January 2007 Project Highlight[1]

Collaborative Research: Language Processing Technology for Electronic Rulemaking

Stuart Shulman, shulman@pitt.edu
Eduard Hovy, hovy@isi.edu
Jamie Callan, callan@cmu.edu
Stephen Zavestoski, smzavestoski@usfca.edu

Many people today — including news analysts, opinion pollsters, advertisers, and government regulation writers — need to interpret, structure, and rapidly master large quantities of opinion-based text. Our eRulemaking Research Group (http://erulmaking.ucsur.pitt.edu) focuses on the federal government's several thousand regulation writers, employed in some 200 agencies, who formulate, in a tightly scripted procedure, the rules and regulations that define the details of our laws. U.S. law requires that regulatory agencies invite comments from the public about proposed regulations, and respond in the final regulation to all substantive issues raised in comments. An agency may receive several hundred thousand form letters by email, academic studies and manuscripts of a few thousand pages, and anything in between. Processing comments quickly and accurately is expensive and challenging, especially when the comment volume is large. Our research develops text processing technology that assists in the analysis of such collections, for example content-based clustering, near-duplicate detection, stakeholder detection, opinion identification, and extractive summarization. These new technologies are deployed in a Web-based prototype, *The Rule-Writer's Workbench*, for government and social science partners to use and evaluate.

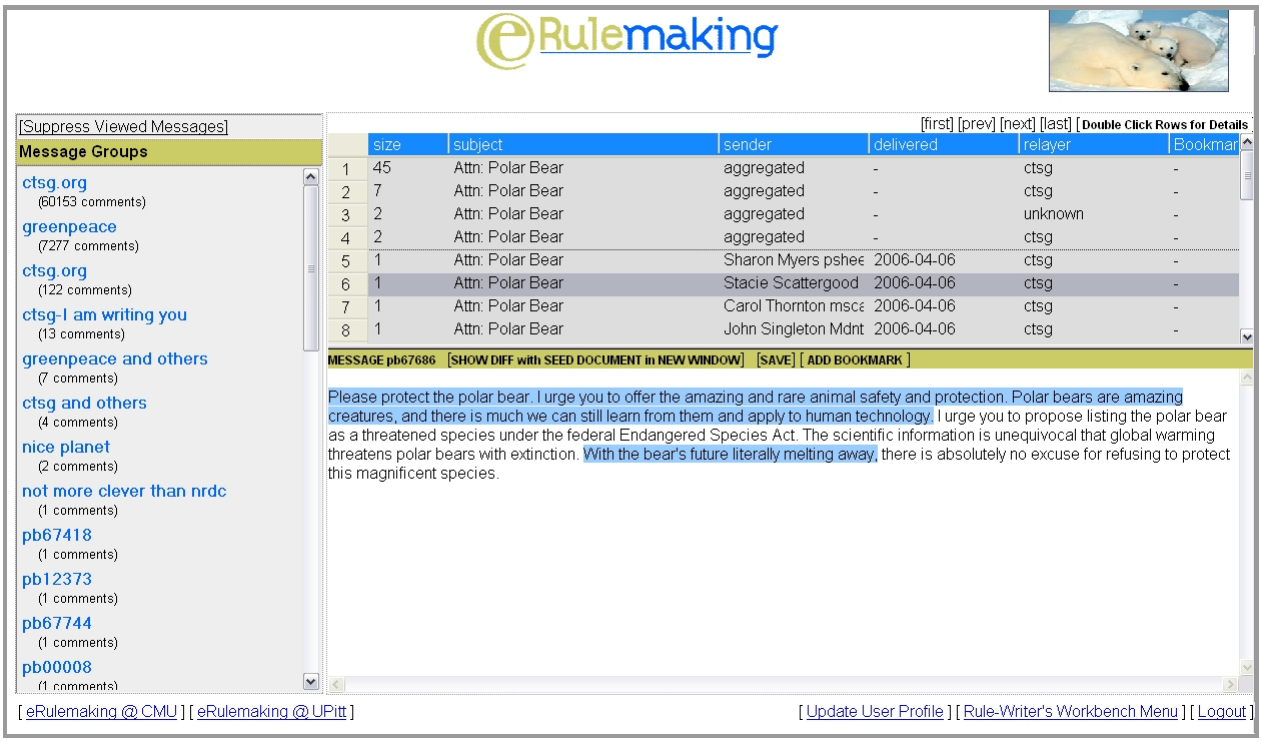
This project, funded under the NSF's Digital Government program, started in October 2004. In September 2006, we convened a series of three usability tests in cooperation with the Bureau of Labor Statistics (BLS) in Washington DC. Three successive groups of ten volunteer subjects drawn primarily from five federal agencies (USDA, DOT, EPA, FCC, and NSF) tested three toolsets in mock analyses of actual public comments submitted to the EPA. Tool 1, the testing baseline, offered simple search and browsing capability. Tool 2 clustered and counted identical and near duplicate comments and highlighted unique text passages inserted by commenters into interest group form letter e-mails. Tool 3 sorted passages of text into sub-topics, opinions and keywords, presenting the results in a variety of tables that allowed the testers to drill down into the relevant chunks of text in a systematic manner. After a 90 minute test in the computer lab, each group spent approximately one hour in a focus group with the project director providing feedback about the tools and guidance for future development. The subjects also discussed the role of public participation in the changing digital landscape of American government. Analysis of the transcripts produced in these focus groups has advanced our work significantly.

One surprising finding was that Tool 1, our presumed baseline, in fact constitutes a more usable and functional system than many rule writers currently have on their desktop. Tool 3 was judged to show significant promise, but is not yet sufficiently easy to use. Comments about a 'better version' of Tool 3 became the basis for rethinking the role of the user in a topic/opinion classification system and triggered a new proposal, submitted to NSF in late 2006.

Most encouraging was the unanimous praise for the logic, function, and maturity of Tool 2, which makes it possible for a single person to review and set aside examples reflecting hundreds or thousands of identical form letters in a matter of minutes. Once the 'dupes' are processed with the cluster counts displayed using Tool 2, the 'near-dupes' are made easily accessible for

browsing and the unique text added to a form letter is automatically highlighted. Testers agreed this tool was ready for use, but also made careful observations about how it might be improved.

Since these test sessions, the best ideas about improving Tool 2 have been incorporated into a modified Rule-Writer's Workbench. The Web-based system is currently being used by Fish & Wildlife Service (FWS) personnel in Anchorage Alaska who are collecting public comment on the proposed listing of polar bears as 'threatened' under the Endangered Species Act. In response to the proposed listing, several environmental groups have launched "Action Alerts" and generated over 100,000 emails in the first eight days of the 90-day comment period, most of which seek to sustain the listing but also to link the cause explicitly to global climate change. FWS personnel will use the new prototype to review, bookmark, and annotate the full set of e-mails received. The system will create a defensible legal record that logs the time and date that each unique comment was reviewed, as well as accurate counts of the exact duplicates. Less time will be spent finding the unique comments and more time will be spent reviewing the passages of text inserted by interest group members.



Screenshot of the DURIAN interface for sorting form letters and finding unique text

[1] This research was made possible with funding from the National Science Foundation (IIS 0429293, 0429102, 0429360 & 0429243). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation.