

SGER: A Testbed for eRulemaking Data

There is a pressing need for synergy between social and computational scientists that will make baseline data available to a growing eRulemaking community of researchers and practitioners. The goal of this project is to merge existing Information Technology (IT) capabilities with social science knowledge in the development of an Internet-based testbed for data gathered through www.regulations.gov. The testbed will also serve as a pilot demonstration of experimental tools that rule-writers, researchers, and the public may eventually use when sorting through potentially voluminous public comment. Since the process of rule-writing is complex and exacting, information technology can assist rule-writers and the public as they seek to develop better and more durable regulations.

A key aspect of rule-writing involves text. Rulemaking involves text collection, organization, categorization, retrieval, cross-indexing, and summarization. The text may come from existing regulations, studies, articles, or the public. In creating regulations, rule-writers need enhanced text management capabilities to help them organize and process large amounts of text. When providing input, the public could use text management capabilities to help them locate the discussion threads most relevant to their concerns. It is here that we see the greatest potential for synergy between IT and social science research. In this SGER we will explore the potential for social science to shape the direction and purpose of the IT research.

We identify a particular set of IT capabilities at the heart of the text processing needs of both the rule writers and the public, and propose a testbed environment that incorporates prototypes of these capabilities. The testbed will allow:

1. IT and social science researchers begin studying what specific technology best suits the needs of this text management problem;
2. rule-writers to begin learning about the capabilities that sophisticated new IT offers them by encouraging their input with a working prototype;
3. select members of the interested public to explore and shape prototype IT.

SGER Project Team

Government Partner: Oscar Morales, EPA, director of the federal eRulemaking initiative and manager of the website www.regulations.gov. Mr. Morales and his staff will provide appropriate data for the testbed, structured access to the site development staff, guidance about legal constraints, and knowledge of IT functionality and deployment requirements.

IT Partners: Dr. Jamie Callan, CMU, and Dr. Eduard Hovy, USC/ISI, who will provide tailored versions of their text processing technology, integrated and parameterized, to work on the testbed data.

Social Science Partners: Dr. Stuart Shulman, Drake University, and Dr. Stephen Zavestoski, University of San Francisco, who will work with members of the government and the public to solicit feedback that will be used to refine the testbed.

An eRulemaking Testbed: Toward Synergy Across Disciplines

Two of the PIs have been fostering a social science research community focused on the acquisition and analysis of public comment data from eRulemaking for the last two years (Shulman 2003; Shulman et al. 2003; Zavestoski & Shulman 2002). One goal for this SGER is to attract the interest of IT researchers, particularly from the Information Retrieval, Text Mining, and Natural Language Processing sub-fields. A prerequisite for expanding the research community is stable, widely-available datasets that can be used to evaluate new techniques. The creation of an eRulemaking testbed, which will provide data of use to computer and social scientists, will establish a common ground from which future interdisciplinary collaborations can be launched.

Shulman and Zavestoski will draw on their ongoing analysis of comments submitted during previous rulemaking processes to help Callan and Hovy begin shaping the clustering, threading, and summarization tools. The goal will be to increase the usefulness of the testbed for social science researchers, members of the public, and administrators of federal agencies when they are evaluating the effectiveness of eRulemaking. Shulman will focus on defining the types of data on end-user characteristics that will be most useful to future eRulemaking research. Expanding on the research needs specified at the January 2003 eRulemaking workshop at the Kennedy School of Government and an informal survey of other social scientists, Shulman will develop a typology of the end-user data that researchers can use to measure the impact of eRulemaking on citizen-government interaction. Possible end-user characteristics that the testbed will need to be able to identify include, but are not limited to, the length of time users spend at various www.regulations.gov pages and the number of pages visited. As the desired types of end-user data are identified, they will be conveyed to the IT partners who will incorporate the ability to gather information about these characteristics into the testbed.

Zavestoski will focus on identifying the data needs of researchers concerned with the impacts of eRulemaking on the content and quality of electronic public comments. He will identify discursive democratic priorities to ensure that the prototype clustering, threading, and summarization tools provide meaningful data access and analysis functionality. Together, Shulman and Zavestoski will supervise undergraduate research assistants, who will conduct comprehensive reviews of the social science literature on public participation processes. The results of this review will be used to refine the search for novel data amenable to collection by Mr. Morales and his eRulemaking team at the EPA.

