstein, Patricia Hastings, Craig Jenkins, Dylan Balch-Lindsay, Churl Oh, Jon Pevehouse, Kevin Quinn, Phil Schrodt, and Langche Zeng for helpful discussions; Valerie Abejuro, Sam Cook, Dan Epstein, Andrew Holbrook, Shane Jensen, Orit Kedar, Adrian Ma, Amit Pathare, and Mikhail Pryadilnikov for research assistance; and the National Science Foundation (IIS-9874747), the National Institutes of Aging (P01 AG17625-01), the Weatherhead Center for International Affairs, and the World Health Organization for research support. The data described in this article are available at http://GKing.Harvard.edu.

International Organization 57, Summer 2003, pp. 617–642
© 2003 by The IO Foundation.
DOI: 10.1017/S0020818303573064

mous, and insufficiently mined, treasure of information about the international system.

Because even reading this much information, much less understanding it all, is physically impossible for individual researchers, we need some summary of it. One reasonable summary would come from reading a necessarily small, selective fraction of the materials. This could provide a very deep understanding of a very small fraction of the materials. Another summary would come from analyzing a quantitative coding, which produces a more shallow understanding of a much larger fraction of materials. Optimally, and ultimately, scholars should use both approaches—to identify "important" areas (by some definition) via a quantitative approach, and then to examine the specific cases identified in more depth via a more traditional qualitative approach.

Our focus here on quantitative events data seeks to shore up this part of the optimally combined quantitative-qualitative approach. Quantitative events data collectors have made enormous progress over the years, but it is probably fair to say that the approach has not yet been able to fulfill its promise, and results from events data analyses have not always caught on in the academic community.[1] Indeed, one good measure of the problems is that qualitative scholars have not yet been very interested.

In recent years, however, a fairly massive change has been occurring in the event data community. For decades, scholars have coded events data by hand, resulting in many individual collections—for example, Rummell's Dimensions of Nations,[2] the Conflict and Peace Data Bank (COPDAB)[3] and its continuation as the Global Event-Data System (GEDS),[4] the Behavioral Correlates of War (BCOW) project,[5] and McClelland's World Events Interactions Survey (WEIS).[6] For a time, these data collections were used in a fairly large fraction of studies in quantitative international relations. During the 1980s and 1990s, however, as these collections became outdated, academic use steadily dropped.[7]

One problem is that collecting these data "by hand" usually means long, tedious, and painstaking work conducted by dozens of undergraduate and graduate students reading newspapers and coding events into many categories. By and large, these hand-coding efforts have now ended; many have been replaced with projects where computer programs "read" news reports, extract information, and produce quantitative data from them.[8] The advances in computer software, speed of computer hardware, and knowledge of computational linguistics have converged, making

1. Laurance 1990.
2. Rummell 1975.
3. Azar 1982.
4. Davies and Daniel 1994.
5. Leng and Singer 1988.
6. McClelland 1978; see Schrodt 1995 for a review.
7. Laurance 1990.
8. Schrodt and Gerner 1994.

at least some people think that an automated events data approach is becoming feasible. Prominent articles that use automated events data have started to appear.[9]

Although many more scholars may have begun to consider this data source, there exists no evaluation of any program conducted independently of the authors of that program. Independent evaluation is of course the gold standard in science generally, and in computational linguistics in particular—which has for the most part focused on applications other than international conflict—and it should be here too. We have no formal connection with any group creating automated events data.

To conduct our evaluation, we were forced to develop a new research design and new methods to accompany it. These efficient methods enable us to evaluate data in which the relative frequency of events is far from uniform (for example, military conflict occurs far less frequently than diplomatic communiques). We have reason to believe that these methods will be of use in other applications in computational linguistics, but without something like them, we would be unable to perform any serious, unbiased evaluation without an immense expenditure of time and research resources.

Our results surprised us. The computer program we evaluated was able to extract information from Reuters news reports on a level equal to trained Harvard undergraduates—and this was for a short-term application. Computer programs do not get tired, bored, and distracted, and so in the long run the program would certainly outdo any human coder that would be feasible for a researcher to recruit. These results suggest that, perhaps for the first time, automated information extraction programs are ready for primetime in the analysis of international events data.

The Software, Event Ontology, and Data

The Virtual Research Associates, Inc. (VRA)[10] Reader is a software tool that processes data either directly from the Reuters Business Briefing (RBB) newswire, or from a precompiled database of RBB news stories. The Reader extracts the first sentence, or lead, from RBB articles and attempts to deliver a compact quantitative summary of all the events that are described in the lead. This takes advantage of a common practice in journalism, in which reporters learn to write lead sentences that summarize the key points in their article. Lead summaries are thus represented by the program as database records with fields for the *source* and *target* actors of each event and a numerical code for the *type of event* that occurs between the actors. The type of event is coded into a 157-category typology called

9. For example, see Goldstein and Pevehouse 1997; Schrodt and Gerner 2000; and Goldstein et al. 2001.

10. See ⟨http://www.vranet.com⟩. Accessed 14 April 2003.

the Integrated Data for Events Analysis (IDEA)[11] (The record also contains about forty other fields providing geopolitical information on the actors and more detailed characterizations of the event.) For example, here are two sample leads with source actors (S), target actors (T), and numerical IDEA categories annotated:

> *Russian artillery*[S] south of the Chechen capital Grozny *blasted*[223] *Chechen positions*[T] overnight before falling silent at dawn, witnesses said on Tuesday.

> *Israel*[S] said on Tuesday it *sent humanitarian aid*[073] to *Colombia*[T] where a *massive earthquake*[S] last week *killed*[96] at least *938 people*[T] and injured 400.

The VRA reader can map the specific actors and targets identified above to higher-level categories that provide more meaning in comparative analysis. For example, the program outputs will indicate that "Chechen positions" is a place in "Chechnya," which is part of "Russia." IDEA codes 223, 073, and 96 denote military engagement, humanitarian aid, and natural disaster, respectively. Reuters leads are written to provide a precis of the full news story, so it is very common for news leads to contain multiple events. See Table 1 for the other IDEA category codes and definitions.[12]

The Reader's native representation of the type of events is given by their numerical IDEA codes, which we very briefly summarize in Table 1 (the actual documentation of each event type is more detailed).[13] IDEA codes are an example of an ontology,[14] which in computer science as well as philosophy is a description of the kinds of things that can occur or exist in some domain of knowledge; it is typically hierarchically organized and is intended to be mutually exclusive and exhaustive. The IDEA codes created by VRA constitute an ontology because they are a hierarchically organized typology of all that can happen in the field of international relations.

IDEA rearranges and substantially extends McClelland's WEIS ontology. WEIS is organized around twenty-two "cue" categories (01 to 22), which are high-level descriptions of events such as requests, threats, denials, and military actions. More specific event forms are subtypes of cue categories and are denoted by an extra digit. For example, in WEIS, 09 is a request, 091 is a request for information, and 093 is a request for material assistance. IDEA extends the cue categories up to 99 (with gaps, this introduces twelve new cue categories), and provides much finer-grained substructure to cover events involving nonstate actors, mass protest behavior, economic activity, natural disasters, and biomedical phenomena. IDEA provides particularly detailed representation for WEIS category 22—which corre-

11. See Bond et al. 2001.

12. Very occasionally the news lead contains no true event. For example, on 21 December 1990, the lead sentence of a story from Pyongyang was: "Electric guitarists in billowy white dresses, actors 'full of endless hope and romance' and troupes of singing 'flower buds'—this is showtime, North Korean style."

13. See ⟨http://www.vranet.com/IDEA/⟩. Accessed 14 April 2003.

14. Sowa 1999.

sponds to the use of force—by adding two more levels of hierarchical structure to the coding scheme, thus allowing four-digit event codes.

IDEA codes can be aggregated into the WEIS ontology, though few categories have exact one-to-one matches. Consequently, although IDEA is the Reader's native ontology, the software can also generate WEIS codes as output. In addition, the Reader's documentation provides a mapping of IDEA categories onto the Protocol for the Assessment of Nonviolent Direct Action (PANDA)[15] codes, and onto Goldstein's[16] conflict-cooperation scale.[17]

The structure of the IDEA categories depends on several factors. They are intended to be congruent with preexisting WEIS categories, and the level of detail is driven in part by theoretical interests—as in the categories of mass protest behavior, and in part by the interests of VRA's clients—as in the highly articulated subtree describing subtypes of the use of force. Consequently, IDEA is relatively unbalanced, and some categories, such as economic and legal activity, are very sparsely described.

