COMPUTER GENERATED EVENTS DATA FROM ONLINE SOURCES:

"SOME ASSEMBLY REQUIRED"*

Devika Subramanian
Department of Computer Science
devika@rice.edu

Richard J. Stoll
Department of Political Science
stoll@rice.edu

Note: both authors are also affiliated with the Center for the Scientific Study of International Conflict and Cooperation and the Computer and Information Technology Institute

March, 2004

Rice University
Box 1892
Houston, TX 77251-1892

Paper prepared for the Annual Convention of the International Studies Association, Montreal, Quebec, March 17-20, 2004.

***The research described in this paper was funded by the National Science Foundation, award EIA-0219673.**

ABSTRACT

This paper discusses a portion of our research that is designed to predict the outbreak of serious international conflict. This research uses events data constructed from online media sources. This paper is a progress report on the development of REDE, the Rice Events Data Extractor. This is a set of computer programs to extract events data from online news accounts. These programs are based on research and ideas from the discipline of artificial intelligence. While REDE is not yet a reality, we believe that it will be a valuable tool for researchers who wish to develop their own events data collections.

INTRODUCTION

Events data are a potentially rich source of data for the testing of a wide variety of models of comparative foreign policy. We use the word "potentially" because to make use of these data researchers need to be able to generate appropriate data for the states and the time frame that is of interest to them. Fortunately, we are in an era when it is becoming increasingly possible to generate events data with computer programs, rather than the human coding that was used to generate the original events datasets. In this paper, we describe our efforts to develop REDE, the Rice Events Data Extractor. It is a computer program to code events data from online news sources. This is part of a larger effort at Rice University to predict the onset of serious interstate conflict through the use of computer-generated events data.

EVENTS DATA: RATIONALE, HISTORY AND LIMITATIONS

One of the pioneers of events data, Charles McClelland, described the significance of events data for the study of international relations as follows:

> International conduct, expressed in terms of event data, is the chief dependent variable of international relations research. ... [This] starting point is provided as readily by the ordering principle of classical diplomatic history as by the basic concepts of general system analysis. Thus, we may assert that the prime intellectual task in the study of international relations is to account for actions and responses of states in

> international politics by relating these to the purposes of statecraft or, alternatively, we can say that the problem is to account for the relations among components of the international system by analyzing the characteristics of the various components of that system by tracing recurring processes within these components. [Both definitions] carry about the same information and involve nearly the same range of choices of inquiry and analysis.
>
> (McClelland, 1970: 6)

Even if one does not fully agree with McClelland, it is easy to see that events data can be extremely useful. In international conflict research, the two most frequently studied dependent variables are war and the outbreak of serious conflicts short of war.[1] This focus is understandable because these variables tap extremely serious behavior by states. But if one wants to study less extreme forms of foreign policy behavior (either as the dependent variable of a study or as an independent variable), then events data are essential.

Many scholars have at least a passing familiarity with the two most well known events data collections, COPDAB (the Conflict and Peace Data Bank; see Azar, 1980), and WEIS (the World Events Interaction Survey; see McClelland, 1976). While there are many differences between these two data sets, the manner in which both were collected is very similar. A large number of well-trained coders were needed to read through news accounts and code events. This type of data collection was only feasible if the researcher had enough resources to train coders and then employ then over an extended period of

[1] The most commonly used war dataset is the one developed by the Correlates of War Proejct. The most commonly-used dataset that taps serious conflicts below the level of war is the Correlates of War Project Militarized Interstate Dispute (MID) dataset. Both are documented and available at http://cow2.la.psu.edu/.

time. Consequently, the major events data collections ended when it was no longer possible to obtain large research grants to continue these projects.

The major events data collection projects left a concrete and useful legacy: the datasets they produced. But there are limitations to these data. One limitation is the time span; COPDAB data go from 1948 to 1978, and WEIS (at least the original collection) run from 1966 to 1978. The limited time spans meant that some questions (for example, changes in state behavior after the Cold War) cannot be examined. As well, the coding schemes used in both COPDAB and WEIS may not be appropriate for every research agenda.

COMPUTER GENERATED EVENTS DATA

Three recent developments offer the prospect of allowing individual researchers to collect – even design - their own events data. One development is the accessibility of online media reports. The second is availability of programming tools to create events data. The third is the development of new schemes to organize and classify events data.

An increasing number of media outlets maintain websites that allow access to their news stories. In some cases, not only are current issues available, but there are also archives which can be accessed. While some of these media outlets require fees to obtain access (some which are very expensive), many can be accessed at no cost – although some require that a scholar register to obtain complete access (for example, registration is necessary to obtain extensive access to *The New York Times* website). We should note that media outlets that maintain websites are not restricted to the United States or even

the western world. We believe that this trend will continue, and more and more current (and archival) media sources will be directly accessible to scholars.

The second development is the growth of computer programs to extract events data from electronic news sources. Some have been developed primarily for commercial reasons (for example, the VRA® Knowledge Manager system; see Bond et al.., 1997, or the Virtual Research Associates (2003a) website: http://vranet.com/index.html). But there are also publicly available systems that any academic can obtain. The most prominent example of publicly available systems are TABARI and its predecessor, KEDS, developed at the University of Kansas by Phil Schrodt and his colleagues (see Schrodt, 2001, or the KEDS (2003) website: http://www.ukans.edu/~keds/project.html). Consequently, the availability of computer programs makes events data collection by single scholars or small groups feasible.

Finally, recent years have seen the development of new events data coding schemes. One example of such a new coding scheme is PANDA (the Protocoal for the Assessment of Nonviolent Direct Action) and its successor IDEA (Integrated Data for Events Analysis; see Virtual Research Associates, 2003b). A second example is CAMEO (Conflict And Mediation Event Observations; see Gerner, Schrodt, Yilmaz, and Abu-Jabr, 2002). Researchers may find that one of these newer schemes is more suitable to their needs than the "traditional" schemes of WEIS or COPDAB.

As can be seen, these three developments work together to open up new vistas for research. For example, researchers can add to existing events data collections, perhaps even extending them back through earlier periods of time. Another possibility is that a researcher could develop her own event data coding scheme, and collect events data

based on that scheme. Undoubtedly scholars will come up with even more creative possibilities.

THE RICE EVENTS DATA EXTRACTOR (REDE)

We have a large scale project to predict the onset of serious international conflict (see Subramanian and Stoll, 2002a, 2002b). As part of this project, we will be collecting a great deal of events data from multiple online media sources. As part of this process, we have decided to build our own event data extractor: the Rice Events Data Extractor (REDE). This will be built on different principles than TABARI.[2] We feel that the research community will benefit from having choices for event extraction. We believe that some researchers will find that TABARI is the superior tool for their needs, while others will turn to REDE. We begin by a brief discussion of the underlying principles of TABARI and then discuss our plans (and progress to date) for REDE.

TABARI

The development of TABARI is an impressive achievement. It has allowed researchers at Kansas and elsewhere to generate and/or use new versions of events data. TABARI (Textual Analysis By Augmented Replacement Instructions) is the open source C++ successor to KEDS. The program, documentation, and source code can be download from the KEDS website (KEDS, 2004a; see also Schrodt, 2001). Since it is

[2] We have at least one thing in common with TABARI. When REDE is finished, it too will be made available to anyone who wants to use it.

described there as well (KEDS, 2004b), we will offer only a brief description to highlight the differences between TABARI and REDE.

TABARI (like its predecessor KEDS) relies on sparse parsing (Schrodt, 2001: 6). That is, rather than fully parsing a sentence, TABARI determines whether there is a codeable event in the sentence. TABARI looks for verb phrases (using its verb dictionary) which reflect the categories of event that are in the event scheme (for example, the event categories in WEIS). It then uses a variety of heuristics (along with its noun dictionary) to identify the actor and target of the event.

The strengths and weaknesses of TABARI lie in its simplicity. It relies heavily on its verb and noun dictionaries. These dictionaries are built by hand (i.e., by researchers themselves). And it is necessary to adjust them if the researcher moves from one spatial-temporal domain to another. Building or modifying these dictionaries can be very laborious. But this is very important; if appropriate items are not included in the dictionaries, then events will be missed. As well, TABARI has some difficulties if the structure of the sentence is not straightforward.[3] This may not be a serious drawback if the source of the events data is written in a clear and straightforward fashion (as we might expect with test written for a newswire).

REDE

Our approach is more complicated than that used by TABARI. Note that complication can be both a good and a big thing. In principle, REDE should have more

[3] Our experiments with TABARI indicate that it has trouble if the sentence is written in passive voice or contains a number of verbs.

capabilities than TABARI. But it also may have more problems. As well, TABARI has one very clear advantage over REDE: it can be downloaded and run today, while REDE is still under development.

There are two phases to REDE. The first is a filtering phase. During this phase, we seek to eliminate text that does not contain potential events. The second is the event extraction phase. In this phase, text that is likely to contain events is examined, and if events are found, they are coded. Like TABARI, we will have to develop dictionaries to identify actors, targets, and event types. As well, at several points in the process it is necessary to train parts of the program so it can learn to identify various elements important to the identification of events.

Building Dictionaries for Actors, Targets, and Events.

We will identify a list of nouns (or noun phrases) and verbs (or verb phrases) that represent the actors, targets, and event types of interest. But we hope to do this without resorting to extensive manual coding. We will make use of techniques and procedures from natural language processing (NLP).

To develop the list of actors (nouns and noun phrases), we will rely on techniques developed in the field of named entity recognition. This is a well-developed field of study, although there has been little work in this field on identifying the names of countries and political organizations that serve as actors and targets for event data. The approach involves building statistical models (usually at the word or phrase level) from training data in which actors are marked up. The problem is usually modeled as a

sequential learning task in which each token (word) in the text is to be labeled as part of a named entity or not. Factors that are used to make this decision include the capitalization of the token, the part of speech tags of a window of tokens before and after the token to be labeled, whether or not tokens before and after the current one are labeled as named entities, etc. (Zhang and Johnson, 2003). The probabilistic model learned from data is of the form Prob(factor|token is named entity) and Prob(factor|token is not a named entity), for each factor. Then Bayes rule is used on unseen documents to predict the most likely label for a token given the factors associated with that token.

Traditionally, the naïve Bayes assumption is made so that factors are treated conditionally independent of one another given the classification. The best named entity recognition performance is obtained by combining feature-based methods with dictionaries of known named entities. Another combination approach involves building classifiers that learn different probabilistic models from the same training data (e.g., conditional random fields (McCallum and Li, 2003) and using weighted voting to compute the final label (Florian, Ittycheriah, Jing and Zhang, 2003). Such methods are able to identify actors in text with a false positive rate of 20 percent and a false negative rate of 20 percent. We are extending these techniques and specializing them to the domain of political events and aim to improve the accuracy of these methods event further. At this point we expect we will still need manual tweaking of the resulting list to eliminate residual errors, but are very hopeful that the bulk of the development of the noun (actor and target) list can be done in an automated fashion.

In order to identify events we also need to develop a list of verbs and verb phrases. Unfortunately unlike the identification of appropriate noun and noun phrases, in

which we can build upon the research on named entity recognition, there is no corresponding field of study which can be used to identify verbs (and verb phrases). However, we believe a statistical machine learning approach can be used to build the verb dictionary.

We will mark up a large number of news items (about a year's worth of stories from a source), some of which contain events that we wish to code, and some of which do not contain events. We will look for verbs and verb phrases, and calculate the probability that particular verbs and verb phrases are associated with particular events. This approach has a good chance of working if there are only a limited number of ways in which events are described.

Filter Phase.

Although we are exploring various options, we believe that we will restrict the input to one or more of the following elements of a news story: (a) the headline, (b) the first sentence of the first paragraph and/or (c) the sentences of the first paragraph.[4] It is our experience examining stories from a variety of news sources that if there are codeable events in a story, they occur near the beginning of the story.

Regardless of the unit to be examined (hereafter "unit" will refer to the particular element - headline, lead sentence, first paragraph – that is to be examined for codeable events), the first step is to determine if the unit is likely to contain a codeable event. Our goal is to filter out many false units (i.e., units that contain no codeable event) as

[4] REDE will have the ability to work with a variety of units. In our own research we will settle on one particular unit to code.

possible. The filter is also built around a naïve Bayes model (i.e., the same sort of model that we will use to identify phrases for the dictionaries). But in this case, we look for words that have a high probability of being in units that contain codeable events, and words that have a low probability of being in units that contain codeable events. The words and their calculated probabilities are then used to filter units. Only those units that successfully pass through this filter are examined for codeable events.

Coding Phase.

The coding phase begins with part of speech tagging of the unit. Part of speech tagging is a well developed area of natural language processing (NLP).[5] Given the part of speech markup, we will identify which nouns (and noun phrases) and verbs (and verb phrases) are potentially parts of events to be coded. This will be done by matching nouns (and noun phrases) and verbs (and verb phrases) to the lists (dictionaries) we have generated.

Like sparse parsers, part of speech tagging can have problems with passive voice. One problem with most forms of parsing is that there is a certain amount of ambiguity in their results. That is, the same sentence can be parsed in more than one way. But a stochastic parser (Charniak, 1997) will indicate the most likely parse of the sentence. We believe that using this type of parser will significantly improve our ability to extract events from news stories.

Tools.

[5] In particular, Brills' parser (Satta and Brill, 1996) is considered to be state of the art.

We are not building our events data generator from scratch. We are using GATE (2004) to develop our software. GATE provides basic software components for the processing of human language. GATE provides a number of tools written in Java. Some can be used as stand alone programs, while others can be used as components of user-written programs.

THE CURRENT STATUS OF REDE

REDE is very much a work in progress. As is typical of programming projects, we have not made as much progress as we would have wished. Nevertheless we have moved forward and believe that REDE will become a reality in the not too distant future. What follows is a status report on REDE.

Filter

The current version of the filter has been tested on a set of 6543 stories from AP in 1998. We selected only stories concerning the Middle East, using a search command similar to that used by KEDS to build its Gulf dataset.[6] These stories were then tagged by

[6] As shown on the KEDS website (KEDS, 2004c), the NEXIS search string used to identify source texts for their Gulf data set was: (SAUDI! OR SAUDI ARABIA! OR IRAN! OR IRAQ! OR KUWAIT! OR GCC OR OMAN! OR YEMEN! OR QATAR! OR BAHRAIN! OR UAE OR EMIRATE! OR DUBAI! OR ABU DHABI!) AND NOT (SOCCER! OR SPORT! OR OLYMPIC! OR TENNIS OR BASKETBALL OR NBA OR T.STRM OR HEADLINE(HIGHLIGHTS OR (WORLD W/2 OUTLOOK) OR (KEY W/1 FACTS) OR (EVENTS W/1 SCHEDULED) OR (HISTORICAL W/1 CALENDAR)))

political science graduate students, placing each story into one of three categories: story contains one or more codeable incidents, story does not contain codeable incidents, and questionable (the final category was used if the student could not decide whether or not the story contained one or more codeable incidents). Ignoring the third category for the time being, about 10 percent of the stories were tagged as containing one or more codeable incidents, and the remaining 90 percent were tagged as containing no codeable incidents.[7]

These stories were then run through the filter program and the results were compared with the tagging of the graduate students. The filter performed well on this set of stories. Of the stories tagged as codeable by the graduate students, the filter only missed two percent. Of the stories tagged by the graduate students as not codeable, only 10 percent pass through the filter and are classified as codeable.

We find these results to be very encouraging. Of course this is only a single test. We will tag other sets of stories and run the filter program to get a broader assessment of its ability to discriminate stories that contain codeable events from those that do not. Our next step is to run the filter on AP stories from 1999. This is important because different media sources can have different writing styles, and we need to be sure that the filter will work with a variety of sources. Ideally we would like to use a single filter program, but we are prepared to "tweak" the program to optimize it for different sources and create different versions if that is necessary.

There are also ways to modify the filter program that we believe will improve its performance. Currently the program only ignores a few types of words. The list is

[7] Although we have not yet done a systematic analysis of the substance of these stories (we intend to do this analysis in the near future), these stories tend to be in one of the following categories: sports, opinion, business, human interest, domestic politics, and domestic economic.

restricted to articles, contractions, and a few very common verbs. Judicious expansion of this list could help the performance of the filter, although we do not expect this to lead to a dramatic improvement.

Currently the words that we use to classify stories are not ranked. But some words are more likely to discriminate stories with codeable events from those that do not have codeable events. The performance (and speed) of the filter is likely to improve if we rank words by their ability to discriminate stories with codeable events from those without codeable events. Note that a word could be highly useful in discriminating stories either if it is strongly associated with those that contain codeable events, or with those that do not contain codeable events.

There are a variety of ranking schemes that could be used and we will experiment to determine which scheme (or schemes) provides the best performance. One standard criterion is to use an information-theoretic measure called information gain to rank words. A word with high information gain is one that is much more likely to be found in a story that contains a codeable event than in a story with no codeable events. There are other statistical criteria for word selection in text classification that we will explore in our implementation efforts.

Parser

To this point most of our programming effort has gone into the development of the filter. But we have made some progress on the parser. We have developed an interface to the WordNet (2004) lexical database. This is a widely used database of

English nouns, verbs, adjectives and adverbs. Accessing it allows us to get definitions, synonyms, antonyms, and identify the part of speech of a word. This will aid us in determining both the meaning of a word or phrase and its part of speech.

While WordNet is an invaluable resource for identifying single words, the correct identification of actors and event types often involves identifying phrases. This involves the realm of named entity recognition. There are a variety of methods that can be used to do named entity recognition. A recent competition at the conference on computational natural language learning (coNLL 2003) ran a named recognition competition involving twelve different methods. Two of these – one discussed in Florian, Ittycheriah, Jing and Zhang, 2003 and the other discussed in Chieu and Ng 2003 -- significantly outperformed the remaining ten. We will take advantage of the results of this competition and use the more successful methods of named entity recognition in our own work.

One significant open problem that we will have to handle is the appearance (and disappearance) of relevant nouns and noun phrases. For example, until recently references to "Aristide," "President Aristide," "Jean Bertrand Aristide" or similar phrases would refer to the president of Haiti. So sentences that had one of these nouns or noun phrases as the subject or the object would indicate a potential event to be coded. But all that changed on February 29th when he went into exile. In some cases there will be "clues" to the change of status (i.e., "former President Jean Bertrand Aristide"). But correctly interpreting the meaning of these qualifiers may be difficult, and in other cases such clues may be lacking. Consequently we believe it will be necessary to put researchers into the loop to monitor the data collection process, paying particular attention to the emergence of new actors and the "disappearance" of others.

The Cost-Benefit Calculus: How Complicated are the Sentences?

As we have noted, REDE involves a more complicated set of methodologies than those used in TABARI. Assuming that we are able to fully implement what we have described, will it be worth it to make available an alternative tool to TABARI? In one sense the answer is yes. There is benefit to providing additional options for researchers. But how important is this?

The utility of REDE turns on the sentence construction of the media source to be used by the researcher. If the text in the source consists predominantly of simple declarative sentences, this is an optimal environment for TABARI. If the typical sentence in the source is significantly more complicated than that, we believe that REDE will be a better choice. So what is the mix of simple and complicated sentences in online news sources? At this point we have no answer to that question. We are just beginning an automated empirical analysis of sentence structure using Charniak's (1997) stochastic parser to determine this percentage. A manual survey of New York Times stories about the Middle East gathered during March 2002 reveals that simple subject-verb-object constructions form about 15-20 percent of the total sentences in these stories. If the percentage of simple sentences in the sources we seek to code remains at this level, alternative coders such as REDE would be quite useful indeed.

CONCLUSION: SOME ASSEMBLY *STILL* REQUIRED

Events data are a potentially rich source of information for the study of international relations. These data afford an opportunity to study a wide variety of foreign policy behavior. We believe that, with the advent of new events data coding schemes, online media archives, and the development of computer tools to automate the coding of events data, we are entering an era in which individual researchers can design and collect their own events data.

Our contribution to this era is REDE, the Rice Events Data Extractor. The design of REDE builds on a number of areas of artificial intelligence. It is still under development. But we have made progress. We believe that when completed, it will offer an additional tool for those who wish to develop their own events data. As well, it will be an integral part of our own efforts to understand and predict the outbreak of serious international conflict.

REFERENECES

Azar, Edward. 1980. The Conflict and Peace Data Bank (COPDAB) Project. *Journal of Conflict Resolution* 24: 143-252.

Bond, Doug, J. Craig Jenkins, Charles L. Taylor and Kurt Schock. 1997. "Mapping Mass Political Conflict and Civil Society: The Automated Development of Event Data" *Journal of Conflict Resolution* 41,4:553-579.

Charniak, Eugene. 1997. Statistical Parsing with a Context-Free Grammar and Word Statistics. Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97), pp. 598-603.

Chieu, Hai Leong and Hwee Tou Ng. 2003. Named Entity Recognition with a Maximum Entropy Approach. *Proceedings of CoNLL-2003*, Edmonton, Canada, pp. 160-163.

Florian, Radu, Abe Ittycheriah, Hongyan Jing and Tong Zhang, 2003. Named Entity Recognition through Classifier Combination. *Proceedings of CoNLL-2003*, Edmonton, Canada, pp. 168-171.

GATE. 2004. General Architecture for Text Engineering. http://gate.ac.uk/. Accessed 2/15/2004.

Gerner, Deborah, Phil Schrodt, Ömür Yilmaz, and Rajaa Abu-Jabr. 2002. Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for a Post Cold War World. Paper presented at the 2002 Annual Meeting of the American Political Science Association.

KEDS. 2003. The KEDS Project. http://www.ukans.edu/~keds/project.html. Accessed 2/15/2004.

KEDS. 2004a. TABARI Program and Source Code. http://www.ukans.edu/~keds/tabari.html. Accessed 2/15/2004.

KEDS. 2004b. TABARI: Text Analysis By Augmented Replacement Instructions. http://www.ukans.edu/~keds/tabari.info.html. Accessed 2/15/2004.

KEDS. 2004c. Gulf Data Set. http://www.ukans.edu/~keds/gulf.html. Accessed 3/12/2004.

McCallum, Andrew and Wei Li. 2003. Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. *Proceedings of CoNLL-2003*, Edmonton, Canada, pp. 188-191.

McClelland, Charles A. 1970. Some Effects on Theory from the International Event Analysis Movement. mimeo, University of Southern California, February, 1970.

McClelland, Charles A. 1976. *World Event/Interaction Survey Codebook.* (ICPSR 5211). Ann Arbor: Inter-University Consortium for Political and Social Research.

Satta, Giorio and Eric Brill. 1996. Efficient Transformation-Based Parsing. 1996 Conference of the Association for Computational Linguistics. http://www.cs.jhu.edu/~brill/Eff_Pars.ps. Accessed 2/15/2004.

Schrodt, Philip and Deborah Gerner. Forthcoming. Analyzing International Events Data: A Handbook of Computer-Based Techniques.

Schrodt, Philip. 2001. Automated Coding of International Event Data Using Sparse Parsing Techniques. Paper presented at Annual Meeting of the International Studies Association, Chicago, IL, February 2001.

Subramanian, Devika and Richard Stoll. 2002a. Events, Patterns, and Analysis: Forecasting International Conflict in the Twenty-First Century. http://www.cs.rice.edu/~devika/projects/stoll.html. Accessed 2/15/2004.

Subramanian, Devika and Richard Stoll. 2002b. Events, Patterns, and Analysis: Forecasting International Conflict in the Twenty-First Century. Position paper presented at the workshop "Computer-Aided Methods for International Conflict Resolution" held at the Austrian Research Institute for Artificial Intelligence, Vienna, Austria, October 25-26, 2002.

Virtual Research Associates. 2003a. http://vranet.com/index.html. Accessed 2/15/2004.

Virtual Research Associates. 2003b. I.D.E.A. http://vranet.com/idea. Accessed 2/15/2004.

Wordnet. 2004. Wordnet: A Lexical Database for the English Language. Princeton, NJ: Cognitive Science Laboratory, Princeton University. http://www.cogsci.princeton.edu/~wn/. Accessed 2/15/2004.

Zhang, Tong and David Johnson. 2003. A Robust Risk Minimization based Named Entity Recognition System. *Proceedings of CoNLL-2003*, Edmonton, Canada, pp. 204-207.