We propose to assemble one or more large datasets of unstructured text public commentary. These datasets will be packaged with supporting materials and documentation about the regulations being discussed. They will be distributed from a public Web site created for this purpose. The Web site will contain links to introductory papers and threaded discussions on eRulemaking and will be hosted by either CMU or USC/ISI using their existing Web servers. The datasets and the supporting Web site are two of the main deliverables of the proposed grant. As a result of the SGER, the project team will:

1. construct the testbed;
2. determine through focus groups requirements for additional text processing functionality and other characteristics as desired by government staff and the public;
3. create a research plan to develop such functionality, as feasible and appropriate in the research environment, and to study and evaluate its utility and performance; and
4. develop a long-term proposal to submit to the NSF's Digital Government program to perform this research.

Information Technology

Initial discussions with our government partner Mr. Morales and others have indicated the primary need for three core functionalities, each embodied in its own technology:

1. Text retrieval
2. Text clustering and topic threading
3. Text summarization

Providing these functionalities, the testbed will enable a government rule-writer to locate from a mass of texts those most pertinent to his or her current endeavor, to then request that the relevant (sub)topics be threaded, which is especially useful for following an online public commentary discussion, and finally to request summaries to arbitrary lengths of the retrieved and/or threaded material, each summary oriented to whatever topic the rule-writer needs to focus on. Similarly, the testbed allows a member of the public to identify which regulations or other texts are pertinent to his or her current

interest, to follow its creation and evolution via discussion threads, and to take away summaries of the most pertinent aspects of the discovered material. We discuss the three functionalities in turn.

Text Retrieval

Large databases of unstructured public comments are becoming the norm, but so far there has been little support for browsing and searching this type of data. Government analysts and the public have become familiar with Web search engines and Web directories, and will expect these capabilities when they begin using large online public comment databases regularly. However, there has been little research on how current techniques work on this type of data, and there are several reasons to suspect that they might not. While the team working with Mr. Morales at the EPA has planned enhancements for Phase II and III of www.regulations.gov, for now the functionality remains an open question.

Public comment data does not have the hyperlink structure and HTML formatting that Web search algorithms rely on. It is not the carefully-edited newspaper and magazine articles studied in research fora such as TREC, nor is it the well-structured data studied in XML document retrieval fora such as INEX. Indeed, in many ways it looks like a worst-case scenario for current information retrieval algorithms: many short documents that, by definition, address more-or-less the same topic. Little is known about how government analysts and the general public would want to access public comment data, for example, how they would want it organized, and what types of queries they would use to search it. We will bring social science to bear so as to increase the understanding of the types of data organization the government, public, and researchers would need. In so doing, we will strive for synergy between the IT and social science partners that will lead to the best possible uses of IT in the eRulemaking arena.

We propose to work with the EPA to build some large, representative text databases of public comment data, and to investigate the types of access mechanisms that might be useful. We will begin by providing search access using the Lemur Toolkit, an open-source toolkit for language modeling and information retrieval research (Ogilvie and Callan, 2001; <http://www.cs.cmu.edu/~lemur>). Lemur includes a variety of ad-hoc retrieval and text summarization algorithms, and is capable of handling the required data volume “out of the box.” We will investigate how well current techniques support the likely information needs of government analysts and the public, and demonstrate these abilities to the EPA.

Text Clustering and Topic Threading

Both CMU and USC/ISI have extensive experience with text clustering algorithms. USC/ISI’s text clustering package ISICL includes the standard clustering algorithms, such as SLINK, CLINK, k-Nearest Neighbor, k-Means, etc. Our most time-consuming work with regard to clustering will be to identify the clustering algorithm(s) most appropriate to the types of requests posed by users and to the types of source text provided by the government and the public. This work might require a series of comparative experiments and a subsequent characterization of which queries/material should activate which algorithm. The remaining work for clustering is relatively mundane: deploying the algorithms in the testbed architecture, ensuring information flow via APIs, etc.

Topic threading functionality is less well developed than clustering, and will require some joint development by the two IT partners. We will develop a rudimentary topic threading capability for the testbed based on the needs specified by the social science partners, and learn what is involved for this application, so as to guide subsequent research planning for the proposal, and ultimately the actual research. To identify the most strongly apparent topics in the source material, we will use topic signature technology developed at ISI (Lin and Hovy 2000). A topic signature characterizes a topic by a set of regularly co-occurring words, each word with an associated strength. Experiments with several word selection methods over the past five years (including *tf*, *tf.idf*, χ^2 , and the likelihood ratio λ (Dunning 1993)) enable us to determine fairly easily which kind of signature will work well for the application.

Armed with these signatures, and using some of the summarization modules described below to score parts of sentences, we will be able to group together sentences according to their similarity to each

signature. Arranging these groupings in creation order, we will produce a first-cut topic thread. Although this rather simple procedure leaves open many questions and possibilities for tuning, we will produce something that works to some degree and leave the refinement and development of better algorithms for the research stage.

Text Summarization

R&D of text summarization technology has been conducted in the Natural Language Group at USC/ISI since the late 1990s, when interest in automated text summarization re-emerged. During this time Dr. Hovy and colleagues, notably Dr. Chin-Yew Lin, have constructed a number of text summarization systems, including Summarist (Hovy and Lin 1999), NeATS (Lin and Hovy 2002), and an as-yet unnamed system that creates headlines of a text or cluster of texts automatically (Zhou and Hovy 2003). They have also spun off a startup company named Infosqueeze that has commercialized this technology.

Both Summarist (which summarizes a single document at a time) and NeATS (which produces one summary of a cluster of related documents) can generate summaries of arbitrary length, as specified by the user. In addition, the user can provide one or more topics of interest, for example in the form of keywords, according to which summaries are then produced. Both systems produce summaries that are *extractive* (which means they identify and return the most pertinent sentences from the source material, rather than re-writing the material into their own sentences) and *informative* (which means that they provide details, instead of just the general theme or topic of the source text).

In very general terms, the systems work as follows. First, pre-processing modules prepare the input text, by stripping away formatting and other markers, determining part of speech and other information for each word, and standardizing word forms (removing plurals, etc.). Next, several independent modules read the text and provide a goodness score for each word and/or each sentence. Each module considers a unique set of factors. One simple technique rewards sentences containing the concepts of the title; another more complex technique uses so-called topic signatures that take into account the relative information-bearing power of each word in its context, using sophisticated information theoretic measures. Other modules, such as Optimum Position Policy (OPP) employ more complex techniques (Lin and Hovy 1997). Most of these modules can be trained to adapt to different subjects areas (such as politics, biology, sports, medicine, or business) and/or different text genres (such as news articles, medical reports, patents, or web pages). In the third stage, the Adaptive Module Integrator (AMI) module combines the different goodness scores for each sentence, adapting them according to the relative importance of their modules in the input text's domain and genre. This integrator uses machine learning techniques to automatically construct the optimal combination function (Lin 1999). In the final stage, the sentences are internally processed to overcome certain disfluencies, after which they are ranked in the order desired by the user, reformatted, and delivered to the user.

Both Summarist and NeATS combine state of the art research with solid software engineering practice. Both systems have been tested in national trials, sponsored by DARPA and/or ARDA, in which systems from around the world compete on unseen data and are evaluated by government employees. In the SUMMAC evaluation conducted by the Department of Defense in 1996, Summarist was competitive in the generic summarization trial with the best systems taking part (Firmin and Chrzanowski 1999). In the DUC01 trial conducted by NIST, NeATS tied for first place and in DUC02 trial scored second (DUC 2001; 2002).

Our proposed work with regard to summarization is to deploy the systems in the testbed architecture, configure them via APIs to communicate with the other systems, the data sources, and the interface, and to train or otherwise specify the parameters mentioned above (relative weightings of modules, scoring combination function, etc.) to provide good utility for testbed usage.

Text Analysis: Enabling Government Rule Writers and Social Science Researchers

It is likely that government analysts and the general public will need more than simple browsing and search access to large databases of public commentary. We propose to begin investigating the effectiveness of simple text analysis capabilities for characterizing the contents of large databases of text analysis. Our initial focus will be on identifying supporting and opposing comments in order to compare and contrast the justifications, language, and themes employed in each. We will also develop approaches to analyzing text that can measure a commenter's depth of knowledge about a proposed rule. Another development will be techniques for identifying stakeholder communities represented in public comment databases, and for classifying submitted comments according to which stakeholder communities they come from.

We aren't likely to solve all of these problems in a few months, but we will develop a proof-of-concept prototype and demonstrate it to the EPA, so as to begin educating EPA about the types of non-traditional access mechanisms that it should be considering for the future. We will also conduct focus group interviews with government personnel and citizens. Focus groups are directed discussions with researcher-defined groups intended to obtain group members' perceptions in an area of interest (Kreuger 1988; Merton et al. 1990). Per focus group research convention, we will conduct interviews with 2-3 groups of each type (government and citizen), as well as 2-3 groups consisting of a mix of government personnel and citizens.

Project Management

Drake University will serve as the lead institution and Shulman as the project director. Team e-mail, conference calls, and regularly scheduled meetings will facilitate the interdisciplinary cross-fertilization. The four PIs on the project will meet face-to-face with the EPA team in Washington, DC three times over a nine-month period and a fourth time at the dg.o 2003 meeting in Boston. A preliminary late spring 2003 meeting will scope out the technical and content issues involved in the creation of the testbed. A second meeting, hosted by Oscar Morales and the EPA in the summer, will bring together a broader cross-section of agency personnel, select members of the growing eRulemaking research community, interested citizens, a post doctoral research assistant from Iowa State University, and a number of graduate and undergraduate students from American University. At the summer meeting the IT partners will test and develop plans to refine pilot applications using live demonstrations. At the same meeting, the social science partners will conduct the focus group interviews in the wake of the live demonstrations. A final meeting, early in fall 2003, will allow the research team to develop a longer-term proposal for research that will build on the pilot established by the SGER.

We recognize that in an exploratory study it is not practical to propose new basic research. Instead, we propose to assemble from our existing text processing technology whatever functionality is clearly required for the testbed. The IT partners will integrate the technology into one or more prototypes and tune it for expected usage patterns. We expect this work to take around 9 months, performed at the two IT locations by two graduate student research assistants, each one supervised by a co-PI.

Although the PIs have not previously worked together as a group, the foundations for this collaboration are not entirely new. Previous collaboration between members of the NL groups at USC/ISI and CMU were productive. Shulman and Callan have been developing the basis for collaboration since their initial encounter at the dg.o 2001 meeting, and Shulman helped Callan acquire and begin working with the USDA organic public comment dataset last Fall. Shulman and Zavestoski have been refining their research methodology through collaboration for the last two years. Shulman has known Morales for several years, and all four PIs developed a positive working relationship with Morales at the recent KSG workshop on eRulemaking.

Works Cited

- DUC. 2001. *Proceedings of the Document Understanding Conference (DUC) Workshop on Multi-Document Summarization Evaluation*, at the SIGIR-01 Conference. New Orleans, LA. See <http://www.itl.nist.gov/iad/894.02/projects/duc/index.html>.
- DUC. 2002. *Proceedings of the Document Understanding Conference (DUC) Workshop on Multi-Document Summarization Evaluation*, at the ACL-02 Conference. Philadelphia, PA (forthcoming).
- Dunning, T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19, 61–74.
- Firmin, T. and M.J. Chrzanowski. 1999. An Evaluation of Text Summarization Systems. In I. Mani and M. Maybury (eds), *Advances in Automated Text Summarization*. Cambridge, MA: MIT Press, 325–335.
- Hovy, E.H. and C-Y. Lin. 1999. Automating Text Summarization in SUMMARIST. In I. Mani and M. Maybury (eds), *Advances in Automated Text Summarization*. Cambridge, MA: MIT Press, 81–97.
- Kreuger, R.A. 1988. *Focus Groups: A Practical Guide for Applied Research*. London: Sage.
- Lin, C-Y. and E.H. Hovy. 1997. Identifying Topics by Position. *Proceedings of the Applied Natural Language Processing Conference (ANLP-97)*, Washington, DC, 283–290.
- Lin, C-Y. 1999. Training a Selection Function for Extraction. *Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM)*, Kansas City, 1–8.
- Lin, C.-Y. and E.H. Hovy. 2000. The Automated Acquisition of Topic Signatures for Text Summarization. *Proceedings of the COLING Conference*. Strasbourg, France. August, 2000.
- Lin, C-Y. and E.H. Hovy. 2002. From Single to Multi-document Summarization: A Prototype System and its Evaluation. *Proceedings of the 40th Conference of the Association of Computational Linguistics (ACL)*. Philadelphia, PA.
- Merton, R.K., M. Fiske, and P.L. Kendall. 1990. *The Focused Interview: A Manual of Problems and Procedures* (2nd ed.). London: Collier MacMillan.
- Ogilvie, P. and J. Callan. 2002. Experiments Using the Lemur Toolkit. *Proceedings of the 2001 Text Retrieval Conference (TREC 2001)* (pp. 103-108). National Institute of Standards and Technology, special publication 500-250.
- Shulman, S.W. 2003. An Experiment in Digital Government at the United States National Organic Program, forthcoming in *Agriculture and Human Values*.
- Shulman, S.W., D. Schlosberg, S. Zavestoski, and D. Courard-Hauri. 2003. Electronic Rulemaking: New Frontiers in Public Participation, forthcoming in *Social Science Computer Review*.
- Zavestoski, S. and S. Shulman. 2002. The Internet and Environmental Decision-Making, *Organization and Environment* 15, 3 (Fall 2002), 323-327.
- Zhou, L.: and E.H. Hovy. 2003. In prep.