One interesting upshot of the move from human to machine event coding is in the definition of 'event.'[18] When a machine is programmed to read text, it is natural to tie the definition of an event closely to an easily recognized, relatively superficial linguistic structure, because the 'intuitive' inferences a human coder would effortlessly draw when presented with a news lead must either be painstakingly explicated in the program code or avoided altogether. The presence of a particular verb or verb combination typically signals an event category.[19] Consequently, the greatest effort for a researcher wishing to use machine-coded events data is expended constructing a suitable dictionary for their subject matter. This goes some way to automatically realizing Leng and Singer's requirement that their (human) coders "describe the overt moves of the protagonist governments and leave the judgements as to the motives of the actors and the consequences of the sequence of moves to the analysis phase of the research process."[20]

Another unremarked advantage of highly articulated event ontologies such as IDEA is that a researcher has considerable scope for avoiding problems because of the fact that, conditional on his or her substantive theory, many distinct event types can be substituted for one another.[21] The existence of articulated low-level category structure makes it easy to adapt an existing ontology by reaggregating event categories. This process effectively creates a new set of cue categories that are more appropriate to the analyst's theoretical assumptions than the assumptions

15. Bond, Jenkins, Taylor, and Schock 1997.
16. Goldstein 1992.
17. Taylor et al. 1999.
18. Merritt 1994, 21–22.
19. Gerner et al. 1994.
20. Leng and Singer 1988, 158.
21. Most and Starr 1984.

TABLE 1. *The IDEA ontology*

Goldstein	*IDEA*	*Definition*	*Goldstein*	*IDEA*	*Definition*
8.3	072	extend military aid	−2.8	12	accuse
7.6	074	rally support	−3	161	warn
7.6	073	extend humanitarian aid	−3	16	warn
7.4	071	extend economic aid	−3.4	122	denounce or denigrate
6.5	081	make substantial agreement	−3.8	194	halt negotiations
5.4	064	improve relations	−4	1134	break law
5.2	0523	promise humanitarian support	−4	1132	disclose information
5.2	0522	promise military support	−4	1131	political flight
5.2	0521	promise economic support	−4	113	defy norms
5.2	052	promise material support	−4	1123	veto
4.8	083	collaborate	−4	1122	censor media
4.8	08	agree	−4	1121	impose curfew
4.7	05	promise	−4	112	refuse to allow
4.5	051	promise policy or nonmaterial support	−4	111	reject proposal
3.5	0432	forgive	−4	11	reject
3.5	04	endorse or approve	−4.4	2122	political arrest and detention
3.4	093	ask for material aid	−4.4	2121	criminal arrest and detention
3.4	092	solicit support	−4.4	212	arrest and detention
3.4	043	empathize	−4.4	171	nonspecific threats
3.4	041	praise	−4.5	1963	administrative sanctions
3	082	agree or accept	−4.5	1961	strike
2.9	065	ease sanctions	−4.5	196	strikes and boycotts
2.8	054	assure	−4.5	19	sanction
2.8	033	host meeting	−4.9	151	demand
2.5	062	extend invitation	−4.9	15	demand
2.2	0655	relax curfew	−5	201	expel
2.2	0654	demobilize armed forces	−5	20	expel
2.2	0653	relax administrative sanction	−5.2	1813	protest defacement and art
2.2	0652	relax censorship	−5.2	1812	protest procession
2.2	0651	observe truce	−5.2	1811	protest obstruction
2.2	0632	evacuate victims	−5.2	181	protest demonstrations
2.2	063	provide shelter	−5.6	193	reduce or stop aid
2.2	06	grant	−5.8	172	sanctions threat
2.2	0431	apologize	−6.4	175	nonmilitary force threats
2	013	acknowledge responsibility	−6.4	17	threaten
1.9	066	release or return	−6.8	2112	guerrilla seizure
1.9	032	travel to meet	−6.8	2111	police seizure
1.6	0933	ask for humanitarian aid	−6.8	21	seize
1.6	0932	ask for military aid	−6.9	183	control crowds
1.6	0931	ask for economic aid	−6.9	1814	protest altruism
1.6	09	request	−6.9	18	protest
1.5	1011	offer peace proposal	−6.9	174	give ultimatum
1.5	101	peace proposal	−7	2231	military clash
1.5	03	consult	−7	195	break relations
1.2	102	call for action	−7	1734	threaten military war
1.1	01	yield	−7	1733	threaten military occupation
1	031	discussions	−7	1732	threaten military blockade
0.8	10	propose	−7	1731	threaten military attack
0.6	012	yield position	−7	173	military force threat
0.6	011	yield to order	−7.6	1827	military border violation

(continued)

TABLE 1. *Continued*

Goldstein	*IDEA*	*Definition*	*Goldstein*	*IDEA*	*Definition*
0.1	091	ask for information	−7.6	1826	military border fortification
0.1	024	optimistic comment	−7.6	1825	military mobilization
0	99	sports contest	−7.6	1824	military troops display
0	98	A and E performance	−7.6	1823	military naval display
0	97	accident	−7.6	1821	military alert
0	96	natural disaster	−7.6	182	military demonstration
0	95	human death	−8.3	224	riot or political turmoil
0	94	human illness	−8.7	221	bombings
0	72	animal death	−9.2	2236	military seizure
0	27	economic status	−9.2	2123	abduction
0	26	adjust	−9.2	211	seize possession
0	25	vote	−9.6	2228	assassination
0	24	adjudicate	−9.6	2227	guerrilla assault
0	2321	government default on payments	−9.6	2226	paramilitary assault
0	2312	private transactions	−9.6	2225	torture
0	2311	government transactions	−9.6	2224	sexual assault
0	231	transactions	−9.6	2223	bodily punishment
0	23	economic activity	−9.6	2222	shooting
−0.1	094	ask for protection	−9.6	2221	beatings
−0.1	022	pessimistic comment	−9.6	222	physical assault
−0.1	021	decline comment	−9.6	22	force
−0.1	02	comment	−10	2237	biological weapons use
−0.9	141	deny responsibility	−10	2235	assault
−1	14	deny	−10	2234	military occupation
−1.1	0631	grant asylum	−10	2233	coups and mutinies
−2.2	192	reduce routine activity	−10	2232	military raid
−2.2	121	criticize or blame	−10	223	military engagements
−2.4	132	formally complain			
−2.4	131	informally complain			
−2.4	13	complain			

Note: IDEA codes and their definitions ordered by level of conflict on the Goldstein conflict-cooperation scale. For more detailed documentation on each category, including full examples and exceptions, see the up-to-date version of IDEA at ⟨http://www.vranet.com/IDEA/⟩.

that drove the construction of WEIS or IDEA, or those embodied in a joint conflict and cooperation scale.

The Reader was originally developed as an extension of Phil Schrodt and his colleagues' Kansas Events Data System (KEDS),[22] and indeed they deserve credit for pioneering this line of research in political science, developing the first working software programs that code news reports, and producing most of the machine-generated events data used in actual substantive research in the field. Although the VRA Reader also owes much to Schrodt's developments, VRA (and Schrodt) report that the two systems—and TABARI, Schrodt's new open-source reader—do not share code. The main differences between KEDS and the Reader are that the

22. Schrodt, Davis, and Weddle 1994.

parsing mechanism in the Reader is more developed and that it natively generates the more detailed IDEA, rather than WEIS, codes.

The VRA Reader is a commercial product, but almost all academic uses of it are free of charge. In addition, the IDEA protocol is fully in the public domain. VRA also routinely makes their compiled modules (and their low-level ActiveX libraries) available without cost to scholars for noncommercial use. VRA has agreed to give us access to the data generated by the Reader from all Reuters news stories for the entire world during the last decade, excepting the most recent year. This is an extremely rich data source that includes approximately 3.7 million individual events. VRA has also agreed to provide updates to these data, as well as refined historical data, as their program improves—contingent only on the continued cooperation of the news reports' publishers. We are arranging to make these data available to the academic community through an online infrastructure being developed by the Harvard-MIT Data Center[23]; this should be complete by the time this article is published. Discussions are also underway between VRA, the Harvard-MIT Data Center, and news publishers to make available the text of all the original news stories, linked to their quantitative codes.

When we examined VRA products and related programs, and thought about undertaking this project, we felt we needed some procedural changes in their basic setup. We therefore asked VRA to reorganize their event hierarchy so that it was fully articulated and exhaustive, more hierarchically structured, and completely documented (with full descriptions, examples, and codes for every category). We asked them to put it on the Web or in some tabular or graphical format so it could be more easily understood by others, and we requested that they write overview documents that would put as much of their knowledge as possible in writing. Our goal was to evaluate only those parts of the VRA system that were fully replicable. We are certain that if VRA personnel trained the human coders, our evaluation of their machine codes would be even more positive, but this is not the kind of independent, replicable knowledge that the scholarly community needs. Thus we wanted to use only those parts of the system that could be represented in some fixed format, independent of specific people who know it well. VRA graciously complied with all of these requests. Although VRA continually updates and improves their event ontology and Reader, we began our experiment with this fixed version of their products and did not consider any further updates or improvements. Of course, this is both a disadvantage—because this article applies directly to only one version of the program, and an advantage—because VRA can use the results of this article to improve the product. The Appendix provides more information about how the VRA Reader and related programs work.

23. See ⟨http://TheData.org⟩. Accessed 14 April 2003. The data are available at http://GKing.Harvard.edu.

In this article, we compare human and machine performance in assigning the event described by a lead into its correct IDEA category, but not their relative performance in identifying the source and target actors. Although automatically determining who the actors are in an event is not trivial, it is a much easier task than determining which of the detailed event types is exemplified in a sentence. Also, the task is made easier by Reuters themselves, because some methods of accessing RBB leads electronically include a list of actors associated with the lead. Consequently, we evaluate only the Reader's categorization of the event type and condition on the source and target actors being correctly identified. We do not formally evaluate any other information generated by the Reader.

A Rare Events Evaluation Design

The Problem and an Overview of Our Approach

Although interest in international relations typically focuses on conflict and violence, most of what happens on the world's stage is neither particularly confrontational nor cooperative. Thus while we may be particularly interested in explicit apologies or threats of force, neutral comments are considerably more common than either in real events data. For example, of the 45,000 events coded by the VRA Reader from news leads on the former Yugoslavia, 10,605 neutral comments were found, versus only four apologies and thirty-five threats of military attack.

To measure the performance of an information extraction system, we would ideally run it over a representative sample of materials whose event structure is known with certainty. This is the standard procedure in the computational linguistics literature, but it may not be feasible for international relations events. The numbers above suggest that a human coder preparing materials for evaluation will have to code, on average, over 2,500 comments to reach an apology, and approximately 300 comments before reaching a single threat of force. This would require a Herculean level of effort from evaluators in order to have enough events to evaluate. Worse, most coding effort would be spent acquiring more and more events in categories on which we already have plenty of evaluative data, and may still result in insufficient representation for substantively interesting events.

One seemingly reasonable, but inadequate, way to address this problem is to use the extraction system itself to perform the initial coding, pick a representative sample covering all event types (or only those of interest) from that, and then examine these instances to see how often the system assigned the correct code. This intuitive approach is feasible, but using it in the obvious way will generate selection bias due to selecting on the dependent variable. To see this, denote M and T as variables indicating into which IDEA category the *M*achine codes an event, and the *T*rue category to which the event actually belongs, respectively. The quantity of interest is the probability that the machine is correct, or in other words $\mathrm{P}(M = i|T = i)$—the probability that the machine classifies an event into category i given that the true coding is indeed in category i. The full characteriza-

tion of the success of the machine requires knowing $P(M = i|T = i)$ for $i = 0, \ldots, J$, which includes all J IDEA categories and where $i = 0$ denotes the situation in which the machine is unable to classify an event into any category. In the version of IDEA we evaluated, $J = 157$, although the categorization has been expanded subsequently. Expressed more simply, the quantity of interest is the full probability density $P(M|T)$.

The problem with this apparently reasonable approach of using the machine to select is that we are conditioning on M. As such, the proportion of events that are actually in category i among those the machine put in category i gives a good estimate of $P(T|M)$, which is not the quantity of interest. It gives us the probability of the truth being in some category instead of the machine. In fact, the truth is fixed, and it is instead the machine that is uncertain—the object of this inquiry, and the variable for which we need to know a probability distribution. Furthermore, $P(T|M)$ is a systematically biased estimate of $P(M|T)$.

Our approach to this problem is to sample in this biased way, but then to use the logic of rare events statistical analyses and data collection designs[24] to correct the problem. This evaluative research design, and the particular correction we propose, has to our knowledge not been used before, but requires only a straightforward application of Bayes's theorem, all component parts of which are fully known or directly estimable. As a result, our estimation involves no modeling assumptions.

We give details of this methodology in the next section and then offer two extensions. In the following section, we discuss the issues involved in summarizing the performance of machine coding. We then, in the next section, extend these methods to evaluating human coders.

The Procedure We Followed

We chose the collapse of Yugoslavia and subsequent conflict over Bosnia as the test domain for our evaluation. We used 45,000 articles from the RBB newswire that included at least one keyword, such as "Bosnia" or "Yugo," designed to capture the actors in the Balkans conflict. The resulting stories include those written in this area, about this area, or about any other country or actor that interacts with countries in this area. We chose to use this selection process rather than the Reader at this stage of data sampling, because we would not have been able to subsequently correct any effects of using it simply as a filter.[25]

24. See Breslow 1996; and King and Zeng 2001a, 2001b, 2002.

25. We began with all English-language Reuters news during 1991–95, eliminating all "factfiles," pictures, and sports. We then specified countries or regions indexed by Reuters and extracted Yugoslavia (which includes all of the former Yugoslavia before 1992; Serbia and Montenegro after 1991; and sometimes refers to the former Yugoslavia post-1991), Bosnia-Herzegovina, Croatia, Macedonia, and Slovenia. Using country codes as we do to extract country information simply means that the country or region in question appears somewhere in the report.

From the VRA Reader's output we estimate $\mathrm{P}(M)$, the marginal probability distribution of IDEA codes assigned by the Reader, from the frequency distribution of categories. For example, one element of this distribution is $\mathrm{P}(M = i)$, the probability that the machine assigns IDEA code i. This is estimated by the proportion of events that are assigned IDEA category i in the data set.

We then randomly chose five news leads from all those events that VRA put in each IDEA category. In thirty-six cases, fewer than five events were available because IDEA is a very detailed coding scheme. We used all that were available in each of these cases. (In only twelve were no true events available to code.) This covers a small fraction of IDEA codes and should not bias our analyses.

Note that although the Reader typically finds more than one event in a news lead, we randomly chose only one so that the sentence exemplified a particular event type. Thus the machine and the coders were evaluated according to their categorization of that particular event, and not any others that they might also find in the sentence.

We added to this collection twenty-five randomly chosen events for which the Reader assigned a source and/or target actor but could not assign an IDEA category. These leads were included to see what events, if any, were being missed by the Reader during normal operation, and whether those constituted a biased sample of the population. This sampling process yielded 711 news leads containing IDEA categories in approximately equal frequency in the sample and a larger set of null responses.

We then recruited eight expert human coders to categorize the 711 leads and achieve a consensus on the correct IDEA categories. Coders were graduate students from Harvard's Government Department (five), Statistics Department (two) and Medical School (one). Each coder was initially given approximately one-eighth of the leads to code. Because all the leads come from the same series of international events, this partitioning should minimize any biasing effects of extended narrative across the leads. In addition, after individual coding had finished, there were multiple rounds of cross-checks, in which each coder checked another's work. Disagreements were resolved by discussion. We continued this process until we were all convinced that we had correctly classified all 711 events.

This process estimates $\mathrm{P}(T|M)$, the probability distribution of true IDEA categories conditional on the Reader's IDEA category assignment. For example, $p(T|M = \text{`02'})$ is the proportion of times that the stories the Reader coded as IDEA code '02', a neutral comment, actually fell into each of the categories, which is not a quantity of interest. Thus we can estimate, without modeling assumptions, $\mathrm{P}(T|M)$ and $\mathrm{P}(M)$. From these, we can get to the quantity of interest via Bayes's theorem:

$$\mathrm{P}(M|T) = \frac{\mathrm{P}(M,T)}{\mathrm{P}(T)} = \frac{\mathrm{P}(T|M)\mathrm{P}(M)}{\mathrm{P}(T)}. \tag{1}$$

where $\mathrm{P}(T) = \sum_{i=0}^{J} P(T|M_i)\mathrm{P}(M_i)$ is a function of the numerator, and thus easily computed from the two quantities we can estimate.

Summarizing Results

For each true category T, our methods enable us to estimate an entire distribution—the probability that the machine classifies an event in each one of the J categories given that the truth is in category T. We can then repeat this analysis for each category T. In other words, because T has J elements and M contains $J + 1$, the sample space for the set of conditional densities in Equation (1) has $(J + 1)J$ (= 24,806) elements. We could validate every one of these categories if we had sufficient events in each, but this is infeasible. Instead, we use the 711 classified true events to characterize the success of the program by summarizing and averaging.

We do this in several ways. First, we present some simple averages of the proportion correct. That is, for a given true category T, we first summarize the univariate density $\mathrm{P}(M|T)$ with the proportion correct:

$$\text{Proportion Correct} = \frac{\sum_{i=1}^{J} \mathrm{P}(M = i|T = i)w_i}{\sum_{i=1}^{J} w_i}. \tag{2}$$

The only question is what the weight, w_i, should be. The choice is purely normative. If one category is more important for some purpose than another, then the evaluation of that category should count more. Because answers to normative questions are neither right nor wrong, we use three separate weighting schemes. For one, we use a constant weighting scheme, $w_i = 1$ for all i [in which case Equation (2) reduces to a simple unweighted mean], because it is in a sense the most obvious. The issue with this approach is that its normative status depends on the categories included in the IDEA framework—those that happened to be of interest to VRA and their academic users and commercial clients. In some areas, such as judicial politics, there exist very few categories; in others, numerous fine-grained categories have been developed.

A constant weighting scheme is not unreasonable, but it is obviously not the only possibility. Thus we also use a weighting scheme based on the frequency of events that actually occur in the world, $w_i = \mathrm{P}(T_i)$. The advantage of this is that it is not a function of the decisions of VRA in defining the categories. The disadvantage of this scheme is that it implies that rarely occurring, highly conflictual categories are of less importance than frequently occurring, routine discussions. In fact, the purpose of our rare events evaluative research design is to guarantee that the rarest categories, many of which are of most interest to political scientists, would be well represented in our study. As such, we also use something near the opposite scheme, $w_i = \mathrm{P}(T_i)^{-1/2}$, which gives the most weight to the categories that occur least frequently.

Our final method of summarizing the success of this program is to use a weighting scheme that is most closely connected to how political scientists have, at least

in the past, been using these data. To do this, we take the *J*-category IDEA framework and map it onto Goldstein conflict-cooperation scores[26] as given by VRA, resulting in a score for each dyad-day ranging from very conflictual (−10) to very cooperative(10) (see Table 1).[27] This means that we need to evaluate Goldstein scores separately from the basic evaluation of the IDEA coding. After all, when two IDEA codes map to the same Goldstein score, a Reader mistake in assigning an event to one code rather than another has no impact on the overall measure. In addition, with this scale score we can evaluate whether misclassifications were "near to" or "far from" the correct category. We therefore use this mapping to judge the severity of Reader errors. In other words, we use the Goldstein scale as a real-valued loss function, expressing the costs to users of various types of misclassification. As will become clear from the results below, our data contained many more conflictual events then cooperative ones. In fact, the most conflictual event type in the data scored −10, whereas the most cooperative event only reached 8.3. The high level of conflict is due in part to our sampling scheme—we tried to take an equal number of events from each category, and IDEA contains more categories for conflictual events—and in part due to the nature of the subject matter, as our leads describe a state of civil war in the former Yugoslavia.

We denote the continuously valued Goldstein conflict-cooperation score for an event in the true IDEA category i as G_i. We also need the average Goldstein score the machine gives for true category i, which we denote g_i and estimate as follows. First, we denote the Goldstein score for true category i and some machine category j as $G_{j|i}$, which of course may take on different values as j changes for any one true category i. We then need the expectation of $G_{j|i}$ with respect to the distribution of all correct ($j = i$) and incorrect ($j \neq i$) positive responses from the Reader j. This distribution is obtained from Equation (1) by computing the density for codes the machine was able to classify:

$$\mathrm{P}(M|T, M \neq 0) = \frac{\mathrm{P}(M|T)1(M \neq 0)}{\mathrm{P}(M \neq 0|T)}. \tag{3}$$

where $1(M \neq 0)$ is an indicator function equaling 1 if $M \neq 0$ and 0 otherwise. Then the expected value we need is simply:

$$g_i = E(G_{j|i}) = \sum_{j=1}^{J} G_{j|i}\mathrm{P}(M = j|T = i, M \neq 0). \tag{4}$$

26. Goldstein 1992.

27. Ibid. Goldstein created scores only for the WEIS categories. The mapping from IDEA to what we call "Goldstein scores" is merely an attempt to put the more detailed IDEA categories on a [−10,10] scale.

Thus with this information, we have a new way of evaluating the machine by simply comparing the average machine score g_i to the true score G_i for any category i.

Finally, these methods also enable us to characterize which sorts of event categories are missed entirely by the Reader. Significant bias could be introduced into analyses based on the data if the machine were more likely to miss cooperative events than conflictual events, for example. To evaluate this aspect of the Reader's performance, we examine the relationship between G_i and $\mathrm{P}(M = 0|T_i)$, the probability the machine does not generate an IDEA code—and therefore a Goldstein score—at all, when the event is truly in category i.

Comparisons with Undergraduate Coders

In addition to the accuracy measures described above, we may also ask what kind of coding the Reader performs. That is, does it perform in a humanlike manner, and are its errors the sort of errors human coders would make? Or does it make judgments that are peculiar to its software implementations? We address this question by looking at the performance of trained undergraduates when presented with event coding for the first time. Although Harvard undergraduates are not necessarily natural choices for performing event data analysis, they provide an useful knowledge baseline.

We presented three undergraduates with documentation provided by VRA and the 711 lead sentences to code. Two students were from psychology and one from history (with courses in government). None had performed similar tasks in the past. In this part of the evaluation, we were interested in how well undergraduate output correlated with Reader output, irrespective of correctness, and also in how accurate undergraduates would be on the task.

Correlation analysis between the machine M and each Undergraduate score U requires the distributions $\mathrm{P}(U|M)$ and $\mathrm{P}(M)$. These can be computed using the same logic above, after substituting undergraduate codes for true codes. Evaluating accuracy is also not difficult. We require $\mathrm{P}(U,T)$, from which we can easily acquire $\mathrm{P}(U|T)$. However, we cannot simply count the proportion of times each undergraduate assigns a lead to category i when it is in fact in category i, because this ignores the fact that we have sampled the leads themselves using the machine, and must therefore condition on M. On the other hand, we do have access to the relevant conditional distribution $\mathrm{P}(U,T|M = i)$. This is the distribution of undergraduate and true categories, conditioned on the fact the Reader assigns IDEA code i. We can then compute $\mathrm{P}(U, T)$ as a weighted average of these distributions:

$$\mathrm{P}(U,T) = \sum_i \mathrm{P}(U,T|M = i)\mathrm{P}(M = i). \tag{5}$$

and where $\mathrm{P}(U|T)$ is obtained by marginalization.

Results

Proportion Correct

We begin presenting our results in Table 2, which gives the overall proportion correct for the Reader in various ways. These are interesting in and of themselves, but they also allow us to compare the proportion correct for the machine (M) to that for each of our three undergraduate coders ($U^{(1)}$, $U^{(2)}$, and $U^{(3)}$). The results are presented for all the IDEA codes (in the left set of columns) and for the WEIS codes in isolation (in the right set of columns). They are presented with three different weighting schemes [$w = 1$, $w = \mathrm{P}(t)$, and $w = 1/\sqrt{\mathrm{P}(t)}$], and for the original set of 157 IDEA codes (marked "detailed") as well as for only the top level or "cue" IDEA categories (marked "aggregate").

For example, when using constant weighting ($w = 1$), and detailed event codes, the Reader places an event in the correct one of the 157 categories 26 percent of the time (see the number in the upper left of the table). Then, reading across, we can see that the three undergraduates were correct, computed in the same way, 32 percent, 23 percent, and 26 percent of the time, respectively. For this particular method of evaluation, the reader thus falls squarely within the range of our human coders.

When we move from this panel in Table 2 to others, we see some major changes. But the changes are all in the *absolute* levels of performance. For example, as would be expected, the proportions correct for the aggregate categories are higher than the detailed categories, and higher for WEIS categories than for all the categories. The former is true as errors at the detailed level that fall within a single

TABLE 2. *Comparing proportion correct: Machine versus human coders*

	All codes				*WEIS codes*			
	M	$U^{(1)}$	$U^{(2)}$	$U^{(3)}$	M	$U^{(1)}$	$U^{(2)}$	$U^{(3)}$
$w = 1$								
detailed	.26	.32	.23	.26	.25	.44	.25	.37
aggregate	.55	.55	.39	.48	.62	.62	.48	.62
$w = \mathrm{P}(t)$								
detailed	.52	.48	.35	.42	.55	.64	.35	.68
aggregate	.65	.70	.53	.64	.70	.72	.56	.65
$w = 1/\sqrt{\mathrm{P}(t)}$								
detailed	.36	.44	.33	.41	.37	.62	.34	.67
aggregate	.59	.66	.49	.62	.64	.68	.53	.63

Note: Cell entries are the estimated proportion correct for the machine (M) and for each of three trained undergraduate coders ($U^{(i)}$, $i = 1, 2, 3$), with different weightings (w) and codings.

aggregate CUE category are not counted as errors at this aggregated level. The latter is presumably true because international conflict scholars know more about coding international conflict than other parts of this ontology. The absolute levels also differ substantially across weighting schemes (with constant weighting receiving the lowest score). But although the absolute level of performance changes with the normative evaluation criteria, the *relative* performance of the machine and the undergraduates does not. Throughout all these statistics, the key finding is that the performance of the machine is indistinguishable from our human coders.

Proportion Coded

The statistics reported in Table 2 only cover events for which the subject (either the machine or a human coder) was sufficiently certain to be able to code an event. We therefore also need to see the success the subject has in finding events when there are events or deciding that there is no event when there is no event. Summary statistics for each of these situations (and both together) are given in Table 3.

When an event exists in a story, the machine finds it 93 percent of the time, and the undergraduate coders find it 94 percent, 80 percent, and 90 percent of the time, respectively. By this criteria, too, the machine's performance falls right in the middle of the performance of the undergraduates. We also evaluated the ability of the machine to recognize when no event existed. Here, it diverged from the undergraduates significantly, correctly classifying nonevents as nonevents only 23 percent of the time, whereas the undergraduates did much better, with performance ranging from 92 to 100 percent. This is the only failure of the machine we were able to identify in relation to our undergraduate human coders; the result is that any quantitative data set coded by this software will have some observations added to the data that are really not events. We give evidence below that these extra "events" are not more likely to appear in some categories than others, and are thus unlikely to bias any analyses based on these data. Furthermore, relatively few news stories are completely contentless like this, which is why the last row in Table 3 more closely resembles the first than the second. This does not, therefore, appear to be a major issue.

TABLE 3. *Proportion of events classified correctly*

	M	$U^{(1)}$	$U^{(2)}$	$U^{(3)}$
Actual events	.93	.94	.80	.90
Actual nonevents	.23	1.00	.97	.92
Total	.85	.94	.82	.90

Note: Cell entries in the first row are the proportion of events classified correctly for news stories containing events; cell entries in the second row are the proportion of events classified correctly for news stories not containing events; and the third row is the total for both.

Thus the implication of this difference between human and machine coding is that the machine is basically throwing some randomly generated "observations" into our data set that do not belong there, but because these data are unrelated to any measured variable, they should not bias any subsequent inferences. The inefficiency that results would be a concern if the machine and the undergraduates worked at the same rate, because including random observations is equivalent to discarding some observations with real information. However, the machine can code millions of news stories in the time it takes a human being to code only a few, and if we recognize this difference in the abilities, the machine is equivalent in terms of bias and far better in terms of efficiency.

Evaluation by Degree of Conflict and Cooperation

In this section, we evaluate the machine and undergraduate coders by weighting events according to the IDEA categories mapped into the Goldstein conflict and cooperation scale. We are thus comparing the estimated true score, g_i to the machine or undergraduate's classification, G_i. Both of these are on a $[-10,10]$ scale.

Figure 1 gives what might be the most obvious presentation, plotting G_i horizontally by the machine's g_i vertically. The graph has one point for each of the event categories that truly occurred in the data and for which the machine generated an IDEA code.[28] If g_i were known exactly, the best situation for the machine would be if all the points in this graph fell exactly on the 45° horizontal line. However, g_i is not known exactly, and instead is estimated by our rare events evaluative design, with only about five evaluations in each event category; that is, each circle in the graph is estimated on the basis of only about five observations. Thus deviations from the line are due to both (random) estimation error—because our research design uses only a finite number of stories—and pure machine error (as is fully summarized in Tables 2 and 3). The results in Figure 1 show that the points cluster fairly closely around the 45° line. This is further strong evidence in favor of the VRA Reader when evaluated according to this criterion.

For the most part, the deviations around the line in Figure 1 are random, but one small pattern can be seen. This is on the left side of the graph: in the region where $G_i < -6$, the machine seems to fall consistently above the 45° line, indicating that when categorizing very conflictual events, it slightly underestimates how conflictual they are. This pattern can be seen somewhat more clearly in the upper left graph in Figure 2, where we have plotted the true Goldstein score by the deviations. In this graph, the ideal situation would be for the points to fall randomly around a horizontal line at $G_i - g_i = 0$. The points (and the nonparametric-smoothed regression line that we included to highlight the pat-

28. The figure includes 130 rather than 157 conditional distributions because twelve categories did not in fact occur in the data—although the Reader generated them—and a further fifteen were left uncategorized.

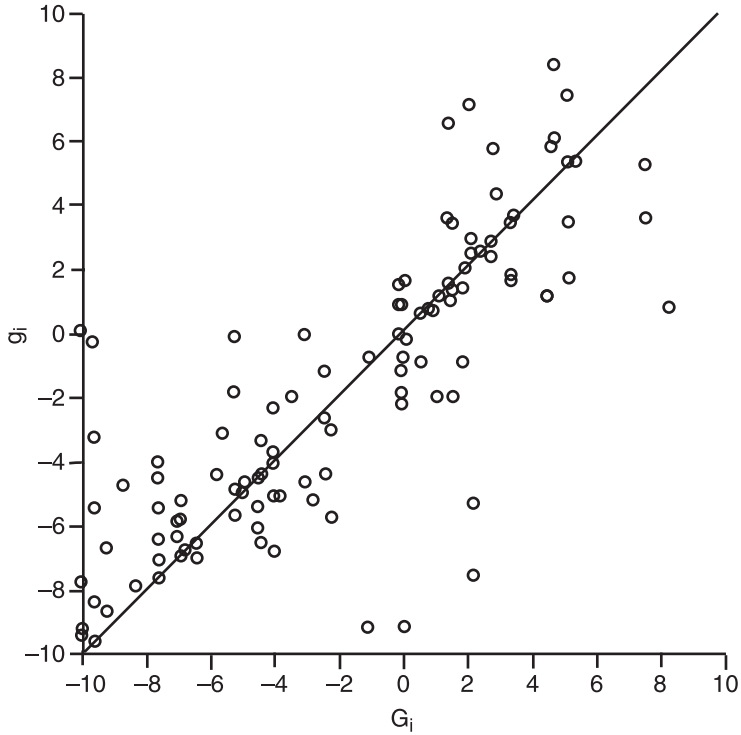

FIGURE 1. *Goldstein scores: true scores (plotted horizontally) by scores estimated on the basis of the machine coding (plotted vertically)*

terns) show this ideal pattern, with the exception of the area on the left of the graph representing the high-conflict categories. The bias there is not large, but it does exist.

The remaining three graphs in Figure 2 give analogous summaries of the performance of our three undergraduate coders. There is some variation in the pattern across the three human coders, but the same overall pattern can be seen. That is, the key result in Figure 2 is that the systematic pattern in the errors made by the machine is also evident in the performance of the human coders. That is, both the machine and the human coders are about as likely to code highly conflictual events as slightly less conflictual than they are. Of course, the variation across the human coders is a disadvantage, because it means that results of human coding are not perfectly reliable and replicable, whereas the machine gives the same result every time, and results are completely replicable no matter who is running the program.

We also wanted to see whether this systematic error pattern might be correctable by statistical analyses. For this to be the case, the error pattern had to be

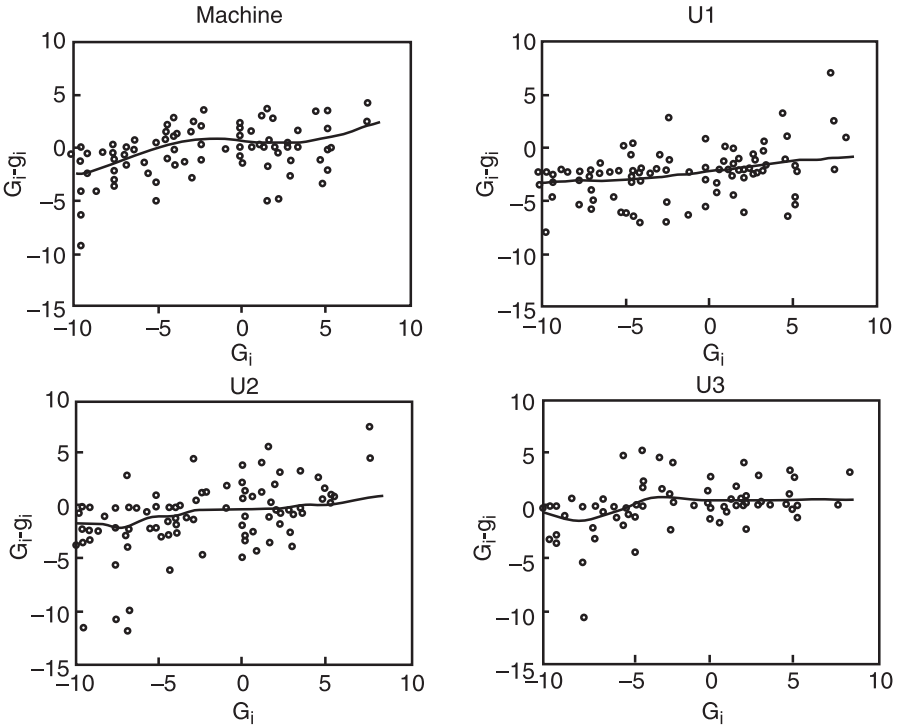

FIGURE 2. *True conflict-cooperation scores (plotted horizontally) by errors (plotted vertically). The machine's performance is in the top left graph and the human coders are in the other three graphs.*

related to the machine (or undergraduate) codes, rather than only to the true event category. Unfortunately, as Figure 3 demonstrates, the estimated codes are not more likely to be biased in one direction or another, and so are not easily correctable by, for example, making the conflictual categories a bit more conflictual. Figure 3 plots the estimated Goldstein score by the error level, but unfortunately reveals no systematic pattern.

Finally, we also ascertain the probability of finding no event as a function of the true event category. If a coding method systematically produces more null codes for some true categories than others, then a form of selection bias would be introduced into any analyses based on the coded data. Figure 4 gives our results, again for the machine in the upper left corner and for our three undergraduates in the other three positions. The results in Figure 4 reveal no systematic patterns in null coding as a function of the event category, for either the machine or the undergraduate coders.

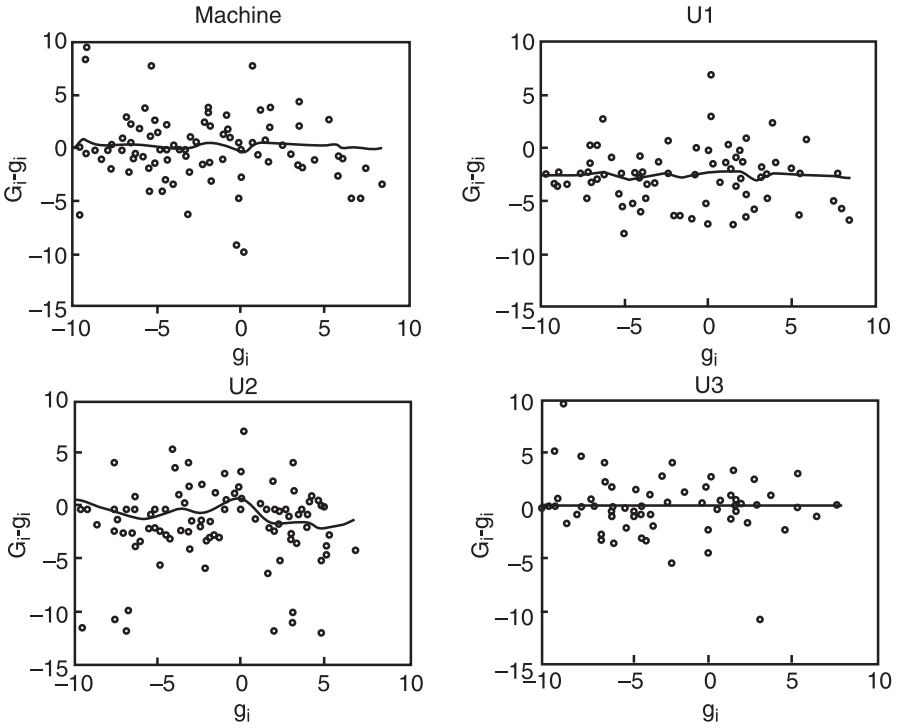

FIGURE 3. *Estimated conflict-cooperation scores (plotted horizontally) by errors (plotted vertically). The machine's performance is in the top left graph and the human coders are in the other three graphs.*

Conclusions

In our view, the results in this article are sufficient to warrant a serious reconsideration of the apparent bias against using events data, and especially automatically created events data, in the study of international relations. If events data are to be used at all, there would now seem to be little contest between the machine and human coding methods. With one exception, performance is virtually identical, and that exception (the higher propensity of the machine to find "events" when none exist in news reports) is strongly counterbalanced by both the fact that these false events are not correlated with the degree of conflict of the event category, and by the overwhelming strength of the machine: the ability to code huge numbers of events extremely quickly and inexpensively.

Although the machine performed approximately equally to our trained human coders in this study, the machine would be far better over the long run. Hiring

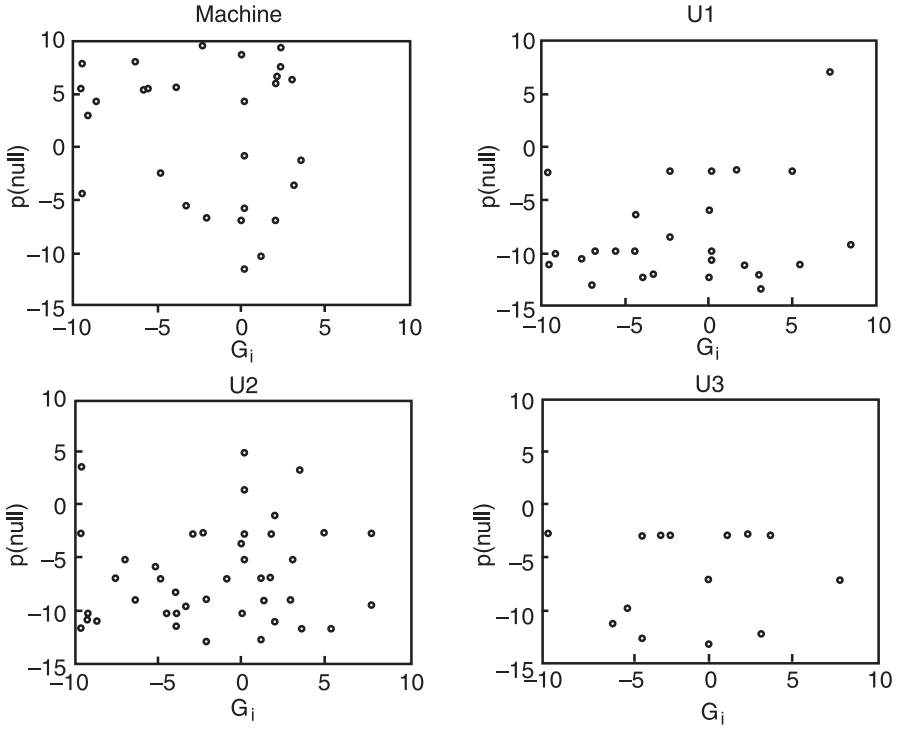

FIGURE 4. *True conflict-cooperation scores (plotted horizontally) by the probability of finding no event (vertically). The machine's performance is in the top left graph and the human coders are in the other three graphs.*

people of the quality we were able to recruit to code many more events than we asked of them is probably infeasible, and doing so for the many years it would take to do this right would undoubtedly reduce performance to levels significantly below that of the machine. Longer-term coding by human coders would result in lower performance, either because we would have to resort to using less-qualified coders or because their attention to the extremely tedious and boring task would wane over time.

Further study is needed to ascertain the precise selection mechanisms involved in using news services to represent actual events in international relations, but it seems infeasible to design a data collection procedure for academic purposes that would be remotely as comprehensive as any major news service. This article provides some reason to be optimistic about our ability to mine this critical information source.

Appendix: The VRA Reader and Related Work in Information Extraction

Because few political scientists have had much experience with this area of research, we describe in this appendix how the VRA Reader and related programs work, and how our article also contributes to the scholarly field of information extraction. We begin with an overview.

Information extraction[29] is a subfield of computational linguistics, the branch of computer science that studies machine processing of natural language.[30] (The field occupies the intersection between linguistics, the study of the form and function of natural languages, and computer science, which is concerned with any kind of data representation and processing that can be described algorithmically and implemented on computers.) Information extraction is a constrained form of natural language understanding in which only prespecified information is acquired from textual data, often by filling in a template. Extraction can be distinguished from the easier task of information retrieval, where, in response to a user's query, the machine tries to return all documents that contain relevant information without specifying what that information is, and the much harder task of text summarization, where there is no prespecified form describing what information is sought and the machine attempts to discover what is relevant and return a precis of the document.

An important development in the field of information extraction occurred in the 1980s when the U.S. Defense Advanced Research Projects Agency organized a series of Message Understanding Conferences (MUC).[31] These tested extraction software (through a series of controlled contests between research groups) on out-of-sample news leads and also helped develop methods of testing. The literature includes analyses of texts about international conflict, such as the third MUC, which was devoted to news leads about Latin American terrorism. However, in this conference, the others in the series, and in most of the literature, input text is carefully chosen to exemplify a very narrow range of subject matter. For example, the set of event types covered in most of the MUC conferences is equivalent to no more than a handful of IDEA categories. The event types studied were also of relatively equal prevalence, and so existing approaches in this literature have not been tested on rare events data, such as the international conflict events studied in political science. The literature also does not include evaluations or methods for evaluation that involve rare events of considerable interest, which we must have for real-world analyses of international conflict, and which we introduce in this article.

Information extraction systems, such as the VRA Reader, are typically organized into three processing stages: tokenization and lexical processing, syntactic processing, and domain analysis.[32] Tokenization and lexical processing involve segmenting words (which is easy for English but difficult for Chinese), part-of-speech tagging (where each word is assigned a grammatical category—proper noun, verb, etc.), and word-sense tagging (where the machine attempts to determine whether "bank" refers to the financial institution or the riverside. Syntactic analysis, or parsing, expresses the grammatical structure of the sen-

29. See Cowie and Lehnert 1996; and Grishman 1997.
30. See Jurafsky and Martin 2000 for a recent overview.
31. See Sundheim 1991, 1992; and Grishman and Sundheim 1996.
32. Appelt and Israel 1999.

tence in a representation that makes its interpretation clearer. For example, even a very crude syntactic analysis of the following sentence can produce two very different analyses:

Border guards saw the lone gunman with a telescopic sight.

The different analyses are illustrated in Figure 5. In this figure, the overarching category is the sentence (S). The sentence is composed of two parts, the subject noun phrase (NP) and a verb phrase (VP). The verb phrase itself decomposes into a verb (V) and an object (NP). PP denotes a prepositional phrase; the PP can be further decomposed into the preposition 'with' and the NP "a telescopic sight," though we do not show the full analysis here.

In the tree structure A1, the PP "with a telescopic sight" relates to the NP "the lone gunman" because they are both subparts of the higher-level noun phrase structure denoted conventionally as NP′. In this analysis the verb takes two arguments, the NP "border guards,"

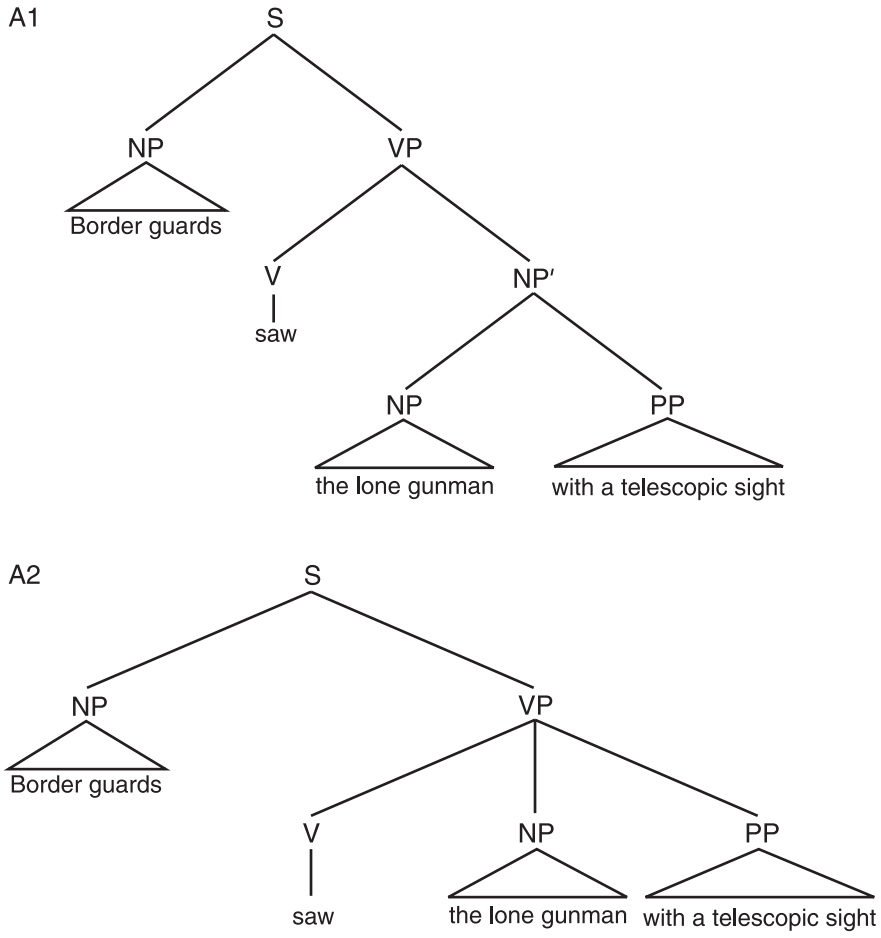

FIGURE 5. *Syntactic analysis*

and NP' "the lone gunman with a telescopic sight." In A2 however, the PP is now a third argument of 'see' and affects our interpretation of the verb itself because it is no longer contained in an overarching NP' structure. This simple difference in syntactic decomposition expresses the fact that in the first analysis, it is the gunman that has the telescopic sight, and in the second analysis, it is the border guards that use the telescopic sight to see the gunman.

Obviously, no machine or human could distinguish interpretation A1 from A2 without additional information such as that from surrounding sentences, but contrast this example with the following superficially similar sentence:

Border guards saw the lone gunman with a bandana.

Although to the human eye it is immediately obvious that this sentence has no similar ambiguity, an extraction system must know a great deal about clothing, human actors, and vision to infer correctly that the second analysis is probably inappropriate for this sentence. Some of this disambiguating knowledge is contained in the lexical semantic database, WordNet.[33] For example, WordNet represents each noun in the database in terms of hyponymy relations:

architect → creator → person → life form → entity.
bandana → handkerchief → piece of cloth → fabric → artifact.

