New Computer Science Applications in Automated Text Identification and Classification for the Social Sciences

A Workshop at Penn State University, August 15-17, 2007

Hosts:
Frank R. Baumgartner, Distinguished Professor of Political Science, frankb@psu.edu
John McCarthy, Professor and Head of Sociology, jxm516@psu.edu

New text-based data resources and tools being developed within the computer and information science communities provide many possibilities for new social science applications. Scholars can envision systematic data of greater consistency, flexibility, historical coverage, and depth of information than previously possible. However, the diffusion of new techniques from the computer science to the social science community has been slow. Social science applications offer important theoretical challenges to computer scientists as the specific variables of interest to political scientists, sociologists, and others differ from other fields. Social scientists have been developing large databases at a rapid pace in recent years. The large established human-coded databases now in existence provide important means by which we can develop and calibrate new computer-based data development tools. Tighter collaborations across these intellectual communities may thus lead to important theoretical and infrastructure advance in both areas.

Our workshop brings together leading computer scientists with political scientists and sociologists with extensive experience in creating large-scale databases. Social scientists will have the opportunity to learn of the latest computer science research relevant to their needs and the computer scientists will learn of the special problems associated with historical research on public policy, social movements, and legislative activities.

Contents:

Note: This workshop is by invitation only. Penn State University has provided funds for domestic travel and local housing expenses for those invited. Baumgartner and McCarthy have submitted a proposal to the National Science Foundation for additional funding which would allow a wider range of participants to be involved. If you are interested in attending or know of people doing research in the area who should be involved, please contact the hosts at the email addresses listed above.

Conference Schedule

(Draft February 12, 2007)

August 15, 2007

Opening Session: 7:00-8:30pm

Welcome and Orientation: Frank Baumgartner and John McCarthy, Penn State
Participant Introductions and Brief Project Descriptions

August 16, 2007

Continental Breakfast: 8:15-9:00

Frontiers of Text Identification and Retrieval: 9:00-10:30

Jamie Callan, Carnegie Mellon University

Break: 10:30-10:45

Automated Classification Systems: 10:45-12:15

Stephen Purpura, Harvard University

Lunch: 12:15-1:30

Frontiers of Automated Text Coding: 1:30-3:00

Eduard Hovy, University of Southern California

Break: 3:00-3:15

Measuring Citizen Input, a Computer Science Political Science Collaboration: 3:15-4:45

Stuart Shulman, University of Pittsburgh, Eduard Hovy, and Jamie Callan

Poster Sessions: 5:00-6:00

Opportunity for Participants to provide more details about large scale Text Projects (e.g. Agendas Projects in various countries, ARDA, social movements and protest projects.)

Dinner: 7:00-9:00

August 17, 2007

Continental Breakfast: 8:15-9:00

Automated Coding of Texts into Unknown or Dynamic Classification Schemes: 9:00-10:30,

Burt Monroe, Penn State

Break: 10:30-10:45

Implementations and Future Collaborations: 10:45-12:30

General Discussion of Future of Network and Potential Collaborations
Led by Frank Baumgartner and John McCarthy, Penn State

Project Description

We propose to organize a conference at Penn State to bring together prominent social scientists currently in the forefront of large-scale and hands-on data collection projects with computer scientists actively working at the frontiers of automated text retrieval, classification, and natural language processing. The social scientists stand to gain in efficiency in completing their future large data collection efforts; the computer scientists, in being presented with new challenging theoretical problems of retrieval and classification that may lead to new technological innovations with potentially wide applications. The many and diverse large existing social science databases that scholars participating in the conference bring to the table for discussion will provide the computer scientists the opportunity to refine, calibrate, and validate their several computer science applications.

Introduction

Exciting developments have taken place over the past decade in the development of vast new text-based databases for the study of a variety of questions across the social sciences, including our own efforts in public policy, social and religious movements, and social protest events. Through the painstaking efforts of dozens of scholars world-wide, we have seen the development of new data resources previously not imagined. The Policy Agendas Project (www.policyagendas.org), for instance, houses data including virtually every congressional action since World War Two (including all bills, laws, and hearings) as well as information on public opinion, presidential activities, Supreme Court decisions, and the federal budget. With the click of a mouse, students, scholars, and members of the public can trace government attention literally to hundreds of public policy issues ranging from water pollution to trade with China. In Sociology, scholars of social movements at Penn State and elsewhere have created new text-based data resources as well, by identifying newspaper reports of protest events and coding event details allowing analyses of protest over long time periods and the linking of protest events to congressional action (e.g., McAdam and Su 2002; Earl et al. 2005) as well as the expansion of the national population of advocacy groups (on-going research by Baumgartner and McCarthy under NSF award SBR–0111611). Penn State sociologist of religion Roger Finke directs the Association of Religion Data Archives (http://www.thearda.com/) which brings together information about religious organizations of all kinds in America and across the globe and is currently assembling text-based data to analyze religious freedom across most nations of the world (Grim and Finke, forthcoming). Text-based datasets have also been developed by sociologists study the evolution of culture (Peterson and Arnaud 2004), ethnographies of labor conflict (Roscigno and Hodson 2004), the framing of social problems (Snow, forthcoming), the dynamics of ethnic conflict (Stathan et al., 2005), and media portrayals of ethnic stereotypes (Gilens 1999). In political science, international relations scholars have developed the large Correlates of War databases (http://cow2.la.psu.edu/); Phil Schrodt's long-standing automated efforts to track newspaper coverage on-line associated with Middle East and other regional rivalries (see for example Schrodt and Gerner 1994); few political scientists have developed systematic ways of using large amounts of media coverage information to study the dynamics of how issues are framed, but one Penn State graduate student is doing exactly that, with NSF support (Boydstun 2007). There is a lot of social science progress but most of this has been done

using very expensive human-coding techniques. This is both a problem and an opportunity for reasons we explain below.

New databases like these are not limited to the United States. Teams of scholars in Canada, Denmark, Belgium, France, the Netherlands, England, Scotland, Spain, Italy, Switzerland, and in the European Union have already begun or are in the process of developing Policy Agendas data bases. And, scholars in Germany, the Netherlands, France and Switzerland have developed and are developing extensive protest event data bases. Many of these national projects have the cooperation and active collaboration of national governments as well as funding from national science agencies. In the Netherlands, the Royal Archive has proposed digitizing the nation's entire record of newspapers, going back to the 1600s as well as the entire legislative record, similarly going back hundreds of years. They seek an academic partner to help categorize and systematize a searchable database of potential use to a wide audience. Google is moving forward with a plan to digitize all congressional hearings, going back to the founding of the Republic. Through our connections with the Library of Congress, we hope to link this Google-led effort to existing Policy Agendas databases on bills, laws, and hearings. In addition, the State of Pennsylvania has sponsored a project to make available the full text of all bills and laws and many legislative authorized studies and reports, as well as abstracts of legislative hearings, executive orders, and state Supreme Court decisions, through a new Policy Database web site on which we are collaborating, linking the site and its classification scheme to the Policy Agendas Project. This will effectively make the site a new public portal to access official legislative documents in digitized form.

Of course, commercial and other databases are typically far more advanced than those generated by social scientists. Vast stores of newly digitized records are continually becoming available through commercial vendors. Considerable work in computer science is focused on these commercial applications. To give one example, web trolling software can allow the manufacturer of a commercial product, say an automobile, to find every comment or review relating to a certain vehicle or model. Once these are identified, the software can determine what characteristic of the car is being discussed: is it the engine, the seat belts, the price, dealer service, expected longevity, crash safety, handling, or what? Finally, the software can determine whether the comments are praising the product's performance in that area or damning it. The value of such an automated text retrieval and analysis system for a manufacturer of commercial products is obvious, and these tools are being developed by many computer scientists.

Social scientists, in spite of their progress in developing new databases, work in almost complete ignorance of the new tools and therefore suffer tremendous disadvantages in terms of efficiency and labor costs. The policy agendas database, for example, has relied almost exclusively on time-consuming and expensive human coding for each and every record in the database; hundreds of thousands so far. And, the Dynamics of Protest project, headed by Susan Olzak, Sarah Soule, Doug McAdam and John McCarthy (supported by NSF), used human labor to read *New York Times* daily editions cover to cover for the 1960 to 1995 period in order to identify protest event stories. Then, the details of each event were coded by teams of research assistants.

The Identification and Classification Problems

Some examples of the kinds of technologies / solutions / applications we have in mind are as follows: Recently, with the help of computer scientists Steven Purpura and Dustin Hillard, major steps have been made toward building computer algorithms to read the full text of digitized documents and to "learn" the complex classification system that is at the heart of the Policy Agendas Project. This classification system identifies approximately 230 distinct topics of government activity, with 20 major topics such as agriculture, defense, energy, and transportation each broken down into a number of more precise subtopics. This system, originally developed in the US, has proved remarkably resilient in international expansions.[1] In a recent test, Purpura and Hillard "taught" their program the content coding system by having it read 100,000 congressional bills along with the topic codes that project workers had assigned. Using various algorithms focusing on the patterns of shared usage of language, the program then generated its own codes, proving to be roughly as accurate as the human coders themselves. Further, the system generates an estimate for each item that its proposed code is correct, based on the degree to which the statistical profile of the words used in the text is clearly identifiable with one and only one subtopic. Most texts have extremely high probabilities of being coded correctly, with a minority generating ambiguous results. The computer algorithm thus may allow human coding resources to be devoted only to that minority of categories where they are needed most. Initial results suggest that the program works well with an initial "seed" of only 10 to 20 percent of the text coded, and that it is over 80 percent accurate in assigning classifications to the remaining items.

There are two fundamental issues associated with computer assisted coding of this type. First is the *classification* (or coding) problem, discussed above. This focuses on knowing, for a given document, what it is about, once a class of documents (such as hearings or bills) has been identified. Through the large-scale comparison of tens of thousands of documents, it appears from the Purpura-Hillard work that this problem can be solved effectively by systematically comparing the patterns of occurrence of words and phrases. At a minimum, the technology promises to reduce the cost of creating an accurate classification system by orders of magnitude, though human coders will still be involved both in generating the "seed" and in checking / revising the original results. Of course, we do not know yet how well this system will work: 1) in other languages; or 2) using data sources where the amount of text is more limited than in the case of US legislation. Will it work well based on only short abstracts as are available for parliamentary questions, for example? Does it work just as well based on abstracts of newspaper articles as on their full text? In sum, a number of extensions are possible and their feasibility as yet is unknown.

[1] In constructing similar projects in other countries, typically fewer than 20 of the total of 230 subtopics have had to be significantly revised, eliminated, or created from scratch. Some US-based policies simply do not exist in other countries (such as our extensive federal networks of public lands and land irrigation systems, Indian affairs questions, or affairs related to the District of Columbia). Similarly, policies such as direct administration of health care services by the national government common in Europe have no place in the US coding system. For over 90% of the classification system, however, the US system is directly transferable to each system where it has been attempted.

The Purpura / Hillard approach to the classification system differs from traditional approaches in computer science since it is based on re-creating a fixed, human-designed, classification system. Other approaches focus on shared word usage inductively to devise the subject categories. Both approaches certainly will have broad application in the social sciences if the tools can be developed and made more widely available. To return to the example above concerning the commercial manufacturing company searching for all comments or reviews on a given product, one can easily see an application in the social sciences relating, for example, to a given political issue. If the issue is the war in Iraq, what aspect of the war is getting attention? Is that focus changing over time? Are people saying positive things or focusing on problems? If the issue is new energy technologies that may be adopted in response to global warming, how much are we discussing nuclear, hydro, biomass, wind, solar, and other technologies, and what types of things are we saying about them? Many political scientists have used simple Lexis-Nexis searches to address these questions, often developing a list of keywords in a seat-of-the-pants manner. More sophisticated research tools could turn this into a major research field. We could know, in real time, the nature of public debate on any given topic and how it is changing over time. Computer scientist Ed Hovy is a leading scholar in the area of natural language processing and has agreed to present at the proposed conference.

The second fundamental issue is that of *identification*. This is of particular interest in the use of media studies, where the problem of locating stories on a given topic requires an enormous investment in human labor. For example, in previous work on social protest activities in the US as well as in Germany, keywords and electronic search terms proved unable to easily identify those stories discussing protest events. Students were hired and scanned the *entire record* of the newspapers looking for stories related to protest events. Developing better technologies for the identification of relevant articles has many possibilities as it would allow one to study vast quantities of information on whatever topic was of interest. This might be particular international events, such as war or armed conflict, energy or global warming, or protest events as in the above example. Computer scientist Jamie Callan is a leading scholar on issues of information retrieval, assessing word count patterns, and evaluating the tone and valence of comments. His current research with Stuart Shulman involves assessing the content of hundreds of thousands of public comments submitted to US federal agencies as part of the "notice and comment" process on proposed rulemaking. Broader applications of these technologies can be very useful.

The scholarly literature that addresses both problems simultaneously is known in computational linguistics as *topic detection and tracking* or *topic modeling*. Here, there is no pre-existing classification scheme. The required approaches draw from the subfield of statistics of *unsupervised learning*, or learning without a teacher. Leading examples outside of social science include the work by David Blei (Princeton) and John Lafferty (Carnegie Mellon), including application to the classification of scientific abstracts, and David Newman, Caitanya Chemudugunta, and Padhraic Smyth (all of UC-Irvine), including application to the classification of newspaper articles. Political scientists Burt Monroe (Penn State), Kevin Quinn (Harvard), and Michael Colaresi (Michigan State), along with computer scientist Dragomir Radev (Michigan), have, under National Science Foundation funding (BCS-0527513), developed such a topic modeling approach for the automated coding of legislative speeches. This work was awarded the 2006 Gosnell Prize for Excellence in Political Methodology.

Computer scientists have developed many tools that are currently being used in government and in the commercial realms. If the technology can be developed to deal with the more complicated problems social scientists deal with, which may require more complicated identification and learning algorithms than those previously developed for commercial product applications, then the work of the computer scientists will be enhanced as they work to develop adaptations to the new theoretical projects provided to them by the social scientists, and social scientists will certainly benefit from the potential of the powerful new tools that may be developed. But the two groups of scholars need to come together in order to assess their mutual and interacting theoretical and technological needs.

Political scientists and sociologists have long been interested in these issues and substantial communities of scholars are actively studying each of these questions internationally. The organizers of this conference are in close contact with many of these scholars. Within the community of information and computer science, the classification and identification problems are also the object of considerable work and expertise. However, until recently these diverse communities have not been brought together. The proposed conference thus has as its primary aim bringing together scholars working on the technical problems of classification and identification with political scientists and sociologists who are developing substantive data bases relevant to citizen mobilization and governmental responsiveness as well as the dynamics of media framing or public issues and problems.

Structure of the Conference

Invited conference participants have been identified on the basis of their record of assembling large text based data sets for analysis. The structure of the conference is aimed at facilitating the sharing information about these manifold resources among social scientists participants and between them and the computer science participants. The basic format includes technical presentations by computer scientists followed by extensive periods for discussion among all participants and the two groups of scholars interact to pursue common intellectual goals.

Accordingly, PIs have scheduled just five presentations, each on a related topic and each with 90 minutes for an overview and substantial discussion and Q&A period. Presentations will range from specific methodologies and specific projects to more general reviews of the state of the art in computer science applications in the areas of information retrieval and natural language processing. We want to know what are the tasks that can be accomplished with high reliability (but which may no longer be interesting research problems for the computer science community) and what is the current frontier of research. The social scientists we have invited and intend to invite have long and deep experience in large scale data collection, and our plan is for the discussion sessions to focus on areas of possible future collaboration, not only with the presenters, but with computer scientists in other countries as well. Our computer science invitees are well networked with European colleagues and we hope to build a network of collaborations. Presentations will be as follows.

Jamie Callan, of the Language Technologies Institute, Carnegie Mellon University, will present on "Frontiers of Text Identification and Retrieval" focusing on the current state of the art in this area. Callan's background is in Information Retrieval and Machine Learning. His recent IR research addresses automatic database selection, high speed adaptive information filtering, algorithms that learn information needs by observing user actions, novelty detection, automatic analysis of gathered information, and question answering. He also studies how industry and government apply information technology to solve "real world" information and knowledge management problems (recent relevant publications include Collins-Thompson and Callan 2007, Shulman et al. 2006, and Sahoo et al. 2006).

Stephen Purpura, Harvard University, will follow with a presentation of his automated classification system and its adaptation to the pre-existing Policy Agendas Project topic system. Purpura was previously a software developer for many years and has more recently come to political science and has been closely involved in efforts to automate the coding of the Congressional Bills Project; this has broad potential applicability for any fixed classification scheme with a "seed" of human coded materials and text; see Purpura and Hillard 2006).

Eduard Hovy, Information Sciences Institute (ISI) of the University of Southern California, will address the state of the art in manual text coding, focusing on the research required to use such coding to produce material that will allow computer systems to learn to perform the same coding automatically. He will use several ongoing projects at ISI and elsewhere, including the OntoNotes project (Hovy et al. 2006) to provide examples of the problems involved.

Stuart Shulman, a University of Pittsburgh political scientist, has been collaborating with Hovy and Callan on a multi-year NSF-supported project building tools to assist federal regulation writers who must evaluate hundreds of thousands of written public comments made in the rule-making processes. Since 1999, Dr. Shulman has acquired a total of 15 datasets comprising over 900,000 public comments (including both electronic and paper submission media). These comments have been made available to the wider research community via an eRulemaking Testbed, which is a web site hosted by Carnegie Mellon University (see also Shulman 2007). Shulman will present an overview of their approach to the issue and a summary of the coding activities conducted under the auspices of the Qualitative Data Analysis Program (QDAP), which he directs at the University of Pittsburgh. His discussion will emphasize the role of manual coding as the bridge between the disciplines and its role as a gateway to future interdisciplinary efforts. Tools developed for eRulemaking, will be refined and made available to the broader community. Shulman hosted a conference on a similar topic to this one at Pittsburgh in 2006, with NSF support. Our conference includes some overlapping participation with that one and presents an extension and continuation of the collaborations begun there. The focus of the earlier conference was on automated categorization schemes, a topic we will also discuss. However, our focus is broader, dealing with identification issues, fixed classification schemes, and unsupervised learning as well. Further, we have organized our conference more about the computer science ideas rather than the applications. However, we should make clear that we see our conference as an important next step in a process of collaboration that was already begun with NSF support in 2006.

Burt Monroe, Penn State, with colleagues Quinn and Colaresi has collaborated on a large NSF-supported project to develop corpora of legislative speech in many countries, and to develop statistical methods for identifying topics, position-taking, and other features of interest to scholars of democratic representation. He will present an overview of their approach and discuss the pros and cons of unsupervised approaches relative to both supervised learning and conventional dictionary-based machine-coding and information retrieval (see Quinn et al. 2006).

Finally, we plan to invite a wide range of political scientists and sociologists, from a variety of fields. Rather than have each one make a substantial presentation, we propose that each will send us in advance information about their data collection projects including their own ideas of how the process might be improved, made more efficient, and the problems they foresee in automating their work. We will turn each of these into a standardized poster-sized presentation for each project, and we will have these posters displayed around the meeting room throughout the conference. In this way all participants can have an understanding of the applied research projects, their problems, scope, and current status. Baumgartner and McCarthy will make short introductions to each of these projects at the beginning; there will be an opportunity to mingle and review the posters; and the questions and discussion during each of the presentations will focus on what technologies and strategies can be applied to what research problems. We have scheduled each of the computer scientists for a relatively long presentation, 90 minutes each, and have scheduled no time at all for the participants to discuss their own projects since the large number of participants makes this impossible in the time available. We want the focus to be on the social scientists learning about the technology, and the discussion to be addressed to issues of application to the various research problems of the several projects. The intense interest among invitees in learning more about these technologies convinces us that this format will encourage lively discussion and interchange.

Follow-up and Subsequent Collaborations

PIs want to ensure that the community of social scientists we assemble becomes integrated with computer scientists. This will help each group intellectually and increase the likelihood of future collaborative projects and each participants own chances of being able to adapt and adopt some of the new technologies in their own subsequent work. As indicated above, we believe we also have significant and interesting research problems to offer to future collaborations with the computer scientists. To facilitate continuing interactions, Penn State will support two quarter-time graduate assistants during the 2007–08 academic year, one in political science and one in computer science, to focus on developing new tools and to work with our participants to help them develop connections and to explain their research problems in a manner of interest equally within the two professional communities. Our goal is that two years after the conference a number of working collaborations should be in place. Some may be focused on a single country and a single research problem and others may be organized through much larger international networks.

Qualifications of the PIs and Host Institution

Baumgartner and McCarthy have been involved in large-scale data collection projects throughout their careers. Baumgartner co-directs the Policy Agendas Project and McCarthy is involved in that project as well as co-director of the large-scale study of protest and social movement activities. Penn State houses the Association of Religious Data Archives, the

Correlates of War Project, and the Quantitative Social Science Research Institute, each of which will be involved here as well. Baumgartner and McCarthy have raised seed funding from Penn State to support a small workshop version of this conference, but seek to expand this from only the social movements and policy agendas scholars with whom they are personally in contact to a much broader audience. At the same time, we hope to move beyond developing applications for use in specific research projects to developing collaborations that can lead to more generally applicable tools for use across the social sciences.

List of Participants

PIs have already been in contact with a number of scholars who have expressed interest in attending the proposed conference. The list of outside invitees will include approximately 30-35 senior scholars, many from across nations of the European Union. Approximately 10-15 Penn State Scholars will also attend. Penn State University will provide resources to pay the expenses of an additional 10 to15 junior scholars and advanced graduate students to attend the conference. Below is a list of scholars who have already indicated interest in attending the conference and a supplementary list of scholars we intend to invite, which will also be expanded, as well as a tentative list of Penn State scholars who will attend.

Invited:

Edouard Hovy, University of Southern California, Information Sciences Institute (ISI)
Jamie Callan, Carnegie Mellon University, Language Technologies Institute
Stephen Purpura, Harvard University, Kennedy School
Stuart Shulman, University of Pittsburgh, School of Information Science
Bryan D. Jones, University of Washington, Co-Director, Policy Agendas Project
John Wilkerson, University of Washington, Co-Director, Policy Agendas Project
Stefaan Walgrave, University of Antwerp, Director, M2P, Media, Movements, and Politics
Christoffer Green Pedersen, Univesity of Aarhus, Director, Danish Agendas Project
Sylvain Brouard, Cevipof / Sciences Po Paris, Co-Director, French Agendas Project
Peter John, University of Manchester, British Agendas Project
Grant Jordan, University of Aberdeen, Political Science
Gerard Breeman, University of Leiden, Dutch Agendas Project
David Lowery, University of Leiden, Dutch Agendas Project
Francesco Zucchini, University of Milan, Political Science
Laura Chaques, University of Barcelona, Political Science
Frederic Varone, University of Geneva, Political Science
Dieter Rucht, WZB, Berlin, Director of Research Group on Civil Society, Citizenship and Political Mobilisation in Europe
Hans Peter Kriesi, University of Zurich, Sociology
Bert Klandermans, Free University of Amserdam, Sociology
Sarah Soule, Cornell University, Sociology
Craig Jenkins, Ohio State University
Doug McAdam, Stanford University, Sociology
Donatella Della Porta, European University Institute, Sociology
Chris Bader, Sociology, Baylor University (Association of Religion Data Archives)

Penn State Faculty:

Frank R. Baumgartner, Penn State University, Co-Director, Policy Agendas Project
John McCarthy, Penn State University, Professor and Head of Sociology
Roger Finke, Penn State University, Director, Association of Religious Data Archives
Burt Monroe, Penn State University, Political Science
Scott Bennett, Penn State, Political Science (COW Project)
Glenn Palmer, Penn State Political Science (COW Project)
Errol Henderson, Penn State Political Science (COW Project)
Suzie De Boef, Penn State Political Science
David Baker, Penn State, Educational Theory and Policy
Mark Anner, Penn State, Labor Studies.
Lee Ann Banaszak, Penn State Political Science.

References

Blei, David M. and John D. Lafferty. 2006. Dynamic Topic Models. Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh. pp. 113-20.

Boydstun, Amber E. 2007. Agenda Setting and Issue Framing Dynamics on Front Page News. Research in progress based on NSF dissertation award SES–0617492.

Collins-Thompson, K., and J. Callan. 2007. Automatic and human scoring of word definition responses. In *Proceedings of the NAACL-HLT 2007 Conference*. Rochester. Forthcoming.

Earl, Jennifer, S. Soule and J. D. McCarthy. 2005. Protest under Fire? Explaining the Policing of Protest. American Sociological Review 68 (4): 581-606.

Gilens, Martin. 1999. *Why Americans Hate Welfare: Race, Media, and the Politics of Antipoverty Policy.* Chicago: University of Chicago Press.

Grim, Brian and Roger Finke, Forthcoming. De jure and De facto Regulation of Religion: Developing and Assessing International Religion Indices.

Hovy, E.H., M. Marcus, M. Palmer, S. Pradhan, L. Ramshaw, and R. Weischedel. 2006. OntoNotes: The 90% Solution. Short paper. Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2006). New York, NY.

McAdam, Doug and Y. Su. 2002. The War at Home: Antiwar Protests and Congressional Voting, 1965 to 1973. American Sociological Review 67 (5): 696-721.

Newman, David, Caitanya Chemudugunta, and Padhraic Smyth. 2006. Statistical Entity-Topic Models. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia. pp. 680-6.

Peterson, Richard A. and R Arnaud. 2004. The Production of Culture Perspective. Annual Review Of Sociology 30: 311-334.

Purpura, S., and Hillard D. 2006. Automated Classification of Congressional Legislation. Proceedings of the National Conference on Digital Government (dg.o.2006). San Diego, CA.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, Dragomir R. Radev. 2006. An Automated Method of Topic-Coding Legislative Speech Over Time with Application to the 105th-108th U.S. Senate. Paper presented to the Society for Political Methodology. University of California at Davis, July 20-22.

Roscigno, Vincent and R. Hodson. 2004. The Organizational and Social Foundations of Worker Resistance. *American Sociological Review* 69 (1): 14-39.

Sahoo, N., J. Callan, R. Krishnan, G. Duncan, and R. Padman. 2006. Incremental hierarchical clustering of text documents. In *Proceedings of the Fifteenth International Conference on Information and Knowledge Management (CIKM'06)*. ACM.

Schrodt, Phil, and Deborah J. Gerner. 1994. Validity Assessment of a Machine-Coded Event Data Set for the Middle East, 1982-1992. *American Journal of Political Science* 38:825-854.

Shulman, S, E. Hovy, J. Callan, and S. Zavestoski. 2006. Progress in language processing technology for electronic rulemaking (research highlight). Proceedings of the Sixth National Conference on Digital Government Research (pp 249-250). San Diego, CA.

Shulman, Stuart W. 2007. Whither Deliberation? Mass e-Mail Campaigns and U.S. Regulatory Rulemaking. *Journal of E-Government* 3 (3). Forthcoming.
Snow, David A. Forthcoming. Framing Processes and Social Problems.
Stathan, Paul, R. Koopmans, M. Guini and F. Passey.2005. Resilient or adaptable Islam? Multiculturalism, religion and migrants' claims-making for group demands in Britain, the Netherlands and France. *Ethnicities* 5 (4): 427-459.