The notation → should be read as "is a kind of." The information that an architect is a person whereas a bandana is an artifact may be important for syntactic analysis. For example, it would tell us whether a verb, such as "see," requires an animate subject. The Reader makes extensive use of WordNet categories, and VRA has changed and augmented the database to cover specialized words relevant to international relations.

Many extraction systems, including the Reader, perform a "full parse" on every sentence that provides considerably more information than the very crude bracketing shown above. Much of formal—that is noncomputational—linguistics is devoted to discovering appropriate rules and structures for appropriate sentence decomposition, and there is a wide range of possible theories. Any full syntactic analysis will specify a complete hierarchical decomposition of the sentence with bracketing information down to the word or even to the subparts of words; for example "guards" might be decomposed into its stem "guard" and the English plural "s." Conversely, the diagrams in Figure 5 show that at the highest level, the sentence can be broken into a noun phrase "Border guards" and a verb phrase "saw the lone gunman with a telescopic sight." For this simple sentence, there is little higher-level structure, but for news leads that often span fifty words or more, the presence of subclauses, complements, and nested quotes entail a significant processing burden, and an increased risk of ambiguity and bracketing errors at these higher levels of analysis. In some situations, however, simply distinguishing subjects, verbs, objects, prepositional phrases, and reported speech is sufficient to discover basic information about who did what to whom.

Domain analysis is the final step, which includes resolving coreference ambiguity using domain information to disambiguate alternative syntactic analyses. Coreference ambiguity occurs when the same entity is referred to in several different ways, and is a particular problem with names. As an example of the problems involved, consider the following (fictional) text that is quite unambiguous to humans:

33. Fellbaum 1998; see ⟨http://www.cogsci.princeton.edu/~wn/⟩. Accessed 14 April 2003.

> General Electric announced a third quarter loss, claiming they will perform significant restructuring. GE has made similar claims before but it may need to follow through with it this time.

An extraction system must know that "General Electric" is a company, not a military officer. It must also infer that "GE" refers to the same thing as "General Electric." Moreover, it must also infer that "GE" and "General Electric" have the same referent as the first "it," but that the second "it" refers to the restructuring process. These corefering expressions are more common the longer the article, because authors intentionally try to employ a variety of referring expressions to keep the reader's attention. This practice makes for text that is much more comfortable for humans and much harder for machines to read.

References

Appelt, Douglas E., and David J. Israel. 1999. Introduction to Information Extraction Technology. A Tutorial Prepared for IJCAI-99. Available at ⟨http://www.ai.sri.com/~appelt/ie-tutorial/IJCAI99.pdf⟩.

Azar, Edward E. 1982. *Codebook of the Conflict and Peace Databank*. College Park: Center for International Development, University of Maryland.

Bond, Doug, J. Craig Jenkins, Charles L. Taylor, and Kurt Schock. 1997. Mapping Mass Political Conflict and Civil Society: Issues and Prospects for the Automated Development of Events Data. *Journal of Conflict Resolution* 41(4):554–79.

Bond, Doug, Joe Bond, J. Craig Jenkins, Churl Oh, and Charles L. Taylor. 2001. Integrated Data for Events Analysis (IDEA): An Event Form Typology for Automated Events Data Development. Unpublished manuscript, Harvard University, Cambridge, Mass.

Breslow, Norman E. 1996. Statistics in Epidemiology: The Case-Control Study. *Journal of the American Statistical Association* 91 (433):14–28.

Cowie, Jim, and Wendy Lehnert. 1996. Information Extraction. *Communications of the ACM* 39 (1):80–91.

Davies, John L., and Chad K. McDaniel. 1994. A New Generation of International Event-Data. *International Interactions* 20 (1–2):55–78.

Fellbaum, Christine, ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

Gerner, Deborah J., Philip A. Schrodt, Ronald A. Francisco, and Judith L. Weddle. 1994. Machine Coding of Event Data Using Regional and International Sources. *International Studies Quarterly* 38 (1):91–119.

Goldstein, Joshua S. A. 1992. Conflict-Cooperation Scale for WEIS Events Data. *The Journal of Conflict Resolution* 36 (2):369–85.

Goldstein, Joshua S. A., and Jon C. Pevehouse. 1997. Reciprocity, Bullying and International Conflict: Time-Series Analysis of the Bosnia Conflict. *American Political Science Review* 91 (3):515–29.

Goldstein, Joshua S. A., Jon C. Pevehouse, Deborah J. Gerner, and Shibley Telhami. 2001. Reciprocity, Tringularity, and Cooperation in the Middle East, 1979–97. *Journal of Conflict Resolution* 45 (5):594–620.

Grishman, Ralph. 1997. Information Extraction: Techniques and Challenges. In *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, edited by Maria Teresa Pazienza, 10–27. Berlin: Springer Verlag.

Grishman, Ralph, and Beth Sundheim. 1996. Message Understanding Conference 6: A Brief History. In *Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING-96)*, edited by Ralph Grishman and Beth Sundheim, 466–71. Copenhagen.

Jurafsky, Daniel, and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, N.J.: Prentice Hall.

King, Gary, and Langche Zeng. 2001a. Explaining Rare Events in International Relations. *International Organization* 55 (3):693–715.

———. 2001b. Logistic Regression in Rare Events Data. *Political Analysis* 9 (1):137–63.

———. 2002. Estimating Risk and Rate Levels, Ratios, and Differences in Case-Control Studies. *Statistics in Medicine* 21 (10):1409–27.

Laurance, Edward J. 1990. Events Data and Policy Analysis: Improving the Potential for Applying Academic Research to Foreign and Defense Policy Problems. *Policy Sciences* 23 (2):111–32.

Leng, Russell J., and J. David Singer. 1988. Militarized Interstate Crises: The BCOW Typology and its Applications. *International Studies Quarterly* 32 (2):155–73.

McClelland, Charles A. 1978. *World Event/Interaction Survey (WEIS) Project, 1966–1978*. Ann Arbor, Mich.: Inter-University Consortium for Political and Social Research.

Merritt, Richard L. 1994. Measuring Events for International Political Analysis. *International Interactions* 20 (1–2):3–33.

Most, Benjamin A., and Harvey Starr. 1984. International Relations, Foreign Policy Substitutability, and "Nice" Laws. *World Politics* 36 (3):383–406.

Rummell, Rudolph J. 1975. *The Dimensions of Nations*. Beverly Hills, Calif.: Sage.

Schrodt, Philip A. 1995. Event Data in Foreign Policy Analysis. In *Foreign Policy Analysis: Continuity and Change in its Second Generation*, edited by Laura Neack, Patrick J. Haney, and Jean A. K. Hay, 145–66. Englewood Cliffs, N.J.: Prentice-Hall.

Schrodt, Philip A., and Deborah J. Gerner. 1994. Validity Assessment of a Machine-Coded Event Data Set for the Middle East, 1982–92. *American Journal of Political Science* 38 (3):825–54.

———. 2000. Cluster-Based Early Warning Indicators for Political Change in the Contemporary Levant. *American Political Science Review* 94 (4):803–18.

Schrodt, Philip A., Shannon G. Davis, and Judith L. Weddle. 1994. Political Science: KEDS—A Program for the Machine Coding of Event Data. *Social Science Computer Review* 12 (4):561–88.

Sowa, John F. 1999. *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Pacific Grove, Calif.: Brooks Cole.

Sundheim, Beth. 1992. Overview of the Fourth Message Understanding Evaluation and Conference. In *Proceedings of the Fourth Message Understanding Conference*, edited by Beth Sundheim, 3–22. San Mateo, Calif: Morgan Kaufmann.

Sundheim, Beth, ed. 1991. *Proceedings of the Third Message Understanding Conference*. San Mateo, Calif.: Morgan Kaufmann.

Taylor, Charles Lewis, Joe Bond, Doug Bond, J. Craig Jenkins, and Zeynep Benderlioglu Kuzucu. 1999. Conflict-Cooperation for Interstate and Intrastate Interactions: An Expansion of the Goldstein Scale. Paper presented at the 40th Annual Convention of the International Studies Association, February, Washington, D.C. Available at ⟨http://www.ciaonet.org/isa/trc01/⟩.